On the Distance of Databases

Gyula O.H. Katona[1], Anita Keszler[2], and Attila Sali[1]

[1] Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences,
Budapest, P.O.B. 127, H-1364 Hungary
{ohkatona,sali}@renyi.hu
[2] Distributed Events Analysis Research Group,
Computer and Automation Research Institute,
Hungarian Academy of Sciences,
1111 Budapest, Kende u. 13-17, Hungary
keszler@sztaki.hu

Abstract. In the present paper a distance concept of databases is investigated. Two database instances are of distance 0, if they have the same number of attributes and satisfy exactly the same set of functional dependencies. This naturally leads to the poset of closures as a model of changing database. The distance of two databases (closures) is defined to be the distance of the two closures in the Hasse diagram of that poset. We determine the diameter of the poset and show that the distance of two closures is equal to the natural lower bound, that is to the size of the symmetric difference of the collections of closed sets. We also investigate the diameter of the set of databases with a given system of keys. Sharp upper bounds are given in the case when the minimal keys are 2 (or r)-element sets.

Keywords: distance of databases, keys, antikeys, closures, poset, Hasse diagram.

1 Introduction

We are trying to investigate questions like "when two databases are the same?" or "are two databases similar?". For instance if we add or delete a row of the database we say that it changes only the instance of the given database. But in a strict sense this change may change the dependency structure, so we might say that a new database has been obtained. Another example is when a new attribute is added to the database, all rows are completed with the value of the new attribute. Is this another database?

If two databases should be considered different, how different they are? One needs a notion of *distance* between databases. If such a notion is introduced one can ask interesting questions like the following one. Knowing the distance between two databases what will be the distance between the merged database and one of the original ones?

Papers of Müller et.al. [9,10] treat this question from the point of view of conflicting copies of scientific databases. Today, many scientific databases overlap in

their sets of represented objects due to redundant data generation or data replication. For instance, in life science research it is common practice to distribute the same set of samples, such as clones, proteins, or patient's blood, to different laboratories to enhance the reliability of analysis results. Whenever overlapping data is generated or administered at different sites, there is a high probability of differences in results. These differences do not need to be accidental, but could be the result of different data production and processing workflows. For example, the three protein structure databases OpenMMS [2], MSD [3], and Columba [11] are all copies of the Protein Data Bank PDB [1]. However, due to different cleansing strategies, these copies vary substantially. Thus, a biologist is often faced with conflicting copies of the same set of real world objects and with the problem of solving these conflicts to produce a consistent view of the data. Thus, Müller et.al propose a distance concept that is similar to the edit distance of strings. They study the problem of efficiently computing the update distance for a pair of relational databases. In analogy to the edit distance of strings, the update distance of two databases is defined as the minimal number of set-oriented insert, delete and modification operations necessary to transform one database into the other.

In the present paper we take a data-mining oriented approach. We are mostly interested in the apparent dependency structure of a database, that is all functional dependencies that are satisfied by the given instance. We answer the questions posed above only for a very modest special case. It will be supposed that two database are the same if they have the same number of attributes and the system of functional dependencies are identical. The distance is introduced only between two databases having the same number of attributes.

Functional dependencies lead naturally to closure operations on the set of attributes $\mathcal{R}$. A subset $A \subseteq \mathcal{R}$ is *closed* if $A \to B$ implies $B \subseteq A$. Thus we take the approach of Burosch et.al. [4] by considering the *poset of closures* as a model of changing database. We will define the distance of two instances as the distance of the closures they generate, in that poset. We will see that this distance of two database instances is equal to the minimum number of tuples that are needed to be added to or removed from one instance to obtain the other instance of the database schema. In Section 2 we show that the largest distance possible is attained between the minimal and maximal elements of this poset. It will turn out that the distance of two closures is in fact *equal to* the obvious lower bound, that is to the size of the symmetric difference of the sets of closed sets. In Section 3 we consider the question that if something is known about the database instance, what is the diameter of the space database instances, that is the set of closures that satisfies the given information. This could be interpreted as a datamining question, we have some information given, and we wonder what is the size of the search space of databases based on that information. Our particular interest is in the case when the number of minimal keys is given. We give an upper bound in the general case. It is also interesting if the minimal keys all have the same cardinality r. We give a sharp upper bound for the diameter

in the case $r = 2$. Finally, Section 4 contains conclusions and future research directions.

Some combinatorics notation we use that may not be well known for database people are as follows. $[n]$ denotes the set of first n positive integers, that is $[n] = \{1, 2, \ldots, n\}$. If Z is a set, then 2^Z denotes the collection of all subsets of Z, while $\binom{Z}{r}$ denotes the set of all r-element subsets of Z. These latter notations are mostly used in case of $Z = [n]$.

2 Poset of Closures

In what follows a schema $\mathcal{R}$ is considered fixed and every instance $\mathbf{r}$ is considered together with *all* functional dependencies $A \to B$ such that $\mathbf{r} \models A \to B$. For a set of attributes $A \subseteq \mathcal{R}$ the *closure* of A is given by $\ell(\mathbf{r})(A) = \{a \in \mathcal{R} \colon \mathbf{r} \models A \to a\}$. It is well known that the function $\ell(\mathbf{r}) \colon 2^{\mathcal{R}} \longrightarrow 2^{\mathcal{R}}$ is a *closure* that is it satisfies the properties

$$\begin{aligned} &A \subseteq \ell(A) \\ &A \subset B \Longrightarrow \ell(A) \subseteq \ell(B) \\ &\ell(\ell(A)) = \ell(A). \end{aligned} \tag{1}$$

Since constant columns are not really interesting, we assume that $\ell(\emptyset) = \emptyset$. Attribute set A is *closed* if $A = \ell(A)$. It is known that the family of closed attribute sets forms an intersection-semilattice. It was observed in [4], and is also well known consequence of the fact that functional dependencies can be expressed in First Order Logic by universal sentences, that if $\mathbf{r} \models A \to B$, then $\mathbf{r}' \models A \to B$ holds for any $\mathbf{r}' \subset \mathbf{r}$. That is a valid functional dependency stays valid if a record is removed from the database. This suggested the investigation of the poset of closures as a model of changing databases. Closure ℓ_1 is said to be *richer* than or equal to ℓ_2, $\ell_1 \geq \ell_2$ in notation, iff $\ell_1(A) \subseteq \ell_2(A)$ for all attribute sets. Let $\mathcal{F}(\ell)$ denote the collection of closed attribute sets for closure ℓ. It was proved in [4] that

Proposition 2.1. *$\ell_1 \leq \ell_2$ iff $\mathcal{F}(\ell_1) \subseteq \mathcal{F}(\ell_2)$.*

If $\mathbf{r_1}$ and $\mathbf{r_2}$ are two instances of schema $\mathcal{R}$, then $\ell(\mathbf{r_2})$ is richer than $\ell(\mathbf{r_1})$ can be interpreted as follows. In $\mathbf{r_2}$ there are more subsets of attributes that only determine attributes inside them, that is in $\mathbf{r_2}$ we need more attributes to determine some attribute, so $\mathbf{r_2}$ conveys 'more information" in the sense that the values of tuples are more abitrary.

The covering relation was also characterized. ℓ_2 *covers* ℓ_1, if $\ell_1 \leq \ell_2$ and for all ℓ' such that $\ell_1 \leq \ell' \leq \ell_2$ either $\ell' = \ell_1$ or $\ell' = \ell_2$.

Proposition 2.2 ([4]). *ℓ_2 covers ℓ_1 iff $\mathcal{F}(\ell_1) \subseteq \mathcal{F}(\ell_2)$ and $|\mathcal{F}(\ell_2) \setminus \mathcal{F}(\ell_1)| = 1$.*

Proposition 2.2 shows that the poset of all closures over a given schema $\mathcal{R}$, $\mathbf{P}(\mathcal{R})$, is *ranked*: its elements are distributed in *levels* and if ℓ_2 covers ℓ_1, then ℓ_2 is in the next level above ℓ_1's one. Let $|\mathcal{R}| = n$. Since $\emptyset$ and $\mathcal{R}$ are both closed for any closure considered, we obtain [4] that the *height* of $\mathbf{P}(\mathcal{R})$ is $2^n - 2$.

Let $\ell(\mathbf{r})$ denote the closure obtained from using all functional dependencies satisfied by instance $\mathbf{r}$. Since removing a record or adding a record to an instance changes the rank (level) of $\ell(\mathbf{r})$ by at most one, the distance of two instances are defined as follows.

Definition 2.1. *Let $\mathbf{r}$ and $\mathbf{r}'$ be two instances of schema $\mathcal{R}$. Their distance $d(\mathbf{r}, \mathbf{r}')$ is defined to be the graph theoretic distance of $\ell(\mathbf{r})$ and $\ell(\mathbf{r}')$ in the Hasse diagram of $\mathbf{P}(\mathcal{R})$. That is, the length of the shortest path between points $\ell(\mathbf{r})$ and $\ell(\mathbf{r}')$ using only covering edges.*

It is easy to check that $d(\mathbf{r}, \mathbf{r}')$ satisfies the triangle condition. The following proposition shows that the height of $\mathbf{P}(\mathcal{R})$ is an upper bound of the largest possible distance between two instances of the schema $\mathcal{R}$. Since the distance of instances is defined using the distance of closures, we allow the ambiguity of speaking of distance of an instance and a closure, as well.

Proposition 2.3. *Let $|\mathcal{R}| = n$. Then $d(\mathbf{r}, \mathbf{r}') \leq 2^n - 2$ for any two instances of the schema $\mathcal{R}$.*

Proof. Let ℓ_m be the minimum element of $\mathbf{P}(\mathcal{R})$ furthermore let ℓ^M be the maximum element. Then

$$d(\mathbf{r}, \mathbf{r}') \leq d(\mathbf{r}, \ell_m) + d(\ell_m, \mathbf{r}') \tag{2}$$

and

$$d(\mathbf{r}, \mathbf{r}') \leq d(\mathbf{r}, \ell^M) + d(\ell^M, \mathbf{r}') \tag{3}$$

Thus,

$$2\, d(\mathbf{r}, \mathbf{r}') \leq d(\mathbf{r}, \ell_m) + d(\mathbf{r}, \ell^M) + d(\ell_m, \mathbf{r}') + d(\ell^M, \mathbf{r}') = 2^n - 2 + 2^n - 2. \tag{4}$$

□

For any instance $\mathbf{r}$ the rank of $\ell(\mathbf{r})$ in $\mathbf{P}(\mathcal{R})$ is $d(\ell_m, \mathbf{r}) = |\mathcal{F}(\ell(\mathbf{r}))| - 2$.

The following is the main result of this section.

Theorem 2.1. *For any two instances $\mathbf{r}$ and $\mathbf{r}'$ of the schema $\mathcal{R}$ we have*

$$d(\mathbf{r}, \mathbf{r}') = |\mathcal{F}(\ell(\mathbf{r})) \,\Delta\, \mathcal{F}(\ell(\mathbf{r}'))|, \tag{5}$$

where $A \Delta B$ denotes the symmetric difference of the two sets, i.e., $A \Delta B = A \setminus B \cup B \setminus A$.

In order to prove Theorem 2.1 we need to recall the following result.

Theorem 2.2 ([5]). *A collection $\mathcal{F}$ of subsets of $\mathcal{R}$ is the collection of closed sets of some closure $\ell(\mathbf{r})$ for an appropriate instance $\mathbf{r}$ of $\mathcal{R}$ iff $\emptyset, \mathcal{R} \in \mathcal{F}$ and $\mathcal{F}$ is closed under intersection.*

Proof (of Theorem 2.1). According to Proposition 2.2 the size of $\mathcal{F}(\ell(\mathbf{r}))$ changes by one when we traverse along a covering edge in the Hasse diagram of $\mathbf{P}(\mathcal{R})$.

This immediately gives that the left hand side of (5) is at least as large as the right hand side, that is the distance of two instances is lower bounded by the size of the symmetric difference of the respective families of closed sets.

In order to prove the inequality in the other direction we have to find a way to move from $\ell(\mathbf{r})$ to $\ell(\mathbf{r}')$ using $|\mathcal{F}(\ell(\mathbf{r})) \,\Delta\, \mathcal{F}(\ell(\mathbf{r}'))|$ covering edges. That is, according to Theorem 2.2 we have to move from $\mathcal{F}(\ell(\mathbf{r}))$ to $\mathcal{F}(\ell(\mathbf{r}'))$ by successively removing a set of $\mathcal{F}(\ell(\mathbf{r})) \setminus \mathcal{F}(\ell(\mathbf{r}'))$ or adding a set of $\mathcal{F}(\ell(\mathbf{r}')) \setminus \mathcal{F}(\ell(\mathbf{r}))$ so that the property of being closed under intersection is preserved in each step. Note, that $\emptyset, \mathcal{R}$ are members of both closed sets system and the operations done do not change this.

First, we "peel off" sets of $\mathcal{F}(\ell(\mathbf{r})) \setminus \mathcal{F}(\ell(\mathbf{r}'))$ one by one. Let $F \in \mathcal{F}(\ell(\mathbf{r})) \setminus \mathcal{F}(\ell(\mathbf{r}'))$ such that $F \neq \mathcal{R}$, but there is no $F' \in \mathcal{F}(\ell(\mathbf{r})) \setminus \mathcal{F}(\ell(\mathbf{r}'))$ such that $F \subsetneq F' \subsetneq \mathcal{R}$. F can be removed from $\mathcal{F}(\ell(\mathbf{r}))$ if it is not an intersection of two closed sets both different from F. However, if $F = F' \cap F''$, then by the maximality property of F, both $F', F'' \in \mathcal{F}(\ell(\mathbf{r}'))$ yielding the contradiction that $F \in \mathcal{F}(\ell(\mathbf{r}'))$, as well.

Repeating the step above as long as $\mathcal{F}(\ell(\mathbf{r})) \setminus \mathcal{F}(\ell(\mathbf{r}'))$ is nonempty we arrive at $\mathcal{F}(\ell(\mathbf{r})) \cap \mathcal{F}(\ell(\mathbf{r}'))$. Now, we have to add the sets of $\mathcal{F}(\ell(\mathbf{r}')) \setminus \mathcal{F}(\ell(\mathbf{r}))$ one-by-one. In order to do so, let G be a minimal element of $\mathcal{F}(\ell(\mathbf{r}')) \setminus \mathcal{F}(\ell(\mathbf{r}))$ with respect to set containment. If $F \in \mathcal{F}(\ell(\mathbf{r})) \cap \mathcal{F}(\ell(\mathbf{r}'))$ then $F \cap G$ is a proper subset of G, or G itself. Since $\mathcal{F}(\ell(\mathbf{r}'))$ closed under intersection, $F \cap G \in \mathcal{F}(\ell(\mathbf{r}'))$. If it is a proper subset of G, then by the minimality of G, $F \cap G \in \mathcal{F}(\ell(\mathbf{r}))$, holds as well. In either case, adding G to $\mathcal{F}(\ell(\mathbf{r})) \cap \mathcal{F}(\ell(\mathbf{r}'))$, the collection remains closed under intersection. □

3 Diameter of Collection of Databases with the Same Set of Minimal Keys

It was investigated in [4] when does the system of keys determine uniquely the closure, that is the system of functional dependencies. A subset $K \subseteq \mathcal{R}$ is a *key* if $K \to \mathcal{R}$, and it is *minimal* if no proper subset of K has this property. In other words, K is a minimal key for instance $\mathbf{r}$ if there are no two rows of $\mathbf{r}$ that agree on K, but K is minimal with respect to this property. Let $\mathcal{K}(\mathbf{r})$ denote the system of minimal keys of instance $\mathbf{r}$. It is clear that $\ell(\mathbf{r})$ uniquely determines $\mathcal{K}(\mathbf{r})$, since $A \to B$ holds iff $B \subseteq \ell(A)$. On the other hand, $\mathcal{K}(\mathbf{r})$ does not determine $\ell(\mathbf{r})$. A simple example is the following. Let $\mathcal{R} = \{a, b, c, d\}$, and let the family of keys be $\mathcal{K} = \{\{a, c\}, \{a, d\}, \{b, c\}, \{b, d\}\}$. Then closure ℓ_1 has $\mathcal{K}$ as system of keys, where ℓ_1-closed sets are $\emptyset, \{a, b\}, \{c, d\}, \{a, b, c, d\}$. On the other hand, $\ell_2 > \ell_1$ has the same key system, where additionally the one-element subsets are ℓ_2-closed, too.

In order to determine the space of closures with a given minimal key system we need to introduce the concept of *maximal antikeys*. A subset $A \subset \mathcal{R}$ is a maximal antikey if it does not contain any key, and maximal with respect to this property. The collection of antikeys for a minimal key system $\mathcal{K}$ is usually denoted by $\mathcal{K}^{-1}$. This is justified by the fact that minimal keys and maximal

antikeys determine each other, respectively [5]. In fact maximal antikeys are maximal sets that do not contain any key and keys are minimal sets that are not contained in any antikey. Clearly, both minimal key systems and maximal antikey systems form inclusion-free families of subsets of $\mathcal{R}$, that is no minimal key/antikey can contain another minimal key/antikey. For a set system $\mathcal{A}$ of subsets of $\mathcal{R}$ let $\mathcal{A}\downarrow = \{B \subseteq \mathcal{R}\colon \exists A \in \mathcal{A} \text{ with } B \subseteq A\} \cup \{\mathcal{R}\}$. Furthermore, let $\mathcal{A}_\cap = \{B \subseteq \mathcal{R}\colon \exists i \geq 1, A_1, A_2, \ldots A_i \in \mathcal{A} \text{ with } B = A_1 \cap A_2 \cap \ldots \cap A_i\} \cup \{\mathcal{R}\}$. That is, $\mathcal{A}\downarrow$ is the down-set generated by $\mathcal{A}$ appended with $\mathcal{R}$ and $\mathcal{A}^\cap$ is the set system closed under intersection generated by $\mathcal{A}$.

Theorem 3.1. *Let $\mathcal{K}$ be an inclusion-free family of subsets of $\mathcal{R}$. Then the closures whose minimal key system is $\mathcal{K}$ form an interval in the poset of closures $\mathbf{P}(\mathcal{R})$ whose smallest element is the closure with closed sets $\mathcal{K}_\cap^{-1}$ and largest element is the closure with closed sets $\mathcal{K}^{-1}\downarrow$.*

Proof. Let us suppose that $\mathbf{r}$ is an Armstrong-instance of $\mathcal{K}$ and let A be an antikey. For any $b \in \mathcal{R} \setminus A$, $A \cup \{b\}$ is a key, thus $A \cup \{b\} \to \mathcal{R}$ holds. If $A \to b$ held, then by the transitivity rule $A \to \mathcal{R}$ would hold, contradicting to the antikey property of A. Thus $\ell(\mathbf{r})(A) = A$ for every antikey $A \in \mathcal{K}^{-1}$. Since $\mathcal{F}(\ell(\mathbf{r}))$ is closed under intersection, $\mathcal{K}_\cap^{-1} \subseteq \mathcal{F}(\ell(\mathbf{r}))$ follows. On the other hand, if $\ell(\mathbf{r})(X) = X$ holds for some $X \subsetneqq \mathcal{R}$, then X cannot contain any key. Thus, there exists a containment-wise maximal set $A \supset X$, that still does not contain any key. This implies that A is an antikey, hence $X \in \mathcal{K}^{-1}\downarrow$. It is easy to check, that both $\mathcal{K}_\cap^{-1}$ and $\mathcal{K}^{-1}\downarrow$ are closed under intersection and contain both, $\emptyset$ and $\mathcal{R}$. □

Corollary 3.1. *The diameter, that is the largest distance between any two elements of the collection of closures with given key system $\mathcal{K}$ is $|\mathcal{K}^{-1}\downarrow| - |\mathcal{K}_\cap^{-1}|$.* □

Corollary 3.1 determined the diameter if S was the set of databases (closures) with a given set of minimal keys. But the result, in this generality cannot be more than algorithmic.

In what follows we try to give a more precise, numerical answer in several special cases.

3.1 Unique Minimal Key

In this subsection we treat the case when the database has a unique minimal key. This is indeed very frequent case in practice.

Theorem 3.2. *The diameter of the set of closures having exactly one minimal key A where $0 < |A| = r < n$ is $2^n - 2^r - 2^{n-r}$.*

Proof. Let $A \subseteq B \subsetneqq \mathcal{R}$. Then $A \to \mathcal{R}$ implies $B \to \mathcal{R}$, therefore B cannot be a closed set. The members F of a family of closed sets $\mathcal{F}$ with unique minimal key A satisfy either $A \not\subseteq F$ or $F = R$.

Let $a \in A$ be an attribute and $B = \mathcal{R} - \{a\}$. If B is not a closed set, then $B \to \mathcal{R}$, that is, B is a key, which does not contain A, hence there should exist another minimal key. This contradiction gives that $B = \mathcal{R} \setminus \{a\}$ is a closed set, $B = \mathcal{R} \setminus \{a\} \in \mathcal{F}$. Since $\mathcal{F}$ is closed under intersection, every set satisfying $F \supseteq \mathcal{R} \setminus A$ must be closed.

We can conclude that the family of closed sets $\mathcal{F}$ satisfies the following conditions.

$$\text{If } F \supseteq \mathcal{R} \setminus A \text{ then } F \in \mathcal{F}, \tag{6}$$

$$\text{if } F \supseteq A, F \neq R \text{ then } F \notin \mathcal{F}, \tag{7}$$

$$\emptyset \in \mathcal{F}. \tag{8}$$

$\mathcal{F}$ is also closed under intersection, but it is not needed for the proof of the upper estimation. On the other hand it is easy to see that if $\mathcal{F}$ satisfies these conditions then there is a closure in which the family of closed sets is $\mathcal{F}$ and A is the only minimal key.

We have to find the maximum size of the symmetric difference of two families $\mathcal{F}_1$ and $\mathcal{F}_2$ satisfying (6)–(8). The symmetric difference does not contain the sets under (6) (7) and (8). The number of subsets of $\mathcal{R}$ satisfying (6) is 2^r (they are in one-to-one correspondance with subsets of A). The number of subsets of $\mathcal{R}$ satisfying (7) is $2^{n-r} - 1$ (they are in one-to-one correspondance with subsets of $\mathcal{R} \setminus A$ except $\mathcal{R} \setminus A$ itself). Therefore we have

$$|\mathcal{F}_1 \,\Delta\, \mathcal{F}_2| \leq 2^n - 2^r - 2^{n-r} + 1 - 1 = 2^n - 2^r - 2^{n-r}. \tag{9}$$

Choose $\mathcal{F}_1$ containing all sets satisfying (6) (7) and (8), while let $\mathcal{F}_2$ have all the sets satisfying (6) and (8). It is easy to see that these families are closed under intersection and satisfy (9) with equality. □

Remark 3.1. The diameter will be the smallest if $r = 1$, then it becomes about half of the diameter of the space without restriction given by Proposition 2.3. This coincides with our expectation: the "smaller" key is "stronger" in the sense that it determines more, the diameter of the space of possible databases becomes the smallest.

Remark 3.2. It is interesting to compare this result with the case when it is only known that A is a minimal key, but we do not know if there are other keys or not. That means we only have (7) and (8) as restrictions. Then the diameter is, of course, larger: $2^n - 2^{n-r} + 1$.

3.2 Upper Bound for Non-uniform Minimal Key System

In this subsection we assume that $\mathcal{R} = \{1, 2, \ldots n\} = [n]$, for the sake of convenience. Let $\binom{[n]}{r}$ denote the collection of r-subsets of $[n]$. Let $\mathcal{M}$ be a non-empty, inclusion-free family. Define

$$\mathcal{D}(\mathcal{M}) = \{H : \exists M \in \mathcal{M} \text{ such that } H \subseteq M\}, \tag{10}$$

$$\mathcal{U}(\mathcal{M}) = \{H : \exists M \in \mathcal{M} \text{ such that } H \supseteq M\}. \tag{11}$$

The *characteristic vector* $v(A)$ of the set $A \subseteq [n]$ is a 0,1 vector in which the i-th coordinate is 1 iff $i \in A$. If $A \in \binom{[n]}{r}$ then $v(A)$ contains exactly r 1's. $b(A)$ is the integer obtained by reading $v(A)$ as a binary number. Now an ordering "$<$" is introduced among the elements of $\binom{[n]}{r}$. Let $A, B \in \binom{[n]}{r}$ then $A < B$ iff $b(A) < b(B)$. This ordering is called *lexicographic*.

Define the (r, ℓ)-*shadow* of a family of r-element sets $\mathcal{A} \subseteq \binom{[n]}{r}$ for $\ell < r$:

$$\sigma_{r,\ell}(\mathcal{A}) = \{H : |H| = \ell, \exists A \in \mathcal{A} \text{ such that } H \subset A\}. \tag{12}$$

Proposition 3.1 ([7,8]). *If $\mathcal{A}$ consists of some lexicographically first members of $\binom{[n]}{r}, \ell < r$ then $\sigma_{r,\ell}(\mathcal{A})$ is a family of some lexicographically first members of $\binom{[n]}{\ell}$.*

Theorem 3.3 (Shadow Theorem, [7,8]). *If $\mathcal{A} \subseteq \binom{[n]}{r}, |\mathcal{A}| = m$ then $|\sigma_{r,\ell}(\mathcal{A})|$ is at least as large as the (r, ℓ)-shadow of the family of the lexicographically first m members of $\binom{[n]}{r}$, that is, the size of the (r, ℓ)-shadow attains its minimum for the lexicographically first r-element sets.*

If $\mathcal{A} \subseteq 2^{[n]}$ is a family then $\mathcal{A}_r$ denotes the subfamily consisting of all r-element members:

$$\mathcal{A}_r = \mathcal{A} \cap \binom{[n]}{r}. \tag{13}$$

The *profile vector* of the family $\mathcal{A} \subseteq 2^{[n]}$ is $p = (p_0, p_1, \ldots p_n)$ where $p_r = p_r(\mathcal{A}) = |\mathcal{A}_r|$.

Lemma 3.1. *Let $\mathcal{M}$ be a non-empty inclusion-free family of subsets of $[n]$ with fixed $|\mathcal{M}| \geq n$. Then $|\mathcal{D}(\mathcal{M})|$ attains its minimum for a family satisfying the following conditions with some $2 \leq r \leq n$.*

$$p_n = \ldots = p_{r+1} = p_{r-2} = \ldots = p_1 = p_0 = 0, \tag{14}$$

$$\mathcal{M}_r \text{ consists of the lexicographically first } r - \text{element subsets}, \tag{15}$$

$$\mathcal{M}_{r-1} = \binom{[n]}{r-1} \setminus \sigma_{r,r-1}(\mathcal{M}_r). \tag{16}$$

We do not claim that this is the only optimal solution. On the other hand, if $|\mathcal{M}| \leq n$ then the best construction consists of $|\mathcal{M}|$ pieces of 1-element sets.

Proof. Suppose that $p_n = \ldots = p_{r+1} = 0, p_r > 0$. Consider $\mathcal{D}(\mathcal{M})_{r-1}$. Its size is at least $|\sigma_{r,r-1}(\mathcal{M}_r)|$ what is minimum, by the Shadow Theorem, if $\mathcal{M}_r$ consists of the lexicographically first r-element subsets. By the proposition, $\sigma_{r,r-1}(\mathcal{M}_r)$ is a family of some first $r-1$-element subsets. $\mathcal{M}_{r-1}$ is disjoint from $\sigma_{r,r-1}(\mathcal{M}_r)$,

since $\mathcal{M}$ is inclusion-free. $|\mathcal{D}(\mathcal{M})_{r-2}|$ is at least $|\sigma_{r-1,r-2}(\sigma_{r,r-1}(\mathcal{M}_r) \cup \mathcal{M}_{r-1})|$ by the Shadow Theorem with equality if $\sigma_{r,r-1}(\mathcal{M}_r) \cup \mathcal{M}_{r-1}$ is "lexicographically first", that is, if $\mathcal{M}_{r-1}$ is the "continuation" of $\sigma_{r,r-1}(\mathcal{M}_r)$ in the lexicographic ordering. Continuing in this way, we can see that $\mathcal{D}(\mathcal{M})_\ell$ will be minimum for a fixed profile for the following construction. Choose the lexicographically first p_r r-element sets, the lexicographically first $r-1$-element sets following $\sigma_{r,r-1}(\mathcal{M}_r)$, the lexicographically first $r-2$-element sets following $\sigma_{r-1,r-2}(\sigma_{r,r-1}(\mathcal{M}_r) \cup \mathcal{M}_{r-1})$, and so on. Since this construction does not depend on ℓ, this construction minimizes also $|\mathcal{D}(\mathcal{M})|$. Now, that we know what structure minimizes $|\mathcal{D}(\mathcal{M})|$ for a fixed profile of $\mathcal{M}$, we need to show that the best profile is when only two set sizes occure.

Suppose that

$$|\mathcal{D}(\mathcal{M})_s| < \binom{n}{s}, |\mathcal{D}(\mathcal{M})_{s-1}| = \binom{n}{s-1}, \ldots, |\mathcal{D}(\mathcal{M})_1| = \binom{n}{1}, |\mathcal{D}(\mathcal{M})_0| = 1 \tag{17}$$

holds for some integer $1 \leq s \leq r$.

If $s = r$, we are done, $\mathcal{M}$ has only r and $r-1$-element sets, ordered according to the statement of the Lemma. Otherwise suppose that $|\mathcal{D}(\mathcal{M})|$ is minimum for the given size $|\mathcal{M}|$ and $r-s$ is the smallest possible. Let A and B be the lexicographically last member of $\mathcal{M}_r$ and the lexicographically first non-member of $\mathcal{M}_s$. Replace A with B. All proper subsets of B are in $\mathcal{D}(\mathcal{M})$, therefore this operation cannot increase $|\mathcal{D}(\mathcal{M})|$. Repeat this step until either $\mathcal{M}_r$ becomes empty, or $\mathcal{M}_s$ "full": $|\mathcal{D}(\mathcal{M})_s| = \binom{n}{s}$. In both cases, the difference $r-s$ becomes smaller. This contradiction finishes the proof. □

Theorem 3.4. *Let $\mathcal{K}$ be a non-empty inclusion-free family of subsets of $[n]$, where $|\mathcal{K}| \geq n$ is fixed. Furthermore, let $S(\mathcal{K})$ denote the set of all closures in which the family of minimal keys is exactly $\mathcal{K}$. Then*

$$\mathsf{diam}(S(\mathcal{K})) \leq 2^n - |\mathcal{U}(\mathcal{K}^*)|, \tag{18}$$

where $\mathcal{K}^*$ consists of some lexicographically last sets of size s and all the $s+1$-element sets not containing the selected s-element ones, for some $0 \leq s \leq n-2$ and $|\mathcal{K}^*| = |\mathcal{K}|$.

Proof. It is obvious that the members of $\mathcal{U}(\mathcal{K}) \setminus \{[n]\}$ are not closed sets in a closure belonging to $S(\mathcal{K})$.

Define $X = \bigcap_{K \in \mathcal{K}} K$ and let x be an element of X. If $[n] \setminus \{x\}$ is not closed then it is a key. It must contain a minimal key as a subset. This contradiction shows that $[n] \setminus \{x\}$ is closed. The intersection of closed sets is closed [5], therefore all sets containing $[n] \setminus X$ are closed. Denote the family of these sets by $2^X + ([n] \setminus X)$. (This notation may sound peculiar, but reflects the idea, that any set containing $[n] \setminus X$ consists of the union of a subset of X and the set $[n] \setminus X$.)

Concluding, if $\mathcal{F}$ is a family of closed sets in a closure from $S(\mathcal{K})$, then the followings are true.

$$\mathcal{F} \cap (\mathcal{U}(\mathcal{K}) \setminus \{[n]\}) = \emptyset, \tag{19}$$

$$\mathcal{F} \supseteq 2^X + ([n] \setminus X). \tag{20}$$

Therefore, if $\mathcal{F}_1$ and $\mathcal{F}_2$ are two families of closed sets of two closures from $S(\mathcal{K})$, then $\mathcal{F}_1 \,\Delta\, \mathcal{F}_2$ cannot contain $\emptyset$, $[n]$ and members of $\mathcal{U}(\mathcal{K}) \setminus \{[n]\}$ and $2^X + ([n] \setminus X)$:

$$\mathcal{F}_1 \,\Delta\, \mathcal{F}_2 \subseteq 2^{[n]} \setminus \mathcal{U}(\mathcal{K}) \setminus (2^X + ([n] \setminus X)) \setminus \{\emptyset\}. \tag{21}$$

Hence we have, considering that the only common element of $\mathcal{U}(\mathcal{K})$ and $2^X + ([n] \setminus X)$ is $[n]$, that

$$|\mathcal{F}_1 \,\Delta\, \mathcal{F}_2| \leq 2^n - |\mathcal{U}(\mathcal{K})| - 2^{|X|}. \tag{22}$$

By Theorem 2.1 the left hand side is the distance of the two closures, (22) gives an upper bound on the diameter. Thus, to find a valid estimate, we need to minimize

$$|\mathcal{U}(\mathcal{K})| + 2^{|X|} \tag{23}$$

for fixed $|\mathcal{K}|$.

Let $\mathcal{K}^- = \{\overline{K} : K \in \mathcal{K}\}$, where $\overline{K}$ denotes the complement of K. It is easy to see that $\mathcal{U}(\mathcal{K})^- = \mathcal{D}(\mathcal{K}^-)$ and $X = [n] \setminus \bigcup_{M \in \mathcal{K}^-} M$. Therefore the minimum of (23) can be found minimizing

$$|\mathcal{D}(\mathcal{M})| + 2^{n - |\bigcup_{M \in \mathcal{M}} M|} \tag{24}$$

for fixed $|\mathcal{M}|$. If $\bigcup_{M \in \mathcal{M}} M \neq [n]$ then replace a member of $\mathcal{M}$ that is covered by the union of the other members,by a 1-element set in $[n] \setminus \bigcup_{M \in \mathcal{M}} M$. If no member of $\mathcal{M}$ is covered by other members, then each one has an "own" element that is not contained in any other member of $\mathcal{M}$. Since $|\mathcal{M}| \geq n$, that is impossible, so a covered member must exist. This operation obviously does not increase (24). Repeated application of this step shows that $\bigcup_{M \in \mathcal{M}} M = [n]$ can be supposed. Lemma3.1 determines the minimum of $\mathcal{D}(\mathcal{M})$, the substitution $\mathcal{M}^- = \mathcal{K}$ gives the construction showing the desired upper bound. □

Remark 3.3. If $|\mathcal{K}| = \binom{n}{s}$ then (18) becomes

$$diam(S(\mathcal{K})) \leq 2^n - \sum_{i=s}^{n} \binom{n}{i} = \sum_{i=0}^{s-1} \binom{n}{i}. \tag{25}$$

Indeed, $\binom{n}{s} = |\mathcal{K}| = |\mathcal{K}^*|$ by Theorem 3.4. There is only one possibility to have $\mathcal{K}^*$ of the structure given by Theorem 3.4 of this size, namely if $\mathcal{K}^* = \binom{[n]}{s}$.

3.3 Uniform Minimal Key Systems

If all keys are one-element sets, then $\mathcal{K}^{-1}$ consists of a single set A, thus $\mathcal{K}^{-1}{\downarrow}$ consists of all subsets of A and $\mathcal{R}$, while $\mathcal{K}_\cap^{-1}$ consists of two sets, A and $\mathcal{R}$, i.e. the diameter is $2^{|A|} - 1$. Hence, we start with the special case when the minimal keys have size 2.

Let D be a closure whose minimal keys have exactly two elements. $G = ([n], E)$ be the graph where $[n] = \{1, 2, \ldots, n\}$ stands for the set of attributes of D and $\{i, j\} \in E (i \neq j)$ is an edge of the graph iff $\{i, j\}$ is not a minimal key in D. That is, $\mathcal{K}$ is equal to $\binom{[n]}{2} - E$. Let $|E| = e$. The set of closures having $\binom{[n]}{2} - E$ as the set of minimal keys will be denoted by $S_2(G)$. We want to give an upper estimate on $\mathsf{diam} S_2(G)$ depending only on e, that is, actually we give upper bound for

$$s_2(e) = \max_{\{G=([n],E):\ |E|=e\}} \mathsf{diam} S_2(G). \tag{26}$$

First we consider the case when G has one non-trivial connected component.

Theorem 3.5. *If* $e = \binom{t}{2} + r$, *where* $0 < r \leq t$, *then*

$$\mathsf{diam} S_2(G) \leq \begin{cases} 2^t + 2^r - 4 & \text{if } r < t \\ 2^{t+1} - 2 & \text{if } r = t \end{cases} \tag{27}$$

for a graph G *whose connected components are isolated vertices except for one component. Furthermore, this bound is sharp.*

Proof. Let $D \in S_2(G)$ where $G = ([n], E)$ is a graph with $|E| = e$ edges. Since the family of minimal keys is $\mathcal{K} = \binom{[n]}{2} - E$, the family of maximal antikeys $\mathcal{K}^{-1}$ consists of sets containing no key (that is, no edge in $\binom{[n]}{2} - E$ and maximal for this property). Then the members of $\mathcal{K}^{-1}$ are maximal complete subgraphs in G. These are called the *cliques* of G. $\mathcal{K}^{-1}\downarrow$ consists of all complete subgraphs of G, while $\mathcal{K}_{\cap}^{-1}$ consists of those complete subgraphs that are intersections of cliques. We will show that

$$|\mathcal{K}^{-1}\downarrow| \leq 2^t + 2^r + n - t - 1. \tag{28}$$

We apply the following theorem of Erdős [6].

Theorem 3.6 (Erdős, 1962). *Let* $G = (V, E)$ *be a connected graph of* e *edges. Assume, that* $e = \binom{t}{2} + r$, *where* $0 < r \leq t$. *Then the number of complete* k*-subgraphs* $C_k(G)$ *of* G *is at most*

$$C_k(G) \leq \binom{t}{k} + \binom{r}{k-1}. \tag{29}$$

This estimate is sharp.

Note that Theorem 3.6 is valid for all k, since if $k > \max(t, r+1)$, then both binomial coefficients are 0, and no complete subgraph of that size could exist.

Since $\mathcal{K}^{-1}\downarrow$ consists of all complete subgraphs of G, so we just have to sum up (29) for all k in the non-trivial component of G, and add the number of isolated vertices. That is

$$|\mathcal{K}^{-1}\downarrow| = \sum_{k \geq 0} \binom{t}{k} + \sum_{k \geq 0} \binom{r}{k-1} + n - t - 1, \tag{30}$$

that results in (28). The optimum construction takes a complete graph on t vertices and add an extra vertex connected to r vertices of the complete graph. If $r < t$, then the nontrivial component of G cannot be a complete graph, so there are at least two maximal cliques. Then $\emptyset$, the two maximal cliques, and their intersection is in $\mathcal{K}_\cap^{-1}$. Furthermore, the isolated vertices are maximal cliques themselves, so they are contained in $\mathcal{K}_\cap^{-1}$, that is $|\mathcal{K}_\cap^{-1}| \geq 4+n-t-1$. Applying (28) and Corollary3.1 the upper bound in (27) follows. □

If we have more than one non-trivial component of G, then Erdős' theorem does not apply. In fact, for small e, we can have better construction. For example, if $e = 2$, then the best graph consists of two independent edges. In the Appendix we give a proof of a general upper bound. With the notations of Theorem 3.5 it says that $s_2(e) \leq 2^{t+1} - 2$.

We can have some upper bound in the case of r-uniform minimal key system. Let D be a closure whose minimal keys have exactly $r(\geq 2)$ elements. $H = ([n], \mathcal{E})$ be the hypergraph where $[n] = \{1, 2, \ldots, n\}$ stands for the set of attributes of D and the r-element set $R \in \binom{[n]}{r}$ is a hyperedge of the hypergraph H, that is a member of the family $\mathcal{E}$ iff R is not a minimal key in D. That is, $\mathcal{K}$ is equal to $\binom{[n]}{r} \setminus \mathcal{E}$. We also suppose that $|\mathcal{E}| = e$. The set of closures having $\binom{[n]}{r} \setminus \mathcal{E}$ as the set of minimal keys will be denoted by $S_r(H)$. We want to give an upper estimate on $\mathsf{diam}S_r(H)$ depending only on e, that is, actually we will give an upper estimate on

$$\max_{\{H=([n],\mathcal{E}):\ |\mathcal{E}|=e\}} \mathsf{diam}S_r(H). \tag{31}$$

Theorem 3.7. *If* $e \leq \binom{a}{r}$ *then* $\mathsf{diam}(S_r(H)) \leq 2^a + e2^r$.

Proof. Let $D \in S_r(H)$ where $H = ([n], \mathcal{E})$ is a graph with $|\mathcal{E}| = e$ hyperedges. Since the family of minimal keys is $\mathcal{K} = \binom{[n]}{r} \setminus \mathcal{E}$, the family of antikeys $\mathcal{K}^{-1}$ consists of sets containing no key (that is, no edge in $\binom{[n]}{r} \setminus \mathcal{E}$ and maximal for this property). Then the members of $\mathcal{K}^{-1}$ are vertex sets of maximal complete subhypergraphs in H. That is sets $B \subset [n]$ such that $\binom{B}{r} \subset \mathcal{E}$ but for all $B' \supsetneq B$ $\binom{B'}{r} \setminus \mathcal{E} \neq \emptyset$. These are called the *(hyper)cliques* of H.

We have to prove that the number of sets of the vertices of H which are subsets of at least one hyperclique and are not intersections of those is at most $2^a + e2^r$. That is, $|\mathcal{K}^{-1}\downarrow| - |\mathcal{K}_\cap^{-1}| \leq 2^a + e2^r$. We will actually prove something stronger, namely we will show that $|\mathcal{K}^{-1}\downarrow| \leq 2^a + e2^r$.

Suppose first $0 < i \leq r$ and consider the number of i-element subsets of the hypercliques. Such a set must be a subset of an r-element set which is either $\in \mathcal{E}$ or is a subset of a larger maximal non-key. Therefore, in the worst case their number is at most $e\binom{r}{i}$.

Suppose now $r < i$. Let $A_1, \ldots, A_m$ be the family of i-element subsets, whose all r-element subsets are in $\mathcal{E}$ (They are not necessarily cliques!) If $m > \binom{a}{i}$ then by the Shadow Theorem the number of r-element subsets (hyperedges) is $> \binom{a}{r} \geq e$. Indeed, the we are considering here (i, r)-shadows, whixh is minimalized by

the lexicographically first m i-sets. If $m > \binom{a}{i}$ then these lexicographically first sets contain all i-subsets of $\{1, 2, \ldots, a\}$, so their (i, r)-shadows contain all r subsets of $\{1, 2, \ldots, a\}$. By the strict inequality $m > \binom{a}{i}$, some i-sets containing $a + 1$ are also in the lexicographically first m, so some r-sets containing $a + 1$ are in the shados. This contradiction shows that $m \leq \binom{a}{i}$.

Add up these maximums:

$$e \sum_{i=1}^{r} \binom{r}{i} + \sum_{i=r+1}^{a} \binom{a}{i} \leq 2^a + e2^r. \tag{32}$$

□

4 Conclusions, Further Research

In the present paper we have introduced a distance concept of databases. It is data-mining based, that is we start with the collection of functional dependencies that a given instance $\mathbf{r}$ of schema $\mathcal{R}$ satisfies. Two databases are considered to be the same, if their numbers of attributes agree and they satisfy exactly the same collection of functional dependencies. It has turned out that this concept fits nicely with the poset of closures as a model of changing databases, which was introduced sometimes ago.

Our research concentrates on how much different two databases are. On the other hand, Müller et. al. discussed distance of databases from the point of view how much work one needs to do in order to synchronize them. Their approach is algorithmic, our approach is more theoretical.

We have done the first steps by determining the largest possible distance between two databases of the same number of attributes. Next, we determined the distance of any two databases by showing that it is the size of the symmetric difference of the collections of closed sets. Then we investigated the diameter of the set of databases with a given system of (minimal) keys. This led to interesting discrete mathematics problems. Namely, given a hypergraph $H = (V, \mathcal{E})$, what is the number of complete subhypergraphs that are *not* intersections of *maximal subhypergraphs*? We have given good upper bounds in the case of ordinary graphs and k-uniform hypergraphs, when the number of hyperedges is fixed. Further research topic is to improve these bounds and find those key systems that achieve the extremal values. We in fact *conjecture* that if minimal keys are 2-element sets, and the number of keys is of form $\binom{n}{2} - \binom{a}{2}$, then the maximum diameter is $2^a - 2$ and it is attained when the two-element sets that are non-keys form a complete graph on a vertices.

Of course, our distance concept is restricted in the sense that it takes into account only the system of functional dependencies. This model can basically be extended in two directions. On the one hand it could be refined in the way distinguishing two databases when their system of functional dependencies are identical, but they are different in some other sense, for instance some other dependencies are satisfied in one of them while they are not satisfied in the other

one. The other direction of extension is the "generalization". Then the distance is introduced for databases with different numbers of attributes, as well.

This can be imagined in the following way. The "space" of all possible databases is given with a hypothetical distance in this space. In our model we forgot about the distance between two databases with the same system of functional dependencies (and, of course the same number of attributes) considering them the same and setting their distance 0, that is they are considered to be one "point" in the space. We introduce the distance between the databases within the set of the databases with the same number of attributes. We do not define here the distance between databases with different numbers of attributes, that is databases in distinct sets. Then the "real" or "hypothetical" distance could be a combination of the distance between databases in different sets and the distance between "refined elements" within one "point".

On the other hand, one may try to find a distance concept that takes into account that a database can have several different statuses during its lifetime. A first and very strict identity definition could be to declare instances $\mathbf{r}$ and $\mathbf{r}'$ of schemata $\mathcal{R}$ and $\mathcal{R}'$, respectively, being the same if there exist one-to-one mappings Φ: $\mathbf{Attr} \to \mathbf{Attr}$ and Ψ: $\mathbf{Dom} \to \mathbf{Dom}$ such that

$$\Phi(\mathcal{R}) = \mathcal{R}' \text{ and } \Psi(\mathbf{r}) = \mathbf{r}'. \tag{33}$$

However, this does not allow adding or deleting records, or modifying the schema. These latter two could be incorporated by saying that $\mathbf{r}$ and $\mathbf{r}'$ of schemata $\mathcal{R}$ and $\mathcal{R}'$ are instances of the same database if instead of (33), only

$$\Phi(\mathcal{R}) \subseteq \mathcal{R}' \text{ and } \Psi(\mathbf{r}) \subseteq \mathbf{r}' \tag{34}$$

or

$$\Phi(\mathcal{R}) \subseteq \mathcal{R}' \text{ and } \Psi(\mathbf{r}) \supseteq \mathbf{r}' \tag{35}$$

is required. However, while (33) is overly restrictive, (34) and (35) are too loose. They would allow the empty database to be the same with any other database. Thus, another future research topic is finding the proper balance between (33), and (34) and (35).

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research 28(1), 235–242 (2000)
2. Bhat, T.N., et al.: The PDB data uniformity project. Nucleic Acid Research 29(1), 214–218 (2001)
3. Boutselakis, H., et al.: E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. Nucleic Acid Research 31(1), 458–462 (2003)
4. Burosch, G., Demetrovics, J., Katona, G.O.H.: The Poset of Closures as a Model of Changing Databases. Order 4, 127–142 (1987)
5. Demetrovics, J., Katona, G.O.H.: Extremal combinatorial problems in relational data base. In: Gecseg, F. (ed.) FCT 1981. LNCS, vol. 117, pp. 110–119. Springer, Heidelberg (1981)

6. Erdős, P.: On the number of complete subgraphs contained in certain graphs. Publ. Math. Inst. Hung. Acad. Sci. VII, Ser. A3, 459–464 (1962), http://www.math-inst.hu/~p_erdos/1962-14.pdf
7. Katona, G.: A theorem on finite sets. In: Theory of Graphs, Proc. Coll. held at Tihany, 1966, Akadémiai Kiadó, pp. 187–207 (1968)
8. Kruskal, J.B.: The number of simplices in a complex. In: Mathematical Optimization Techniques, pp. 251–278. University of California Press, Berkeley (1963)
9. Müller, H., Freytag, J.-C., Leser, U.: On the Distance of Databases, Technical Report, HUB-IB-199 (March 2006)
10. Müller, H., Freytag, J.-C., Leser, U.: Describing differences between databases. In: CIKM 2006: Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, pp. 612–621 (2006)
11. Rother, K., Müller, H., Trissl, S., Koch, I., Steinke, T., Preissner, R., Frömmel, C., Leser, U.: COLUMBA: Multidimensional Data Integration of Protein Annotations. In: Rahm, E. (ed.) DILS 2004. LNCS (LNBI), vol. 2994. Springer, Heidelberg (2004)

A Appendix

Here we give a general upper bound for $s_2(e)$, through a series of lemmata. We suppose that $G = ([n], E)$ is a graph with e edges, where $e \leq \binom{a}{2}$. The number of subsets of $[n]$ which span a complete graph, not equal to $\emptyset$ and a clique is $c(G)$. Our final aim is to prove $c(G) \leq 2^a - 2$.

Theorem A.1. *If $9 < a$ and the number of edges e in G is at most $\binom{a}{2}$ then the number of subsets spanning a complete graph in G, not counting the empty set and the cliques is at most $2^a - 2$, that is $s_2(e) \leq 2^a - 2$.*

Remark A.1. If $e = \binom{a}{2}$ then the complete graph shows that our estimate is sharp. The statement of the theorem is also true when $a \leq 9$. This can be shown with ugly case analysis.

Remark A.2. With the notations of Theorem 3.5, $a = t + 1$.

Lemma A.1. *If G contains a K_{a-1} then $c(G) \leq 2^a - 2$.*

Proof. K_{a-1} contains $\binom{a-1}{2}$ edges, at most $a - 1$ are left. Denote the vertex set of K_{a-1} by A.

Suppose first that the graph M determined by the edges not in the K_{a-1} is connected and is not a tree. Then the number of vertices covered by them, that is, the number of vertices of M is at most $a-1$. The number of subsets of the vertices of M is at most 2^{a-1}. Since a the vertex set of a complete subgraph in G is either a subset of A or of the vertex set of M, we obtained $c(G) \leq 2^{a-1}-2+2^{a-1} = 2^a-2$.

If M is connected, but is a tree, the previous consideration does not work only when the number of edges of M is $a - 1$. Denote the set of vertices of M not in A by B, the subgraph of M spanned by B will be denoted by M_B. If $|B| = 1$ then G is a complete graph on a vertices, the statement is obvious. Suppose that M_B is connected and $|B| = b \geq 2$. Denote the set of vertices in

A adjacent to $i \in B$ by A_i. These sets $A_i (1 \leq i \leq b)$ are disjoint, because M is a tree. The number of edges of M_B is $b-1$. The sum of the sizes of A_i is $a-1-(b-1) = a-b$. The complete subgraphs not in A are either 2-element subsets of B (edges of M_B) or a vertex $i \in B$ plus a subset of A_i. Their total number is $\sum_{i=1} 2^{|A_i|} + b - 1 \leq 2^{a-b} + b - 1 + b - 1$. The maximum of the right hand side in the interval $1 \leq b \leq a-1$ is 2^{a-1}, this case is also settled.

Suppose now that M_B consists of $k \geq 2$ components: $N_1, \ldots, N_k$. The number of edges of N_i is denoted by f_i. On the other hand, let the number of edges having at least one end in N_i be $a_i - 1$. Here $\sum_{i=1}^{k} (a_i - 1) = a - 1$. Every new complete graph must contain one vertex from B but cannot contain vertices from distinct components N_i. Consider those new complete graphs containing at least one vertex from N_i. The statement of the previous paragraph can be repeated replacing $a-1$ by $a_i - 1$ and b by f_i. Therefore the number of these complete graphs is at most $2^{a_1 - 1}$. The total number of complete graphs is at most

$$\sum_{i=1}^{k} 2^{a_i - 1}. \tag{36}$$

Here $a_i = 1$ is impossible because then N_i would be an isolated point. Then 2^{a-1} is an upper estimate on (36), like in the previous cases.

Suppose now that M consists of $k \geq 2$ components: $M_1, \ldots, M_k$. The number of edges of M_i is denoted by m_i. Here $\sum_{i=1}^{k} m_i \leq a-1$. The result of the previous sections can be used: the number of new complete graphs in the ith component is at most 2^{m_i}. Altogether: $\sum_{i=1}^{k} 2^{m_i}$. Since every m_i is at least 1, therefore they cannot exceed $a-k$. This case can be finished exactly like the previous one. □

The *difference* of two complete graphs is the $|V_1 - V_2|$ where V_1 and V_2 are the two vertex sets.

Lemma A.2. *Suppose that G contains no K_{a-1} but it contains (at least) two K_{a-2} with the vertex sets V_1 and V_2, respectively. Then one of the followings holds:*

$$|V_1 - V_2| = 1, \tag{37}$$

$$|V_1 - V_2| = 2, \tag{38}$$

$$|V_1 - V_2| \geq 3 \text{ and } a \leq 9. \tag{39}$$

Proof. Let $|V_1 - V_2|$ be denoted by i. Then the total number of edges in the two K_{a-2} is

$$2\binom{a-2}{2} - \binom{a-2-i}{2} \leq \binom{a}{2}. \tag{40}$$

Easy algebra leads to the inequality

$$0 \leq i^2 - i(2a-5) + 4a - 6. \tag{41}$$

This holds iff i is not in the interval determined by the solutions of the quadratic equation:

$$\frac{2a-5\pm\sqrt{(2a-5)^2-16a+24}}{2}=\frac{2a-5\pm\sqrt{4a^2-36a+49}}{2}. \tag{42}$$

Here $2a-11<\sqrt{4a^2-36a+49}$ holds when $9<a$. Using this inequality we obtain strict upper and a lower estimate on the "smaller" and the "larger" roots, respectively:

$$3=\frac{2a-5-(2a-11)}{2},\quad 2a-8=\frac{2a-5+(2a-11)}{2}. \tag{43}$$

This proves that $9<a$ implies $i<3$. (The other estimate becomes $a+1<i$ when $9<a$ what is impossible.) □

Lemma A.3. *Suppose that $5\le a$ and G contains no copy of K_{a-1}, but contains a pair of K_{a-2}'s with difference 2. Then $c(G)<2^a-2$.*

Proof. It is easy to see that there might be at most 4 extra edges which can be added to the two K_{a-2}'s. The number of subsets spanning complete graphs in them, not counting themselves and the empty set, is $2^{a-2}-3+32^{a-4}$.

First suppose that there is no edge between V_1-V_2 and V_2-V_1, where V_1 and V_2 are the vertex sets of the K_{a-2}'s. Then all of the 4 new edges have one vertex not in $V_1\cup V_2$. It is easy to see by case analysis, that the maximum number of sets spanning a complete graph is 16, using the extra 4 edges. The total number of sets in question is at most $2^{a-2}-3+32^{a-4}+16\le 2^a-2$ when, say $5\le a$.

Suppose now that there is exactly one edge between V_1-V_2 and V_2-V_1. This addition creates a new K_{a-2}. It also adds $2^{a-4}-1$ new complete graphs. the remaining 3 edges may add at most 8. Altogether: $2^{a-2}-3+32^{a-4}+2^{a-4}-1+8=2^{a-1}+4$ what cannot be more than 2^a-2 when $4\le a$.

If there are two adjacent edges between V_1-V_2 and V_2-V_1 then they form a K_{a-1} contradicting our assumptions. Therefore if we suppose that there are at least two edges between V_1-V_2 and V_2-V_1 then it is possible only when there are exactly two of them and they have no common vertex. The so obtained graph contains $2^{a-1}+2^{a-4}-5$ proper complete graphs. The remaining two edges may add 4 more complete graphs: $2^{a-1}+2^{a-4}-1\le 2^a-2$. □

Lemma A.4. *Suppose that $4\le a$, G contains no copy of K_{a-1}, contains no pair of K_{a-2}s with difference 2, but contains 3 copies of K_{a-2} with pairwise difference 1. Then $c(G)\le 2^a-2$.*

Proof. Take two of the complete graphs. Let K denote the intersection of their vertex sets. ($|K|=a-3\ge 2$.) The vertex sets have the respective forms $K\cup u$ and $K\cup v (u\ne v)$. It is easy to see that the vertex set of the third K_{a-2} cannot contain either of u and v, therefore it also has the form $K\cup w$ where $(w\ne u),(w\ne v)$. Denote the graph obtained as their union by $K(3)$.

The number of edges of $K(3)$ is $\binom{a-3}{2}+3(a-3)=\binom{a}{2}-3$, that is, only 3 edges remained.

The number of proper complete subgraphs is $2^{a-3}+32^{a-3}-4=2^{a-1}-4$. The addition of 3 edges may create at most 8 complete graphs. $2^{a-1}+4\leq 2^a-2$ holds. □

Proof (of Theorem A.1). Start like in the proof of Theorem 3.5 The number of 1-element sets in question is at most $2e$, the number of 2-element subsets is e. The number of i-element subsets (≥ 3) is at most $\binom{a}{i}$. The only novelty here is the treatment of the large sets. We saw in Lemmata A.1-A.4 that the statement of the theorem holds when there is a K_{a-1} or at least three K_{a-2} in G. If we suppose the contrary then the number of $a-2$ and $a-1$-element sets in question is 0, the number of such $a-3$-element subsets is at most $2(a-2)$ (subsets of the two possible K_{a-2}). The total number of sets is at most

$$3e+\sum_{i=3}^{a-4}\binom{a}{i}+2(a-2)=3e+2^a-1-a-\binom{a}{2}-\binom{a}{a-3}+2(a-2)-$$
$$-\binom{a}{a-2}-\binom{a}{a-1}-1$$
$$=2^a-6+\left(2e-2\binom{a}{2}\right)+\left(e-\binom{a}{3}\right). \qquad (44)$$

Here $e\leq\binom{a}{2}$ by definition, $e\leq\binom{a}{3}$ is an easy consequence when $5\leq a$. (44) can be upper estimated by 2^a-6. □