Extremal Combinatorial Problems in Relational Databases

by

Yuriy Hulovatyy

Submitted to

Central European University

Department of Mathematics and Its Applications

In partial fulfillment of the requirements for the degree of
Master of Science

Supervisor: Dr. Gyula O. H. Katona

Budapest, Hungary

2012

Contents

Chapter 1

Introduction

Data is one of the most important assets in the modern world. Consequently, there is a practical need for a robust, secure, and effective way to store and manage it. This can be achieved by using a solid and reliable organized collection of data called a *database*.

The way data is organized in a database has a profound effect on how easily and effectively it can be managed. It is desirable to make it simple and intuitive, but at the same time powerful and flexible. There are different methods of describing data in databases that have been presented since their appearance in the late 1960s.

One of the most important data models is the *relational model* introduced by Codd [8] in 1970, which made a significant change in databases. Its mathematical foundation is based on set theory and predicate logic. In this model, we are presented with a view of data organized as tables. Each row of a table contains an instance of data for the categories specified by the columns. Thus, a database consists of a collection of tables containing data which can be manipulated by using a limited set of operations. A database that conforms to the relational model is called a *relational database*.

In contrast to other data models of its time, the relational model was much less complicated, more flexible, and independent of the physical storage methods. So, being simple and powerful, it soon became very popular. As a result, relational databases grew to be widely used. By the 1990s, they were the standard.

Relational databases have been an active area of research since the 1970s. In particular, they are interesting from the combinatorial point of view. There are many extremal combinatorial questions that arise from the relational model. For example, what is the maximum number of minimal keys a table with a fixed number of columns may have? Another interesting question is to find a matrix with the minimum number of rows that can represent a given closure operation.

In this thesis, we discuss some extremal combinatorial problems in the theory of relational databases. We explore many of the most important results in this area. Our main purpose is to organize these results and present them in a meaningful and coherent way. In addition to this, we introduce a connection between relational models and secret sharing together with some results.

The thesis is organized in the following way. In Chapter 2, we give a short introduction to the theory of relational databases. Chapter 3 presents closure operations and their matrix representation. Then, in Chapter 4, we show some important inequalities for database parameters. Chapter 5 is concerned with minimum matrix representation of closure operations. Finally, in Chapter 6, we discuss a connection between relational models and secret sharing.

Chapter 2

Relational databases

2.1 Relational model

The notion of a *data model*, which is a notation for describing data or information, is fundamental in the study of databases. The description consists of three parts: structure of the data, operations on the data, and constraints on the data. Usually, there is a limited set of operations that can be performed on the data that includes *queries*, that is, operations that retrieve information, and *modifications*, operations that change the database. This limitation makes possible to describe database operations at a very high level and at the same time have the database management system to implement them effectively.

The relational model is a data model of the utmost importance for databases. It was proposed by Codd [8] in 1970. The main idea of this model is to organize data into collections of two-dimensional tables called *relations*. Thus, a collection of relations is called a *database*. Properties of the objects that are kept in a table are called *attributes*. Each attribute has its *domain*, an elementary type associated with it. The columns of a table are named by the attributes. That is, a column contains information of the same type.

A *tuple* is an ordered set of attribute values. A tuple has one *component* for every attribute of the relation. Each component of a tuple must contain a value belonging to the domain of the associated attribute. So, a tuple represents some object or basic fact.

Tables provide a convenient, but at the same time powerful and flexible way to represent relations. A table has two parts: the *header* (the row containing column names) and the *body* (the rows containing information). Thus, the rows of a table other than the header row represent the tuples of the corresponding relation. The term "relation" refers to the

set of rows which contain information. The set of column names is called the *schema* of the relation. Consequently, the schema specifies how the information is stored in the relation, while the set of the tuples in the relation is the actual information stored in it.

So, a *relation* can be defined as a set of tuples with the same schema. All tuples in a relation have the same arity, since every tuple must have one component for each attribute of the relation. There should be no duplicated tuples, in other words, no two rows are allowed to have equal values in all columns. The order in which the tuples are presented is not important. Similarly, the attributes can also be ordered in different ways. However, the correspondence between the attributes and the columns of the table must be preserved to write the tuples properly, that is, to make each component's value correspond to the correct attribute.

The set of possible values for an attribute is determined by its domain. *Constraints* can be used to further restrict an attribute domain. They provide a way to achieve consistency and implement business rules in a database.

Keys are among the most important constraints that can be placed on relational schemas. A set of attributes of a relation forms a key if it is not allowed for two tuples to have equal values in all of these attributes at the same time. In other words, the tuples can be uniquely identified by their values for the key attributes. In general, there can be several different sets of attributes that can form a key.

Example 2.1. *Consider the following relation containing information about the schedule of courses:*

Attribute

	Id	*Title*	*Lecturer*	*Semester*	*Room*	*RoomCapacity*
Header						
Body	*1*	*Algebra*	*Smith*	*Fall'09*	*101*	*200*
	2	*Algebra*	*Smith*	*Fall'10*	*102*	*150*
	3	*Algebra*	*Jones*	*Fall'11*	*101*	*200*
	4	*Calculus*	*Adams*	*Fall'09*	*101*	*200*
	5	*Calculus*	*Thompson*	*Fall'11*	*102*	*150*
	6	*Complex analysis*	*Adams*	*Spring'10*	*103*	*100*

Tuple ⟵ (row 3)

Value (row 6, *Room*)

The schema of the relation is the following:

$$\{Id, Title, Lecturer, Semester, Room, RoomCapacity\}.$$

The header row contains attribute names. All other rows contain information about the schedule of courses. For example, the first row states that the Algebra course was held by Smith in Fall'09 semester in the room 101 with capacity 200. Similarly, according to the record with Id value 3, the course with the same title and in the same room was also presented in Fall'11 semester, but it was held by a different lecturer.

The column corresponding to the attribute "Lecturer" contains the names of all lecturers for the courses in the table. All attribute values should belong to the corresponding domains. For instance, the values in the "Room" column should be correct room numbers and the values in the "RoomCapacity" column should be nonnegative integers.

Clearly, "Id" is a key, since every record has a unique identification string. At the same time, "Title" is not a key, since two courses having the same title can be held in different semesters. On the other hand, $\{Title, Semester\}$ *is a key, because there can be at most one course with a specific title each semester.*

2.2 Basic notation

We now provide an alternative formulation of the definitions from the previous section, which is used in the following chapters.

One of the simplest ways to represent a relation is to use a matrix. That is, a relation can be considered as an $m \times n$ matrix M which columns correspond to the n *attributes* and rows correspond to the m *records* of the relation. Consequently, a row contains the data of a given object and a column contains the data of a particular type. For example, if the object corresponds to an individual, then its attributes may be the individual's name or date of birth. The entry in a specific row and column is called an *attribute value*. For our example, it may be a specific person's name or date of birth. We assume that the data of two distinct records is different. In other words, all rows of the matrix are distinct.

The set of possible entries in the ith attribute is called the *domain* of this attribute and is denoted by D_i. So, a record can be think of as an element of the direct product $D_1 \times D_2 \times \cdots \times D_n$. Such an element is called a *tuple*. Thus, the whole matrix M can represent the relation $R \subseteq D_1 \times D_2 \times \cdots \times D_n$. Even though all attributes are different, the domains may not be distinct. Since for our purposes the choice of sets D_i is not important, in general, we assume that every D_i is equal to the set of nonnegative integers. For simplicity, we also assume that all attributes are enumerated and refer to a specific attribute by using its number.

Let X denote the set of the columns of M. For $A \subseteq X$, we denote by $M(A)$ the submatrix of M determined by the columns in A. In this case, the data of records is restricted only to the attributes corresponding to the columns in the set A.

The following demonstrates an $m \times n$ matrix M representing a relation:

1	2	...	i	...	j	...	n-1	n	⟵X
a_{11}	a_{12}	...	a_{1i}	...	a_{1j}	...	a_{1n-1}	a_{1n}	
a_{21}	a_{22}	...	a_{2i}	...	a_{2j}	...	a_{2n-1}	a_{2n}	
⋮	⋮	⋱	⋮	...	⋮	⋱	⋮	⋮	
a_{l1}	a_{l2}	...	a_{li}	...	a_{lj}	...	a_{ln-1}	a_{ln}	⟵Tuple
⋮	⋮	⋱	⋮	...	⋮	⋱	⋮	⋮	
a_{m1}	a_{m2}	...	a_{mi}	...	a_{mj}	...	a_{mn-1}	a_{mn}	

$M(\{i, i+1, \ldots, j\})$

The relation has n attributes and m tuples. Clearly, $X = \{1, 2, \ldots, n\}$. All attribute values belong to the corresponding domains, so $a_{li} \in D_i$, $l \in \{1, 2, \ldots, m\}$, $i \in X$.

Definition 2.1. A set $K \subseteq X$ of columns is called a *key* if all rows of the submatrix $M(K)$ are different.

So, there are no two distinct rows of M which have equal attribute values in the columns belonging to a key K. To put it differently, attributes corresponding to a key K uniquely determine the records, that is, for any given values of these attributes there exists at most one record having these values.

Definition 2.2. A key $K \subseteq X$ is called minimal if there is no key K' such that $K' \subset K$.

Thus, no proper subset of a minimal key can be a key itself. We denote the family of all minimal keys by $\mathcal{K}$.

2.3 Functional dependencies

Relations are often used to model real world scenarios. For example, each record can represent an object or a relationship between objects. It may happen that even if some records have the right number of attributes and all attribute values are chosen from the right domains, they could not be in a relation due to some real world facts. For instance, in the relation from Example 2.1 there can be no two different records having the same values in the attributes “Title” and “Semester”. Even though the values are taken from

the right domains, it would contradict to the fact that there can be only one class with a specific title every semester.

There are two kinds of restrictions on relations: restrictions that depend on the semantics of domains and restrictions that depend on the equality or inequality of attribute values.

In the first case, the restrictions depend on the meaning of an attribute. For example, a person's date of birth must be a valid date and a person's height cannot be negative. Such incorrect values usually occur due to errors in entering or computing data. However, these restrictions are of little help in designing database schemas.

In the second case, the restrictions depend not on a particular attribute value, but only on whether two records agree in certain attributes. Consider the relation from Example 2.1. Clearly, if two rows have equal values in the attribute "Room", they must have equal values in the attribute "RoomCapacity" too, since otherwise it would contradict to the fact that each room has a specified capacity. These restrictions have the greatest importance in the design of database schemas.

Definition 2.3. Let $A, B \subseteq X$. We say that B *(functionally) depends* on A if there are no two rows equal in A but different in B. We denote this by $A \to B$. The pair (A, B) is called a *functional dependency*.

So, if $A \to B$, then any two rows in M having equal attribute values in A must also have equal attribute values in B. In other words, attribute values in A uniquely determine attribute values in B. Thus, in some sense, A can be think of as a key in $A \cup B$.

For brevity, in case $a \in X$ is a single column, we write $A \to \{a\}$ simply as $A \to a$.

Example 2.2. *For the relation from Example 2.1, the following functional dependencies hold:*

$$\{Id\} \to \{Id, Title, Lecturer, Semester, Room, RoomCapacity\},$$

since "Id" is a key;

$$\{Title, Semester\} \to \{Id, Title, Lecturer, Semester, Room, RoomCapacity\},$$

since $\{Title, Semester\}$ *is a key;*

$$\{Lecturer, Semester\} \to \{Title\},$$

if we suppose that every lecturer can have at most one course each semester;

$$\{Room\} \to \{RoomCapacity\},$$

since every room has a specified capacity.

However,

$$\{Title\} \not\rightarrow \{Lecturer\},$$

because different lecturers may have the same course. For instance, Algebra course was given by Smith in Fall'09 and Fall'10, but by Jones in Fall'11.

Definition 2.4. A functional dependency (A, B) is called *trivial* if $B \subseteq A$.

Definition 2.5. Let $A, B \subseteq X$. A functional dependency (A, B) is called *basic* if

1. $A \neq B$;
2. $\nexists A' \subset A$ such that $A' \rightarrow B$;
3. $\nexists B' \supset B$ such that $A \rightarrow B'$.

It is easy to see that in a special case of a functional dependency (A, B) when $B = X$, that is, the whole set X functionally depends on A, the set A is a key.

Definition 2.6. The set X with the sets D_i, $i \in \{1, 2, \ldots, n\}$ and the set of logical connections (like functional dependencies) is called the *relation schema*. The relation schema determines the set R^* of all possible tuples.

It is very important to design a good schema. If a schema is not designed properly, then it may lead to several problems, such as *redundancy* (the same information may be repeated in several places), *insertion anomalies* (some data cannot be inserted into the table), *update anomalies* (updates to the table may result in logical inconsistencies), or *deletion anomalies* (deletion of some data makes necessary the deletion of completely different data too). If, however, we know the complete set of functional dependencies for a given application, then these anomalies can be avoided. The methods of determining the degree of vulnerability of a database to anomalies and the ways to avoid them can be found in [18].

Example 2.3. *Consider the relation from Example 2.1. We demonstrate the problems in its design by showing the anomalies.*

The fact that room 101 has capacity 200 is repeated in several places, which means that there is redundancy. In addition, if a new room was built (for example, room 104), but it was not used for any course, we cannot add this information to the database. That is, there is an insertion anomaly. Furthermore, if a room 102 has been enlarged and its capacity is now 200, we will need to update both courses with id 3 and 5, since otherwise

there is a logical inconsistency. So, there is an update anomaly. Finally, if we delete a course with id 6, we will have no record about room 103, but it may still be available for other courses. Thus, there is a deletion anomaly.

Definition 2.7. The set R of actual tuples is called the *instance*.

The instance has to satisfy the conditions of the relation schema, so $R \subseteq R^*$. In general, however, the inclusion is proper.

Functional dependencies are statements about all possible values of a relation schema, not about a particular relation. We cannot deduce which functional dependencies hold for a relation schema from a specific relation. For instance, all functional dependencies hold for an empty relation, but they may not hold in general for a relation schema. It is possible, however, to deduce that some functional dependencies do not hold for a relation schema in general from a particular relation for this schema.

In order to determine functional dependencies, we have to consider the real world meaning of the attributes. Functional dependencies cannot be proved. They are statements about the real world.

2.4 Armstrong's axioms

It is important to make reasoning about functional dependencies. Suppose we know that a relation satisfies certain functional dependencies. Then, we can deduce that the relation has to satisfy other functional dependencies as well.

Definition 2.8. Let $\mathcal{F} = \{A_i \to B_i, i = 1, 2, \ldots, k\}$ be a set of functional dependencies for some relation schema and let $A \to B$ be a functional dependency. $\mathcal{F}$ *logically implies* $A \to B$ (denoted by $\mathcal{F} \models A \to B$) if every relation for the relation schema that satisfies functional dependencies in $\mathcal{F}$ also satisfies $A \to B$.

Definition 2.9. Let $\mathcal{F} = \{A_i \to B_i, i = 1, 2, \ldots, k\}$ be a set of functional dependencies for some relation schema. The *closure* of $\mathcal{F}$, denoted by $\mathcal{F}^+$, is the set of all functional dependencies that are logically implied by $\mathcal{F}$. That is,

$$\mathcal{F}^+ = \{A \to B : \mathcal{F} \models A \to B\}$$

In order to determine keys and understand implications among functional dependencies, it is essential to have inference rules telling how functional dependencies imply others. The rules, due to Armstrong [2], are the following:

Axiom 2.10 (Reflexivity). *If $B \subseteq A \subseteq X$, then $A \to B$.*

Axiom 2.11 (Augmentation). *If $A \to B$, then $A \cup C \to B \cup C \quad \forall C \subseteq X$.*

Axiom 2.12 (Transitivity). *If $A \to B$ and $B \to C$, then $A \to C$.*

These rules are called *Armstrong's axioms*. They are sound and complete. That is, they lead only to true conclusions and can be used to make any valid inference about functional dependencies. The proof of soundness and completeness of Armstrong's axioms can be found in [29].

Example 2.4. *Consider the relation from Example 2.1. We show by using Armstrong's axioms that the fact that $\{Title, Semester\}$ is a key and the functional dependency*

$$\{Lecturer, Semester\} \to \{Title\},$$

as in Example 2.2, imply that $\{Lecturer, Semester\}$ is a key.

By augmenting the given functional dependency with "Semester", we have that

$$\{Lecturer, Semester\} \to \{Title, Semester\}.$$

Then, from transitivity we get that the whole set of attributes functional depends on $\{Lecturer, Semester\}$, which means that it is a key.

Let $A, B, C, D \subseteq X$. By [29], we can derive additional rules from Armstrong's axioms:

Lemma 2.13 (Union rule). *If $A \to B$ and $A \to C$, then $A \to B \cup C$.*

Proof. We have that $A \to B$ and $A \to C$. By using augmentation, we get from $A \to C$ that $A \to A \cup C$. Similarly, by augmenting $A \to B$, we obtain that $A \cup C \to B \cup C$. Finally, by transitivity, $A \to A \cup C$ and $A \cup C \to B \cup C$ imply that $A \to B \cup C$. □

Lemma 2.14 (Decomposition rule). *If $A \to B \cup C$, then $A \to B$ and $A \to C$.*

Proof. We have that $A \to B \cup C$. By reflexivity, we obtain that $B \cup C \to B$ and $B \cup C \to C$. By applying transitivity, we get from $A \to B \cup C$ and $B \cup C \to B$ that $A \to B$ and from $A \to B \cup C$ and $B \cup C \to C$ that $A \to C$. □

Lemma 2.15 (Pseudotransitivity rule). *If $A \to B$ and $D \cup B \to C$, then $D \cup A \to C$.*

Proof. We have that $A \to B$ and $D \cup B \to C$. By augmenting $A \to B$, we get that $D \cup A \to D \cup B$. By applying transitivity, we get from $D \cup A \to D \cup B$ and $D \cup B \to C$ that $D \cup A \to C$. □

Chapter 3

Closure operations

3.1 Closures

Let M be an $m \times n$ matrix and let X be the set of its columns.

Definition 3.1. Let $A \subseteq X$. The *closure* of A is

$$\alpha_M(A) = \{a : a \in X, A \to a\}.$$

The following rules hold for $\alpha_M = \alpha$:

Lemma 3.2 ([15])**.** *Let $A, B \subseteq X$. Then*

$$A \subseteq \alpha(A), \tag{3.1}$$

$$A \subseteq B \implies \alpha(A) \subseteq \alpha(B), \tag{3.2}$$

$$\alpha(\alpha(A)) = \alpha(A). \tag{3.3}$$

Proof. The first property means that $A \to b \quad \forall b \in A$. Clearly, if two rows are equal in A, they must also be equal in every $b \in A$. Thus, (3.1) holds.

In order to show the second property, suppose that $a \in \alpha(A)$, which means that $A \to a$. Consequently, if two rows are equal in A, then they are equal in a too. From $A \subseteq B$ it follows that we can replace A by B in $A \to a$. Thus, $B \to a$, so we have that $a \in \alpha(B)$. Since it holds for any $a \in A$, we get (3.2).

Let us prove the third property. (3.1) implies that $\alpha(\alpha(A)) \supseteq \alpha(A)$, so we have only to show that $\alpha(\alpha(A)) \subseteq \alpha(A)$. Suppose that $a \in \alpha(\alpha(A))$. It means that any two rows

equal in $\alpha(A)$ are also equal in a. If two rows are equal in A, then, by the definition of $\alpha(A)$, they are equal in $\alpha(A)$ too. As a consequence, they must be also equal in a. So, $a \in \alpha(A)$. From the arbitrary choice of a we get that (3.3) holds.

The proof is complete. □

The next lemma easily follows from the definition of closure:

Lemma 3.3 ([15]). *Let $A, B \subseteq X$. Then $A \to B$ if and only if $B \subseteq \alpha(A)$.*

From Lemma 3.2 and Lemma 3.3 we can get the following properties of functional dependencies:

Lemma 3.4 ([15]). *Let $A, B, C, D \subseteq X$. Then*

$$\text{If } A \subseteq C,\ D \subseteq B,\ \text{and } A \to B,\ \text{then } C \to D \tag{3.4}$$

and

$$\text{If } A \to B \text{ and } C \to D,\ \text{then } A \cup C \to B \cup D \tag{3.5}$$

hold.

Proof. We start with (3.4). Lemma 3.3 implies that $B \subseteq \alpha(A)$. Since $D \subseteq B$, we have that $D \subseteq \alpha(A)$. By (3.2) and $A \subseteq C$, we get that $\alpha(A) \subseteq \alpha(C)$. So, $D \subseteq \alpha(C)$. It follows from Lemma 3.3 that $C \to D$. Thus, (3.4) holds.

Let us now prove (3.5). By Lemma 3.3, we have that $B \subseteq \alpha(A)$ and $D \subseteq \alpha(C)$. Consequently, $B \cup D \subseteq \alpha(A) \cup \alpha(C)$. From (3.2) we get that $\alpha(A) \subseteq \alpha(A \cup C)$ and $\alpha(C) \subseteq \alpha(A \cup C)$. Therefore, $B \cup D \subseteq \alpha(A) \cup \alpha(C) \subseteq \alpha(A \cup C)$, which means, by Lemma 3.3, that $A \cup C \to B \cup D$. So, (3.5) holds.

The lemma is proved. □

Definition 3.5. A function $\alpha : 2^X \to 2^X$ is called a *closure operation* if it satisfies (3.1)-(3.3).

In the beginning of this section, we defined the closure α_M assuming that we are given a matrix M. On the other hand, if α is an arbitrary closure operation on an n-element set X, then there exists an $m \times n$ matrix M such that $\alpha_M = \alpha$:

Theorem 3.6 ([2]). *For any given closure operation α there exists a matrix M such that*

$$\alpha_M = \alpha.$$

Definition 3.7. Let α be a closure operation. We say that a matrix M *represents* (*realizes*) α if $\alpha_M = \alpha$ holds.

Clearly, a matrix with a small number rows cannot have a complicated closure. By Theorem 3.6, closures and matrices are equivalent. So, the minimum number of rows in a matrix representing a closure operation can be used to measure the complexity of the closure operation.

Definition 3.8. Let α be a closure operation on X. Then let

$$s(\alpha) = \min\{m : M \text{ is a } m \times n \text{ matrix, } \alpha_M = \alpha\}.$$

In general, it is hard to determine $s(\alpha)$ for an arbitrary closure operation α. However, there are known some nice combinatorial results for specific classes of closure operations.

3.2 Keys and antikeys

It is easy to see that a closure operation determines a class of keys. Indeed, $K \subseteq X$ is a key in a matrix realizing α if $\alpha(K) = X$, that is, the closure of K is the whole set X. As before, the family of minimal keys is denoted by $\mathcal{K} = \mathcal{K}(\alpha)$. Clearly, $\mathcal{K}$ is a Sperner family (see Appendix A), since if $K_1, K_2 \in \mathcal{K}, K_1 \neq K_2$, then $K_1 \not\subset K_2$.

Definition 3.9. A matrix M *represents* (*realizes*) a given Sperner family $\mathcal{K}$ if $\mathcal{K} = \mathcal{K}(\alpha_M)$ holds.

An important question is whether any Sperner family can be the set of minimal keys of some relation. The following theorem provides an answer to this question:

Theorem 3.10 ([11])**.** *$\mathcal{K}$ is the set of minimal keys of some relation if and only if $\mathcal{K}$ is a nonempty Sperner family.*

Proof. We know that the set of minimal keys is a Sperner family. Thus, we have to prove only the other direction.

Let $\mathcal{I} = \{S_1, S_2, \ldots, S_m\} \subseteq 2^X$ be a Sperner family and let $m_i = |S_i| \quad \forall i = 1, 2, \ldots, m$.

We assume that

$$S_i = \{a_{i1}, a_{i2}, \ldots, a_{im_i}\} \quad \forall i = 1, 2, \ldots, m$$

with $a_{ij} \in X$ and $\bigcup_{i,j} a_{ij} = X$.

Let us construct a class of sets $\mathcal{M} = \{A_1, A_2, \ldots, A_t\}$ such that A_j $(j = 1, 2, \ldots, t)$ belongs to $\mathcal{M}$ if and only if

$$A_j \subseteq X \tag{3.6}$$

and

$$A_j \cap S_i \neq \emptyset \text{ for } i = 1, 2, \ldots, m. \tag{3.7}$$

We choose $\mathcal{F}$ as the set of the elements minimal in $\mathcal{M}$:

$$A_j \in \mathcal{F} \Leftrightarrow \nexists A_l \in \mathcal{M} \text{ such that } A_l \subset A_j. \tag{3.8}$$

From (3.6)-(3.8) we have that

$$\max\{m_1, m_2, \ldots, m_m\} \leq |\mathcal{F}| \leq m_1 \cdot m_2 \cdot \ldots m_m \tag{3.9}$$

and

$$A_j \in \mathcal{F} \text{ implies that } 1 \leq |A_j| \leq m. \tag{3.10}$$

Consider subsets $\mathcal{F}_k$ of $\mathcal{F}$ $(k = 1, 2, \ldots, n)$ such that

$$A_j \in \mathcal{F}_k \Leftrightarrow k \in A_j \in \mathcal{F}. \tag{3.11}$$

We claim that if the kth column of the relation is determined by the function f_k, identical with the class of sets $\mathcal{F}_k$, then the class of minimal keys is identical with the system $\mathcal{I}$. This is implied by the following three statements:

1. All of the sets S_i in $\mathcal{I}$ are keys.
2. No proper subset of S_i is a key.
3. There is no such a minimal key that is not in $\mathcal{I}$.

Let us prove the first statement. Since every $A_j \in \mathcal{F}$ is constructed so as to contain at least one element of every S_i, we have that

$$\bigcup_{k \in S_i} \mathcal{F}_k = \mathcal{F}. \tag{3.12}$$

It follows that each S_i contains a key K_i $(i = 1, 2, \ldots, m)$, so it is a key itself.

To prove the second statement, we show that if a key K_i is in S_i, then it equals S_i. For S_i there exists $A \in \mathcal{F}$ with $A \cap S_i = \{a\}$, since every S_j $(j = 1, 2, \ldots, m)$ contains either

a or some $a' \notin S_i$. Consequently,

$$\forall a \in S_i : \bigcup_{k \in A_i \setminus \{a\}} \mathcal{F}_k \subseteq \mathcal{F} \setminus \{A\} \text{ with } A \in \mathcal{F}, \tag{3.13}$$

which implies that no proper subset of S_i can be a key. So, every S_i is a minimal key.

Finally, let us prove the third statement. We have to show that there are no minimal keys in the relation beyond those in $\mathcal{I}$.

By contradiction, suppose that there exists a minimal key K such that it is not in $\mathcal{I}$. Clearly, $S_i \cap \bar{K} \neq \emptyset$ for $i = 1, 2, \ldots, m$. Thus, there exists at least one set $A \in \mathcal{F}$ such that it is not contained in any of the columns determined by K. That is,

$$\bigcup_{k \in K} \mathcal{F}_k \subseteq \mathcal{F} \setminus \{A\}. \tag{3.14}$$

Since K is a minimal key, this is a contradiction. So, there are no minimal keys that are not in $\mathcal{I}$.

The proof is complete. □

The following example from [11] provides an interpretation of the proof of Theorem 3.10:

Example 3.1. *Let* $n = 5$ *and let*

$$\mathcal{I} = \{S_1 = \{1, 2, 3\}, S_2 = \{3, 4, 5\}, S_3 = \{1, 3, 4\}\}.$$

Obviously, $\mathcal{I}$ *is a Sperner family.*

From (3.8) *we have that*

$$\mathcal{F} = \{A_1 = \{3\}, A_2 = \{1, 4\}, A_3 = \{1, 5\}, A_4 = \{2, 4\}\}.$$

We assign a different prime number to every set in $\mathcal{F}$*. So,* p_1 *corresponds to* A_1*,* p_2 *- to* A_2*,* p_3 *- to* A_3*, and* p_4 *- to* A_4*. We assume that* p_1*,* p_2*,* p_3*, and* p_4 *are in ascending order. Every set in* $\mathcal{F}_k$ *has an associated prime number too. We define the function* f_k *as the product of the prime numbers corresponding to the sets in* $\mathcal{F}_k$*. Since*

$$\mathcal{F}_1 = \{A_2, A_3\}, \quad \mathcal{F}_2 = \{A_4\}, \quad \mathcal{F}_3 = \{A_1\}, \quad \mathcal{F}_4 = \{A_2, A_4\}, \quad \mathcal{F}_5 = \{A_3\},$$

we have that

$$f_1 = p_2 \cdot p_3, \quad f_2 = p_4, \quad f_3 = p_1, \quad f_4 = p_2 \cdot p_4, \quad f_5 = p_3.$$

Let us show the rows corresponding to the following prime numbers:

1. $p_1 = 2, \quad p_2 = 3, \quad p_3 = 5, \quad p_4 = 7;$
2. $p_1 = 2, \quad p_2 = 5, \quad p_3 = 7, \quad p_4 = 11;$
3. $p_1 = 3, \quad p_2 = 5, \quad p_3 = 7, \quad p_4 = 11;$
4. $p_1 = 2, \quad p_2 = 5, \quad p_3 = 7, \quad p_4 = 13.$

The matrix is the following:

1	*2*	*3*	*4*	*5*
15	*7*	*2*	*21*	*5*
35	*11*	*2*	*55*	*7*
35	*11*	*3*	*55*	*7*
35	*13*	*2*	*65*	*7*

Every set from $\mathcal{I}$ is a key in the relation. In addition, for any other set which is not a superset of a member of $\mathcal{I}$ there exist at least two rows having the same values in the correspondent columns. Thus, $\mathcal{I}$ is the family of minimal keys of the relation.

Another class determined by a closure operation is the class of maximal non-keys:

Definition 3.11. The maximal non-keys are called *antikeys* and are denoted by

$$\mathcal{K}^{-1} = \{A : \nexists B \in \mathcal{K}, B \subseteq A, A \text{ is maximal with this property}\}.$$

Clearly, $\mathcal{K}^{-1}$ is a Sperner family, since its members are maximal with their property.

By [9], we can use antikeys to build a matrix having a given Sperner family $\mathcal{K}$ as the family of minimal keys. Let $\mathcal{K}^{-1} = \{G_1, G_2, ..., G_m\}$. The matrix has $n = |X|$ columns and $m + 1$ rows. The first row contains only zeros, while for $2 \leq i \leq m + 1$ the ith row contains in the column c either zero, if $c \in G_{i-1}$, or i otherwise.

Suppose that $A \in X$ does not contain any member of $\mathcal{K}$ as a subset. From the definition of $\mathcal{K}^{-1}$ it follows that there exists $i \in \{1, 2, \ldots, m\}$ such that $A \subseteq G_i$. Thus, the first and the $(i + 1)$th rows are equal in A. Consequently, A is not a key.

If, however, $A \supset K \in \mathcal{K}$, then $A \setminus G_i \neq \emptyset \quad \forall i \in \{1, 2, \ldots, m\}$. It implies that for $2 \leq i \leq m+1$ the ith row has the value i in at least one column from A. So, A is a key.

Therefore, the family of keys of the matrix is exactly the collection of all supersets of the members of $\mathcal{K}$. Thus, we have the matrix for which the family of its minimal keys is $\mathcal{K}$.

Example 3.2. *As in Example 3.1, let* $n = 5$ *and let*

$$\mathcal{K} = \{S_1 = \{1, 2, 3\}, S_2 = \{3, 4, 5\}, S_3 = \{1, 3, 4\}\}$$

be a Sperner family. We build a relation having $\mathcal{K}$ *as its family of minimal keys.*

Let us determine the collection of antikeys:

$$\mathcal{K}^{-1} = \{\{1, 2, 4, 5\}, \{1, 3, 5\}, \{2, 3, 4\}, \{2, 3, 5\}\}.$$

The matrix has 5 *columns and* 5 *rows. For example, the row corresponding to {1,3,5} has zeros in the* 1*st,* 3*rd and* 5*th columns, and* 3*'s in all other columns.*

So, the matrix is the following:

1	*2*	*3*	*4*	*5*
0	*0*	*0*	*0*	*0*
0	*0*	*2*	*0*	*0*
0	*3*	*0*	*3*	*0*
4	*0*	*0*	*0*	*4*
5	*0*	*0*	*5*	*0*

Every set from $\mathcal{K}$ *can uniquely determine a row. On the other hand, if a set is not a superset of a member of* $\mathcal{K}$*, then we have at least two rows having zeros in the corresponding positions, so the set is not a key. Thus,* $\mathcal{K}$ *is the family of minimal keys of the relation.*

Lemma 3.12 ([9])**.** *M represents a Sperner family* $\mathcal{K}$ *if and only if for any* $A \in \mathcal{K}^{-1}$ *it has two different rows having the same entries in A and any two rows that are equal in* $K \in \mathcal{K}$ *are equal everywhere.*

Proof. Clearly, if M represents $\mathcal{K}$, then $\mathcal{K} = \mathcal{K}(\alpha_M)$ holds. $A \in \mathcal{K}^{-1}$ implies that $\alpha_M(A) \neq X$, so we get the first condition. In the same way, $K \in \mathcal{K}$ implies that $\alpha_M(K) = X$. Thus, the second condition holds too.

Let us prove the other direction. If both conditions for M and $\mathcal{K}$ are satisfied, then

$$\alpha_M(A) \neq X \quad \forall A \in \mathcal{K}^{-1} \tag{3.15}$$

and

$$\alpha_M(K) = X \quad \forall K \in \mathcal{K} \tag{3.16}$$

hold.

(3.16) and (3.2) imply that $\alpha_M(C) = X$ if $C \supseteq K$ for some $K \in \mathcal{K}$. Assume that C is not a superset of a member of $\mathcal{K}$. In this case, by the definition of antikeys, there exists a set $A \in \mathcal{K}^{-1}$ such that $C \subseteq A$. Consequently, from (3.15) and (3.2) we have that $\alpha_M(C) \neq X$. In other words, $\alpha_M(C) = X$ holds only for the supersets of the members of $\mathcal{K} = \mathcal{K}(\alpha_M)$, which means that M represents $\mathcal{K}$.

The lemma is proved. □

The following definition is analogous to the Definition 3.8, but is defined for a Sperner family instead of a closure operation on an n-element set:

Definition 3.13.

$$s(\mathcal{K}) = \min\{m : M \text{ is a } m \times n \text{ matrix, } M \text{ represents } \mathcal{K}\}.$$

That is, $s(\mathcal{K})$ is the minimal number of records of a relation where the system of minimal keys is $\mathcal{K}$.

3.3 Uniform closure operations

We now introduce a special class of closure operations:

Definition 3.14. The *k-uniform closure operation* on an n-element ground set X is defined by

$$\alpha_k^n(A) = \begin{cases} X, & \text{if } |A| \geq k, \\ A, & \text{if } |A| < k. \end{cases}$$

It is easy to see that α_k^n satisfies (3.1)-(3.3), which implies that it is in fact a closure operation.

We denote the family of all k-element subsets of X by $\binom{X}{k}$. Clearly, if the family of minimal keys $\mathcal{K}$ is equal to $\binom{X}{k}$, then the family of antikeys $\mathcal{K}^{-1}$ is $\binom{X}{k-1}$, since they are the maximal sets not containing k-element subsets of X.

Lemma 3.15 ([9])**.** *Let α be a closure operation on X. Then*

$$\mathcal{K}(\alpha) = \binom{X}{k} \text{ if and only if } \alpha = \alpha_k^n.$$

Proof. If $\alpha = \alpha_k^n$, then $\mathcal{K}(\alpha) = \binom{X}{k}$ trivially follows from the definition of the k-uniform closure operation.

Let us prove the other direction. Suppose that for some $A \subseteq X$ with $|A| < k$ there exists an element $a \in \alpha(A) \setminus A$. Then, there is a set $B \subseteq X$ with $|B| = k$ and $B \supseteq A \cup \{a\}$.

From (3.1) it follows that $\alpha(B \setminus \{a\}) \supseteq B \setminus \{a\}$. (3.2) implies that $\alpha(B \setminus \{a\}) \supseteq \alpha(A)$. Since $a \in \alpha(A)$, we have that $\alpha(B \setminus \{a\}) \supseteq B$.

Thus, (3.2) and (3.3) imply that $\alpha(B\setminus\{a\}) = \alpha(\alpha(B\setminus\{a\})) \supseteq \alpha(B) = X$. Consequently, there exists a set $B \setminus \{a\}$ of size less than k, the closure of which is the whole X. This is a contradiction. It follows that such an $a \in \alpha(A) \setminus A$ cannot exist if $|A| < k$. Therefore, $\alpha(A) = A$.

$\mathcal{K}(\alpha) = \binom{X}{k}$ and (3.2) imply that $\alpha(A) = X$ for $|A| \geq k$.

So, $\alpha = \alpha_k^n$ holds.

The proof is complete. □

In general, there may be more than one closure operation having the same system of minimal keys. However, Lemma 3.15 shows that if the system of minimal keys is the family of all k-element subsets of X, then it uniquely determines the corresponding closure operation.

Chapter 4

Inequalities for database parameters

4.1 Maximum number of minimal keys

In relational databases, keys play a very important role, since they can be used to uniquely identify rows of a relation. The family of minimal keys contains all keys that do not have other keys as their subsets, that is, the keys that are minimal with their property. Thus, this family is of particular interest.

An interesting extremal combinatorial problem related to the family of minimal keys is to determine the maximum number of minimal keys a relation with a fixed number of attributes may have. The following theorem provides the maximum value for the number of members of this family:

Theorem 4.1 ([9])**.** *The maximum number of minimal keys in a relation with n attributes is*

$$\binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Proof. We know that minimal keys form a Sperner family. Therefore, by Sperner's Theorem (see Appendix A), minimal keys cannot contain more than $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ members. □

Let us construct an $m \times n$ matrix M having the maximum possible number of minimal keys. The matrix has n columns, $\binom{n}{\lfloor \frac{n}{2} \rfloor - 1} + 1$ rows, and $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ minimal keys.

The first row contains zeros only. The other rows have $\lfloor \frac{n}{2} \rfloor - 1$ zeros in all possible ways and the rest of the entries of the ith $(2 \leq i \leq \binom{n}{\lfloor \frac{n}{2} \rfloor - 1} + 1)$ row are i's.

For $n = 4$ the matrix is the following:

1	2	3	4
0	0	0	0
0	2	2	2
3	0	3	3
4	4	0	4
5	5	5	0

It is easy to see that if we choose any $\lfloor \frac{n}{2} \rfloor$ attributes, then there will be either all zeros or at least one number $i \neq 0$. In this case, we can uniquely determine a row, since all zeros means that it is the first row, and at least one number $i \neq 0$ implies that it is the ith row. So, any $\lfloor \frac{n}{2} \rfloor$ attributes form a key. If, however, we choose less than $\lfloor \frac{n}{2} \rfloor$ attributes, then they will not form a key, since the first row having all zeros will coincide with some other row in the corresponding submatrix of M. Thus, the family of minimal keys of the matrix is precisely the family of all $\binom{n}{\lfloor \frac{n}{2} \rfloor}$-element subsets of attributes. So, the matrix has n columns and $\binom{n}{\lfloor \frac{n}{2} \rfloor}$ minimal keys.

By using Theorem 4.1, we can make the upper estimate for the number of minimal keys in a relation with a given number of attributes. The following theorem improves this estimation for the bounded domains:

Theorem 4.2 ([28]). *Let $D_i \leq k \quad \forall i \in \{1, 2, \ldots, n\}$ and $k^4 < 2n+1$. Then the number of minimal keys of the relation is at most*

$$\binom{n}{\lfloor \frac{n}{2} \rfloor} - \lfloor \frac{n}{2} \rfloor.$$

In some cases, a subset of X cannot uniquely determine a too large set of attributes. That is, for any functional dependency $A \to B$ the inequality $|B \setminus A| \leq k$ must hold. Consequently, all keys must have at least $n - k$ attributes. Since the minimal keys form a Sperner family, we have to find the largest Sperner family that has members of size at least $n - k$. Thus, we have the following theorem providing the upper bound for the number of minimal keys in this case:

Theorem 4.3 ([16]). *Let all functional dependencies $A \to B$ of a relation satisfy the inequality $|B \setminus A| \leq k$, where $k \leq \frac{n}{2}$. Then the number of minimal keys of the relation cannot exceed*

$$\binom{n}{k}.$$

Proof. Since for $A \to X$ we need A to have at least $n-k$ elements, it follows that in the family of minimal keys $\mathcal{K}$ there are no members of sizes $0, 1, \ldots, n-k-1$. By using the inequality

$$\binom{n}{i} \leq \binom{n}{n-k}, k \leq \frac{n}{2}, n-k \leq i \leq n$$

together with the LYM inequality (see Appendix A), we get that

$$1 \geq \sum_{i=n-k}^{n} \frac{f_i}{\binom{n}{i}} \geq \sum_{i=n-k}^{n} \frac{f_i}{\binom{n}{n-k}} = \frac{\sum_{i=n-k}^{n} f_i}{\binom{n}{n-k}} = \frac{|\mathcal{K}|}{\binom{n}{n-k}},$$

where f_i denotes the number of sets in $\mathcal{K}$ having i elements.

So, $|\mathcal{K}| \leq \binom{n}{n-k} = \binom{n}{k}$.

The theorem is proved. □

4.2 Maximum number of basic functional dependencies

It is possible to deduce all functional dependencies from the basic functional dependencies. Consequently, their number can characterize the complexity of a relation.

Therefore, the problem of finding the maximum number of basic functional dependencies is interesting, since it corresponds to finding the most complex relation with a given number of attributes.

Note that the same problem for all functional dependencies is trivial, because the functional dependency $\emptyset \to X$ leads to an extremal case in which there are 2^{2n} functional dependencies.

Definition 4.4. $N(n)$ is the maximum number of basic functional dependencies in a relation with n attributes.

Since it is possible to construct a matrix having the basic functional dependencies in the form $A \to A \cup \{x\}$ for a fixed attribute x, we have that $2^{n-1} \leq N(n)$. So, we have a lower bound for $N(n)$. In fact, this value is exact for $n = 1, 2, 3, 4$.

However, the actual value of $N(n)$ can be greater than the trivial lower estimate, which is shown in the following theorem:

Theorem 4.5 ([3])**.**

$$2^n \cdot \left(1 - \frac{4}{\log_2 e} \cdot \frac{\log_2 \log_2 n}{\log_2 n} \cdot (1 + o(1))\right) \leq N(n) \leq 2^n \cdot (1 - \frac{1}{n+1}). \tag{4.1}$$

Proof. Let us start with the lower estimate.

Suppose that $q_1, q_2, \ldots, q_k$ are positive integers such that $\sum_{i=1}^{k} q_i = n$.

The matrix Q_i having $(\binom{q_i}{2} + 1)$ rows and q_i columns is defined in the following way. The first row of the matrix contains only zeros. The other rows contain $q_i - 2$ zeros in all possible ways, and the rest of the entries of the ith $(2 \leq i \leq \binom{q_i}{2} + 1)$ row are i's.

Now, let us build the matrix $M = Q_1 \times Q_2 \times \cdots \times Q_k$, the rows of which are all possible combinations of the rows of $Q_1, Q_2, \ldots, Q_k$. That is, we take an arbitrary row of Q_1, then an arbitrary row of Q_2, and so on up to Q_k, and then put them together. Since $\sum_{i=1}^{k} q_i = n$, M has n columns.

Let us show an example of M for $n = 5$, $k = 2$, and $q_1 = 3$, $q_2 = 2$ $(q_1 + q_2 = n)$. In this case, Q_1 is

1	2	3
0	0	0
0	2	2
3	0	3
4	4	0

and Q_2 is

1	2
0	0
2	2

Thus, M is the following:

1	2	3	4	5
0	0	0	0	0
0	0	0	2	2
0	2	2	0	0
0	2	2	2	2
3	0	3	0	0
3	0	3	2	2
4	4	0	0	0
4	4	0	2	2

Q_i^* denotes the set of column indices of matrix Q_i in M. In other words, $Q_i^* = \{q_1 + q_2 + \cdots + q_{i-1} + 1, \ldots, q_1 + q_2 + \cdots + q_i\}$.

The basic functional dependencies (A, B) in M are those $A, B \subseteq X$ that satisfy

$$\min_i\{q_i - |A \cap Q_i^*|\} = 1, \tag{4.2}$$

$$B = A \cup \bigcup_{|A\cap Q_i^*|=q_i-1} Q_i^*. \tag{4.3}$$

Indeed, if the left-hand side of (4.2) is greater than one, then it follows from (4.3) that the functional dependency (A, B) is trivial. On the other hand, if it is equal to zero, consider the set $A^* = A \setminus \{x\}$, where $x \in Q_j^*$ for Q_j^* such that $|Q_j^*| = |A \cap Q_j^*|$. Then (A^*, B) is a functional dependency, so the functional dependency (A, B) is not basic. But if (4.2) holds, then (A, B) is in fact a basic functional dependency.

The number of sets A satisfying (4.2) is equal to the difference between the number of sets satisfying

$$\min_i\{q_i - |A \cap Q_i^*|\} \geq 2$$

and the number of sets satisfying

$$\min_i\{q_i - |A \cap Q_i^*|\} \geq 1.$$

The first number is equal to $\prod_{i=1}^k (2^{q_i} - 1)$ and the second is equal to $\prod_{i=1}^k (2^{q_i} - q_i - 1)$. So, the number of basic functional dependencies in M is

$$\prod_{i=1}^k (2^{q_i} - 1) - \prod_{i=1}^k (2^{q_i} - q_i - 1). \tag{4.4}$$

This gives a lower estimate for $N(n)$.

In the case of previous example, for which $n = 5$, $k = 2$, $q_1 = 3$, $q_2 = 2$, this estimate yields $7 \cdot 3 - 4 \cdot 1 = 17 \leq N(5)$. So, the value of $N(5)$ is greater than the trivial lower bound $2^{5-1} = 16$.

To get the estimate for large n, let

$$q = q(n) = \log_2 n - \log_2 w(n),$$

where

$$w(n) = \frac{1}{\log_2 e} \cdot (\log_2 \log_2 n - \log_2 \log_2 \log_2 n - \log_2 \log_2 e - 1).$$

Define the nonnegative integers $k = k(n)$ and $r = r(n)$ by

$$n = q(n) \cdot k(n) + r(n), \quad 0 \le r(n) < q(n). \tag{4.5}$$

Let us choose q_i's in the following way:

$$q_1 = q_2 = \cdots = q_r = q + 1,$$
$$q_{r+1} = q_{r+2} = \cdots = q_k = q.$$

We have the inequalities

$$(1 - x)^y \ge 1 - xy, \quad 0 \le |x| \le 1, \quad y = 0 \text{ or } y \ge 1 \tag{4.6}$$

and

$$(1 - x)^y \le e^{-xy}, \quad 0 \le |x| \le 1, \quad y \ge 0 \tag{4.7}$$

from calculus.

In addition, by (4.5),

$$\frac{r(n)}{k(n) - r(n)} = \frac{r(n) \cdot q}{n - r(n) - r(n) \cdot q} \le \frac{r(n) \cdot q}{n - q - q^2} = o(1) \text{ for } n \to \infty. \tag{4.8}$$

Consequently, from (4.6) and (4.8) we get that

$$\begin{aligned}
\frac{1}{2^n} \prod_{i=1}^{k} (2^{q_i} - 1) &= \big(1 - \frac{1}{2^{q+1}}\big)^{r(n)} \cdot \big(1 - \frac{1}{2^q}\big)^{k(n)-r(n)} \\
&\ge (1 - \frac{r(n)}{2^{q+1}}) \cdot (1 - \frac{k(n) - r(n)}{2^q}) \\
&= 1 - \frac{k(n) - r(n)}{2^q} - \frac{r(n)}{2^{q+1}} + \frac{r(n)(k(n) - r(n))}{2^{2q+1}} \\
&\ge 1 - \frac{k(n) - r(n)}{2^q} - \frac{r(n)}{2^{q+1}} \\
&= 1 - \frac{k(n) - r(n)}{2^q} - \frac{r(n)(k(n) - r(n))}{2(k(n) - r(n))2^q} \\
&\ge 1 - \frac{k(n) - r(n)}{2^q} - o(1) \cdot \frac{k(n) - r(n)}{2^q} \\
&= 1 - \frac{k(n) - r(n)}{2^q} \cdot (1 + o(1)).
\end{aligned} \tag{4.9}$$

Similarly, by (4.5) and (4.7), we have that

$$\begin{aligned}\frac{1}{2^n}\prod_{i=1}^{k}(2^{q_i}-q_i-1) &= \big(1-\frac{q+2}{2^{q+1}}\big)^{r(n)}\cdot\big(1-\frac{q+1}{2^q}\big)^{k(n)-r(n)}\\ &\le \exp\Big(-\frac{\frac{1}{2}qr(n)+r(n)+qk(n)-qr(n)+k(n)-r(n)}{2^q}\Big)\\ &\le \exp\Big(-\frac{n+k(n)-r(n)-\frac{1}{2}qr(n)}{2^q}\Big). \qquad (4.10)\end{aligned}$$

By (4.5), the expression $k(n)-r(n)-\frac{1}{2}qr(n)$ is positive for sufficiently large n. Therefore, it follows from (4.10) that

$$\frac{1}{2^n}\prod_{i=1}^{k}(2^{q_i}-q_i-1) \le \exp\Big(-\frac{n}{2^q}\Big). \qquad (4.11)$$

Thus, from (4.4), by using (4.9) and (4.11), we have that

$$2^n\big(1-\frac{k(n)-r(n)}{2^q}\cdot(1+o(1))-\exp\big(-\frac{n}{2^q}\big)\big) \le N(n). \qquad (4.12)$$

Finally, by applying

$$\exp\Big(-\frac{n}{2^q}\Big)=\exp\Big(-\frac{n}{2^{\log_2 n-\log_2 w(n)}}\Big)=\exp(-w(n))=\frac{2}{\log_2 e}\cdot\frac{\log_2\log_2 n}{\log_2 n}$$

and

$$\frac{k(n)-r(n)}{2^q}\le\frac{n}{q2^q}\le\frac{2w(n)}{q}=\frac{2}{\log_2 e}\cdot\frac{\log_2\log_2 n}{\log_2}\cdot(1+o(1))$$

to (4.12), we get the left-hand side of (4.1). So, we have the lower estimate of $N(n)$.

Now, we prove the upper estimate of $N(n)$. Let $\mathcal{H}$ be the family of sets A having a pair B such that (A,B) is a basic functional dependency. If a set C satisfies $A\subset C\subset B$ and $|C|=|A|+1$, then $C\notin\mathcal{H}$. Indeed, by (3.2) and (3.3), we get that $\alpha(A)\subseteq\alpha(C)\subseteq\alpha(\alpha(A))=\alpha(A)$, which implies that $\alpha(C)=\alpha(A)$. It follows that C cannot be in $\mathcal{H}$. Such a set C can be obtained from at most n different sets A, so for at least $\frac{|\mathcal{H}|}{n}$ sets C we have that $C\notin\mathcal{H}$. Thus,

$$|\mathcal{H}|+\frac{|\mathcal{H}|}{n}\le 2^n.$$

So,

$$|\mathcal{H}|\le 2^n\cdot\frac{1}{1+\frac{1}{n}}=2^n\cdot\frac{n}{n+1}=2^n\cdot\frac{n+1-1}{n+1}=2^n\cdot(1-\frac{1}{n+1}),$$

which, by definition of $\mathcal{H}$, is equivalent to the right-hand side of (4.1). That is, we have shown the upper estimate of $N(n)$.

The proof is complete. □

Chapter 5

Minimum matrix representations

5.1 Minimum matrix representation of families of minimal keys

It is a natural extremal combinatorial question to ask what is the minimum number of records in a relation having a certain family of minimal keys. That is, it would be interesting to study minimum matrix representations of closure operations. So, the problem is to estimate the minimum number of rows a matrix must have in order to realize a given Sperner family.

Assume that a Sperner family $\mathcal{K}$ on an n-element ground set X is exactly the system of minimal keys for some relation. Then, the problem is to estimate $s(\mathcal{K})$.

The following lemma demonstrates the relationship between the number of rows and the number of antikeys in a relation:

Lemma 5.1 ([14]). *If M is an $m \times n$ matrix that represents $\mathcal{K}$, then*

$$\binom{m}{2} \geq |\mathcal{K}^{-1}|.$$

Proof. Let $A \in \mathcal{K}^{-1}$. Then, by Lemma 3.12, there exist two different rows i, j that are equal in A.

Let $B \in \mathcal{K}^{-1}$ be different from A. Similarly, let i', j' be the corresponding pairs of rows that are equal in B.

If the sets $\{i, j\}$ and $\{i', j'\}$ are equal, that is, i, j and i', j' are the same rows, then these two rows are equal in $A \cup B$ too. Thus, the closure of $A \cup B$ is not the whole X, which

means that $A \cup B$ is not a key. It follows that there exists $C \in \mathcal{K}^{-1}$ such that $C \supseteq A \cup B$. Therefore, since antikeys form a Sperner family, both $C = A$ and $C = B$ must hold at the same time, which contradicts to the assumption that A and B are different members of $\mathcal{K}^{-1}$.

Consequently, different members of $\mathcal{K}^{-1}$ must have different corresponding pairs of rows. So, there must be at least as many pairs of rows in M as there are members in $\mathcal{K}^{-1}$. That is, $\binom{m}{2} \geq |\mathcal{K}^{-1}|$ □

Clearly, Lemma 5.1 provides a lower bound for $s(\mathcal{K})$. The next lemma gives an upper bound for $s(\mathcal{K})$ in terms of the number of antikeys in a relation:

Lemma 5.2 ([11])**.**

$$s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}|.$$

Proof. Let us build a matrix realizing $\mathcal{K}$ and having $1+|\mathcal{K}^{-1}|$ rows as in the construction given after Definition 3.11.

The first row of the matrix consists of zeros only. Every other row corresponds to some member of $\mathcal{K}^{-1}$. That is, for $2 \leq i \leq 1 + |\mathcal{K}^{-1}|$ the ith row contains zeros in all columns that are in the corresponding antikey and i's in all other positions.

Thus, any subset A of X that is a superset of some member of $\mathcal{K}$ uniquely determines a row. Clearly, in this case, for every antikey there is at least one attribute of A which is not in this antikey. So, for $2 \leq i \leq 1 + |\mathcal{K}^{-1}|$ the ith row has the value i in at least one column from A, which implies that A is a key.

On the other hand, for any other subset of X there exists some antikey containing it. Therefore, for $2 \leq i \leq 1 + |\mathcal{K}^{-1}|$ the first and the ith rows are equal in this subset. It means that such a subset is not a key.

So, the matrix has $1 + |\mathcal{K}^{-1}|$ rows and realizes $\mathcal{K}$. Since this is a concrete construction, in general we have that $s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}|$ holds. □

By Lemma 5.1 and Lemma 5.2, we have the following estimates for $s(\mathcal{K})$ in terms of the size of $\mathcal{K}^{-1}$:

$$\sqrt{2}\sqrt{|\mathcal{K}^{-1}|} < s(\mathcal{K}) \leq 1 + |\mathcal{K}^{-1}|.$$

Since antikeys form a Sperner family, by using Lemma 5.2, we can get the following upper bound for $s(\mathcal{K})$:

Theorem 5.3 ([13])**.**

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

Proof. From Lemma 5.2 we have that the number of rows in a matrix realizing $\mathcal{K}$ is at most $1 + |\mathcal{K}^{-1}|$. Since $\mathcal{K}^{-1}$ is a Sperner family, by Sperner's Theorem we have that $|\mathcal{K}^{-1}| \leq \binom{n}{\lfloor \frac{n}{2} \rfloor}$. Thus, $s(\mathcal{K}) \leq 1 + \binom{n}{\lfloor \frac{n}{2} \rfloor}$ holds. □

So, we have estimates for the minimum number of records in a relation with a given family of minimal keys $\mathcal{K}$. However, it is also interesting to get some information about the minimum number of rows needed in the worst case, that is, to estimate the maximum value of $s(\mathcal{K})$, where $\mathcal{K}$ is taken over all Sperner families on X.

Definition 5.4. Assume that $n > 0$. Then let

$$s(n) = \max_{\mathcal{K}} s(\mathcal{K}),$$

where $\mathcal{K}$ is a Sperner family over an n-element set.

In other words, $s(n)$ is the smallest integer for which any Sperner family is realizable by a matrix with n columns and at most $s(n)$ rows.

By Theorem 5.3,

$$s(n) \leq 1 + \binom{n}{\lfloor \frac{n}{2} \rfloor}. \tag{5.1}$$

The following theorem gives the lower bound for $s(n)$:

Theorem 5.5 ([13])**.**

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

Proof. Clearly, any $\mathcal{K}$ can be realized by an $m \times n$ matrix, where $m \leq s(n)$.

We can add $(m+1)$th, $(m+2)$th, ..., $s(n)$th rows to a matrix realizing $\mathcal{K}$ and put completely new different values in them. Thus, the new matrix with $s(n)$ rows also realizes $\mathcal{K}$.

Assume that the integers in the first column of the matrix are $i_1 < i_2 < \cdots < i_r$, where $1 \leq r \leq s(n)$. Let us replace i_j by j for $1 \leq j \leq r$. It is easy to see that this replacement does not change the family of minimal keys. If we repeat the same procedure for all n columns of the matrix, we will have an $s(n) \times n$ matrix and all of its entries will be from $\{1, 2, \ldots, s(n)\}$.

The number of such matrices is $s(n)^{n\,s(n)}$, so the total number of $\mathcal{K}$'s is less than $s(n)^{n\,s(n)}$. Clearly, any family of $\lfloor \frac{n}{2} \rfloor$-element subsets of an n-element set is a Sperner family. Thus, since the number of such families is $2^{\binom{n}{\lfloor \frac{n}{2} \rfloor}}$, we have that

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

Therefore,

$$n\,s(n)\log_2 s(n) > \binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

So,

$$s(n)\log_2 s(n) > \frac{1}{n}\binom{n}{\lfloor \frac{n}{2} \rfloor}.$$

Thus, by (5.1), we have that

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

The theorem is proved. □

So, Theorem 5.5 implies that for any $n > 0$ there exists a Sperner family $\mathcal{K}$ over the n-element set such that

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

Let us consider the minimum number of records in a relation for which the family of minimal keys $\mathcal{K}$ contains only k-element subsets of X, that is, $\mathcal{K} \subseteq \binom{X}{k}$.

Definition 5.6. Assume that $0 < k \leq n$. Then let

$$f_k(n) = \max\{s(\mathcal{K}) : \mathcal{K} \subseteq \binom{X}{k}\}.$$

In other words, $f_k(n)$ is the maximum number rows in a minimum matrix representation of $\mathcal{K}$, taken among all $\mathcal{K} \subseteq \binom{X}{k}$.

The following theorem gives a lower bound for $f_k(n)$:

Theorem 5.7 ([9])**.**

$$f_k(n) \geq \sqrt{2}\binom{2k-2}{k-1}^{\lfloor n/(2k-2) \rfloor/2}.$$

Proof. Let $X = X_1 \cup X_2 \cup \cdots \cup X_q \cup Y$ be a partition of the set of attributes, where $q = \lfloor n/(2k-2) \rfloor$ and $|X_i| = 2k-2$, $1 \leq i \leq q$.

Let

$$\mathcal{K} = \{A \subseteq X : |A| = k,\ A \subseteq X_i \text{ for some } i\}$$

be the family of minimal keys.

By the definition of antikeys, they do not contain any minimal keys as their subsets. Every minimal key is an k-element subset of some X_i, thus the intersection of an antikey with all X_i's can contain at most $k-1$ elements. Since all antikeys must be maximal with their property, we have that

$$\mathcal{K}^{-1} = \{B : |B \cap X_i| = k-1 \text{ for all } i, |B \cap Y| = |Y|\}.$$

It follows that

$$|\mathcal{K}^{-1}| = \binom{2k-2}{k-1}^{\lfloor n/(2k-2)\rfloor},$$

since every antikey contains exactly $k-1$ attributes from every of $q = \lfloor n/(2k-2)\rfloor$ disjoint sets X_i of size $2k-2$.

Lemma 5.1 implies that

$$\binom{m}{2} \geq \binom{2k-2}{k-1}^{\lfloor n/(2k-2)\rfloor}$$

for any $m \times n$ matrix M representing $\mathcal{K}$.

Since m corresponds to the specific family $\mathcal{K} \subseteq \binom{X}{k}$ and $f_k(n)$ is the maximum among all such families, we get that $f_k(n) \geq m$.

Thus,

$$f_k(n) \geq \sqrt{2}\binom{2k-2}{k-1}^{\lfloor n/(2k-2)\rfloor/2}.$$

□

Clearly, for $k = 1$, that is, in case every minimal key is an attribute from X, we have that $f_1(n) = 2$. For $k = 2$, however, Theorem 5.7 implies that $f_2(n) \geq \sqrt{2} \cdot 2^{\lfloor n/2\rfloor/2} > 2^{n/4}$. So, even for the case where all minimal keys are just pairs of attributes, the value of $s(\mathcal{K})$ can be large.

5.2 Minimum matrix representation of uniform closure operations

Let us now consider the problem of minimum matrix representation in case of a specific class of closure operations, namely k-uniform closure operations.

We start with the following lemma which provides a fairly good lower estimate of $s(\alpha_k^n)$:

Lemma 5.8 ([14])**.**

$$\binom{s(\alpha_k^n)}{2} \geq \binom{n}{k-1}.$$

Proof. From Lemma 3.15 it follows that a matrix representing α_k^n has $\binom{X}{k}$ as its family of minimal keys. In this case, we know that the family of antikeys is $\binom{X}{k-1}$.

So, by Lemma 5.1, the statement is proved. □

The lower estimate given by Lemma 5.8 is sharp for $k = 1, 2, n-1$ and is sharp up to a constant depending on k for any k when $n \to \infty$.

In order to get the exact value of $s(\alpha_n^n)$, we need another lemma.

Definition 5.9. Assume that M is an $m \times n$ matrix. Then let $G(M)$ denote the graph whose vertices correspond to the rows of M, and two vertices are connected by an edge if and only if the set of attributes $A \subseteq X$ where the two associated rows are equal is nonempty. Every edge is labeled by the corresponding set A.

Lemma 5.10 ([14])**.** *Assume that M is an $m \times n$ matrix. Let $A_1, A_2, \ldots, A_r$ be the labels along a cycle in the graph $G(M)$. Then*

$$\Big(\bigcap_{i=1, i\neq j}^{r} A_i \Big) \setminus A_j = \emptyset, \quad 1 \leq j \leq r.$$

Proof. Suppose, by contradiction, that the lemma does not hold. Then there exists a column u which is in every A_i, but not in A_j for some fixed j.

Let $v_1, v_2, \ldots, v_r$ be the vertices of the cycle such that the edge (v_i, v_{i+1}) for $1 \leq i < r$ is labeled by A_i and the edge (v_r, v_1) is labeled by A_r.

Since u is in every A_i except A_j, we have that $u \in A_{j+1}$. Therefore, since the edge (v_{j+1}, v_{j+2}) is labeled by A_{j+1}, the rows corresponding to v_{j+1} and v_{j+2} must have equal entries in the column u. Clearly, the same holds for v_{j+2} and v_{j+3}, …, v_r and v_1, …, v_{j-1} and v_j. Thus, the rows corresponding to the vertices $v_{j+1}, v_{j+2}, \ldots, v_r, v_1, \ldots, v_j$ have equal values in the column u. So, since for v_j and v_{j+1} the associated rows are equal in u, it follows that $u \in A_j$, which is a contradiction.

The lemma is proved. □

We are now ready to state the following theorem:

Theorem 5.11 ([14])**.**

$$s(\alpha_1^n) = 2, \tag{5.2}$$

$$s(\alpha_2^n) = \left\lceil \frac{1+\sqrt{1+8n}}{2} \right\rceil, \tag{5.3}$$

$$s(\alpha_{n-1}^n) = n, \tag{5.4}$$

$$s(\alpha_n^n) = n+1. \tag{5.5}$$

Proof. Lemma 3.15 implies that the family of minimal keys in a matrix representing α_k^n is exactly $\binom{X}{k}$, that is, $s(\alpha_k^n) = s(\mathcal{K}(\alpha_k^n)) = s(\binom{X}{k})$.

Let us start with (5.2). From Lemma 5.8 it follows that $s(\alpha_1^n) \geq 2$. A concrete matrix realizing α_1^n and having exactly 2 rows proves the equality. The construction of the matrix is the following:

1	2	...	n
0	0	...	0
1	1	...	1

Obviously, every column is a key, since if the value in a column is zero, it corresponds to the first row, and otherwise it corresponds to the second row.

Now, we show (5.3). Lemma 5.8 implies that

$$\binom{s(\alpha_2^n)}{2} \geq n.$$

If we solve the inequality for $s(\alpha_2^n)$, we get it in the form as in (5.3). To prove the equality, we have to build a matrix realizing α_2^n with exactly $s(\alpha_2^n)$ rows. The construction of the matrix is the following. Every column contains a unique pair of zeros. That is, for any column the pair of rows having zeros in this column is different. The other entries of a row are equal to its ordinal number.

It is easy to see that the entries in any two columns can uniquely determine a row, since all pairs of zeros have distinct positions and all other entries in a row are specific to this row. Furthermore, any single column cannot be a key, because there are two rows having zeros in this column. Consequently, by Lemma 3.12, the matrix represents $\binom{X}{2}$.

For example, for $n = 5$, the matrix has $m = \lceil (1+\sqrt{1+8\cdot 5})/2 \rceil = 4$ rows and may be the following:

1	2	3	4	5
0	0	0	1	1
0	2	2	0	0
3	0	3	0	3
4	4	0	4	0

Let us show (5.4). Again, from Lemma 5.8 we have that $s(\alpha_1^n) \geq n$. The unit matrix of size n proves the equality:

1	2	...	n
1	0	...	0
0	1	...	0
⋮	⋮	⋱	⋮
0	0	...	1

Any $n-1$ columns uniquely determine a row (one row has all values equal to zero and others have one's in distinct positions), while for any $n-2$ columns there are two rows having all zeros in the corresponding positions. Thus, by Lemma 3.12, the matrix represents $\binom{X}{n-1}$.

Finally, we show (5.5). Suppose that some $m \times n$ matrix realizes $\binom{X}{n}$, that is, only the whole attribute set X is a key. Lemma 3.12 implies that for any $A \in \binom{X}{n-1}$ there exists an edge in $G(M)$ which is labeled by A. Since there are n such sets A, there are n corresponding edges. All of these edges are different, because otherwise it contradicts to the fact that X is a key.

Clearly, $(n-1)$-element subsets cannot satisfy the condition of Lemma 5.10, since for any collection of such subsets, the intersection of all of its members except one is the element which is not present in the remaining subset. That is, from Lemma 5.10 it follows that the corresponding edges cannot form a cycle. So, since $G(M)$ contains n edges which do not form a cycle, it has at least $n+1$ vertices, that is, $s(\alpha_n^n) \geq n+1$. The unit matrix of size n with an additional row containing all zeros proves the equality:

1	2	...	n
0	0	...	0
1	0	...	0
0	1	...	0
⋮	⋮	⋱	⋮
0	0	...	1

In this matrix, all rows are different and for any $n-1$ columns there are two rows having all zeros in them. Therefore, by Lemma 3.12, we have that the matrix represents α_n^n.

The theorem is proved. □

For $k = 3$, Lemma 5.8 implies that $s(\alpha_3^n) \geq n$. In fact, this is sharp for $n = 7$ and for all $n \geq 9$:

Theorem 5.12 ([4])**.**

$$s(\alpha_3^n) = n, \quad n = 7 \text{ or } n \geq 9.$$

There is no formula for $s(\alpha_k^n)$ in the general case. However, the following theorem provides its upper and lower estimates:

Theorem 5.13 ([9])**.**

$$\sqrt{2}\Big(\frac{1}{k-1}\Big)^{(k-1)/2} n^{(k-1)/2} < s(\alpha_k^n) < 2^{3k/2} n^{(k-1)/2}, \quad 2 \leq k < n.$$

Proof. From Lemma 5.8 and the properties of binomial coefficients we have that

$$\binom{s(\alpha_k^n)}{2} \geq \binom{n}{k-1} \geq \Big(\frac{n}{k-1}\Big)^{k-1}.$$

That is,

$$s(\alpha_k^n)^2 > 2\Big(\frac{1}{k-1}\Big)^{k-1} n^{k-1}.$$

Thus, we get that

$$s(\alpha_k^n) > \sqrt{2}\Big(\frac{1}{k-1}\Big)^{(k-1)/2} n^{(k-1)/2},$$

which proves the lower bound.

Now, we show the upper bound.

Assume that p is a prime number. We build a set D having $2\lfloor\sqrt{p}\rfloor$ elements and such that for any integer i there exist two elements $d_1, d_2 \in D$ with the following property:

$$i \equiv d_1 - d_2 \pmod{p}.$$

Let us prove that the construction

$$D = \{0, 1, 2, \ldots, a-2, a-1, 2a, 3a, \ldots, (a-2)a, (a-1)a\},$$

where $a = \lceil\sqrt{p}\rceil$, works. We assume that $p > 9$, that is, $a \geq 4$ (other cases can be checked separately).

Suppose that $0 \leq i < p$ and $i = al + r$ for $0 \leq r < a$ and $0 \leq l < a$. If $l = 0$, then $d_1 = r \in D$ and $d_2 = 0$. If $l = 1$ and $r = 0$, then $d_1 = 3a$ and $d_2 = 2a$. If $2 \leq l \leq a - 1$ and $r = 0$, then $d_1 = al$ and $d_2 = 0$. Otherwise, if $1 \leq l \leq a - 2$, then $d_1 = (l + 1)a$ and $d_2 = a - r$.

So, the set D satisfies the property and

$$|D| = 2a - 2 = 2(\lceil\sqrt{p}\rceil - 1) = 2\lfloor\sqrt{p}\rfloor.$$

Let $\mathcal{P}$ be the following set of polynomials:

$$\mathcal{P} = \{c_{k-1}x^{k-1} + c_{k-2}x^{k-2} + \cdots + c_1x + c_0\colon\ c_0, c_1, \ldots, c_{k-2} \in D,\ c_{k-1} \in \{0, 1\}\}.$$

Clearly,

$$|\mathcal{P}| = 2^k\lfloor\sqrt{p}\rfloor^{k-1}.$$

We build a $|\mathcal{P}|\times p$ matrix M realizing α_k^p. Every row of M is associated with a polynomial from $\mathcal{P}$. The columns are enumerated from 0 to $p-1$. The jth entry of the row associated with $z(x) \in \mathcal{P}$ is $z(j) \pmod p$. So, all entries of M belong to $\{0, 1, 2, \ldots, p - 1\}$.

Let us show that M represents α_k^p. By Lemma 3.15, we need to show that its family of minimal keys is $\binom{X}{k}$, $|X| = p$.

To show that any k-element subset of X is a key, suppose, by contradiction, that the rows corresponding to polynomials $z_1(x)$ and $z_2(x)$ have k equal entries, that is,

$$z_1(t_i) \equiv z_2(t_i) \pmod p, \quad 0 \leq t_1 < t_2 < \cdots < t_k \leq p - 1.$$

So, the polynomial $z_1(x)-z_2(x)$ has k different roots. Since the degree of this polynomial is at most $k - 1$, this is a contradiction. Consequently, any k-element subset of X is a key.

Now, we show that any $k - 1$ attributes do not form a key. Let $0 \leq t_1 < t_2 < \cdots < t_{k-1} \leq p - 1$ be arbitrarily chosen integers. We have to find two different rows with equal entries in the columns corresponding to $t_1, t_2, \ldots, t_{k-1}$. Let $w(x)$ be a polynomial having these integers as its roots:

$$w(x) = (x - t_1)(x - t_2)\cdots(x - t_{k-1}) = x^{k-1} + a_{k-2}x^{k-2} + \cdots + a_1x + a_0.$$

By the definition of D, for any a_i there exist $c_i, c_i' \in D$ such that $a_i \equiv c_i - c_i' \pmod{p}$. So, for

$$z(x) = x^{k-1} + c_{k-2}x^{k-2} + \cdots + c_1 x + c_0,$$
$$z'(x) = c'_{k-2}x^{k-2} + \cdots + c'_1 x + c'_0$$

we have that

$$w(x) = z(x) - z'(x).$$

Clearly, $z(x)$ and $z'(x)$ are different members of $\mathcal{P}$ and $z(t_i) \equiv z'(t_i) \pmod{p}$ for $i = 1, 2, \ldots, k-1$, from the definition of $w(x)$. Consequently, for any $(k-1)$-element subset of X there exist two different rows having equal entries in the corresponding positions.

Thus, by Lemma 3.12, we get that M represents α_k^p. This means that

$$s(\alpha_k^p) \leq 2^k p^{(k-1)/2}.$$

By Chebyshev's theorem, for any integer n there exists a prime number p such that $n \leq p \leq 2n$. If we build a matrix M representing α_k^p and omit $p - n$ columns, then the resulting matrix will represent α_k^n. Therefore,

$$s(\alpha_k^n) \leq 2^k p^{(k-1)/2} \leq 2^k (2n)^{(k-1)/2} < 2^{3k/2} n^{(k-1)/2},$$

which proves the upper bound.

The proof is complete. □

Theorem 5.13 determines the asymptotic behavior of $s(\alpha_k^n)$ for a fixed k. However, it provides good estimates only for small values of k. For example, for $k = n/2$, Lemma 5.2 gives a much better estimate.

Chapter 6

Relational models and secret sharing

6.1 Secret sharing

Secret sharing allows a dealer holding a secret to distribute shares of this secret among a group of participants such that only qualified subsets of participants can reconstruct the secret from their shares. It was introduced independently by Shamir [26] and Blakley [5] in 1979.

Let P be the set of n participants. Assume that the *dealer* has the *secret* s chosen from some set S. Each participant receives a *share* of the secret. Subsets of P that can recover the secret are called *qualified*, and the other subsets are called *unqualified*. The collection of qualified sets is called an *access structure* and is denoted by $\mathcal{A} \subset 2^P$. A qualified subset is called *minimal* if it is minimal with its property, that is, none of its proper subsets is qualified. A *secret sharing scheme* for $\mathcal{A}$ is a method by which the dealer distributes the shares of the secret to the participants in such a way that only subsets that are in $\mathcal{A}$ can reconstruct the secret by using the corresponding shares. A secret sharing scheme is *perfect* if any unqualified subset cannot get any information about the secret from the shares of its members.

In order for a secret sharing scheme to exist, $\mathcal{A}$ has to be upward-closed (if $A \in \mathcal{A}$ and $A \subseteq B \subseteq P$, then $B \in \mathcal{A}$), and non-trivial (there is at least one subset in $\mathcal{A}$). If the first property holds, then the collection of minimal qualified subsets uniquely determines the access structure. Informally, this property means that if a subset of participants can recover the secret, then when further participants are added to this subset, it still can

reconstruct the secret. Clearly, since $\mathcal{A}$ is upward-closed and non-empty, the whole set of participants is qualified, that is, $P \in \mathcal{A}$.

In this chapter, we consider only the so-called *threshold* schemes. For these schemes, a subset of participants is qualified if and only if it has enough members. If the qualified subsets are exactly the subsets containing at least k $(0 < k \leq n)$ participants, then the scheme is called a (k, n)-*threshold scheme*. In this case, all members of $\mathcal{A}$ have cardinality greater than or equal to k. It is easy to see that $\mathcal{A}$ is upward-closed and non-empty, so the corresponding secret sharing scheme exists.

As an illustration of a secret sharing scheme, let us consider the (k, n)-threshold scheme proposed by Shamir [26]. This scheme is based on the idea that we need at least k points to define a polynomial of degree $k - 1$.

Assume that $S = \mathbb{Z}_p$ for some prime p, where $p > n$. Let $f(x)$ be a random polynomial over $\mathbb{Z}_p$ of degree at most $k - 1$ such that $f(0) = s$. So, $f(x) = a_{k-1}x^{k-1} + a_{k-2}x^{k-2} + \cdots + a_1x + s$, where $a_1, a_2, \ldots, a_k \in \mathbb{Z}_p$. The share of the ith $(1 \leq i \leq n)$ participant is $s_i = f(i) \pmod p$. Thus, the dealer randomly chooses the values $a_1, a_2, \ldots, a_k \in \mathbb{Z}_p$, defines the polynomial $f(x)$, and distributes the shares to the participants.

By polynomial interpolation, the scheme is correct, that is, any group of at least k participants can recover the secret. Indeed, $f(x)$ has degree of at most $k - 1$, so it can be reconstructed from its values in at least k points. Since every participant receives the value of $f(x)$ in some point and all points are distinct, a group of at least k persons can find the coefficients of $f(x)$, which means that it can recover the value of s.

Let us now show that a group of less than k participants cannot get any information about s. We fix any index set I of size $k - 1$ and any secret s. Consider a set of shares for the participants with indices in I. By polynomial interpolation, there is exactly one polynomial $f(x)$ of degree at most $k-1$ such that $f(0) = s$ and $f(i) = s_i$ for $i \in I$. Thus, any set of shares can result from sharing s. There are p^{k-1} sets of shares for participants with indices in I and p^{k-1} polynomials over $\mathbb{Z}_p$ of degree at most $k - 1$ that have value s at 0, so there is a one-to-one correspondence between them. Since the dealer chooses an appropriate polynomial $f(x)$ in a random way, every set of shares with indices in I is equally likely to result. So, the corresponding group of participants has no information about s. The secret s and the index set I were chosen arbitrarily, which implies that for unqualified subsets their shares cannot give any information about the secret.

Consequently, since any group of at least k participants can recover the secret and any group of at most $k - 1$ participants cannot get any information about the secret, the scheme is a perfect (k, n)-threshold scheme.

6.2 From secret sharing to relational models

Assume that there are n participants and m possible secrets. That is, by using the notation from the previous section, $|P| = n$ and $|S| = m$. The dealer chooses one secret out of m possibilities and sends to each participant a share of the secret.

This can also formulated by using an $m \times n$ matrix M with the column set X. The columns of the matrix correspond to the participants and the rows correspond to the possible secrets. Thus, the secret is the index of the row chosen by the dealer. The share of a person is the secret row value in this person's column. In other words, for the ith participant the share is a_{si}, where s is the index of the secret row. Consequently, the ith person knows that the secret row is one of the rows that have the value a_{si} in the ith column. So, the matrix M is the following:

1	2	...	i	...	n-1	n	⟵X
a_{11}	a_{12}	...	a_{1i}	...	a_{1n-1}	a_{1n}	
a_{21}	a_{22}	...	a_{2i}	...	a_{2n-1}	a_{2n}	
⋮	⋮	⋱	⋮	⋱	⋮	⋮	
a_{s1}	a_{s2}	...	a_{si}	...	a_{sn-1}	a_{sn}	⟵ secret row
⋮	⋮	⋱	⋮	⋱	⋮	⋮	
a_{m1}	a_{m2}	...	a_{mi}	...	a_{mn-1}	a_{mn}	

↑

ith share

In order to have a (k, n)-threshold scheme, we need the matrix M to have the following properties. First, any k out of n columns should uniquely determine a row. Second, for any choice of the secret row and for any $k-1$ columns, there should be at least two rows with the values in these columns equal to the shares of the associated participants. For such a matrix, a group with at least k persons can find the secret, and a group with less than k persons cannot uniquely determine the secret, so we have a (k, n)-threshold scheme. For an illustration, consider the following matrix:

1	2	3
1	1	1
1	2	2
2	3	3
2	1	2
3	2	3
3	3	1

Clearly, any two columns form a key. On the other hand, for any choice of the secret row a participant has two possible rows with equal values in the associated column. That is, the secret cannot be uniquely determined by only one person. Consequently, the matrix corresponds to a $(2, 3)$-threshold scheme.

In this chapter, we consider a slightly different version of the matrix M. The first property of the matrix is the same as before, that is, any k-element subsets of X is a key. The second property, however, is different. The matrix should have at least two distinct rows with equal corresponding values for at least one choice of the secret row, but not necessarily for all m possibilities. In other words, we require that for any $k-1$ columns there should be at least two rows with equal entries in these columns. So, in this case, the matrix M represents $\binom{X}{k}$.

Since the family of minimal keys of M is $\binom{X}{k}$, any k participants can recover the secret row from their shares, so the correctness of the associated scheme holds. However, it may happen that for some choice of the secret row and for some subset of participants with less than $k-1$ members there is exactly one row having the corresponding values. In other words, a particular group of less than $k-1$ participants can recover the secret row for a particular choice of the secret. That is, the privacy of the scheme does not always hold. The following matrix illustrates this for $k = 2$ and $n = 3$:

1	2	3	
0	1	1	
0	0	2	
3	0	0	⟵ secret row
4	4	0	

Any two columns form a key. In addition, for any single column there are two rows with equal values in this column. However, for a choice of the secret row, there is a participant that can determine it. For example, if $s = 3$, then for the second and third persons there are two possibilities, while there is only one row with the value 3 in the first column, so the first person can determine the secret row.

Since the matrix M represents $\binom{X}{k}$, it follows from the definition of antikeys that for every set of at most $k-1$ columns there exist at least two rows with equal values in these columns. So, it may happen that for an unqualified set of participants there are exactly two rows with equal corresponding values. Even though the members of this set do not know the exact row, they have only two possible choices, which is not very good for protecting the secret. To this end, it would be better to have bigger sets of the potential secret rows for unqualified sets of participants.

To achieve this, we require every antikey to have at least l for $l \geq 2$ rows with equal values in the associated columns. For every key, as before, all rows have a unique set of values in the corresponding columns, because every qualified subset must be able to recover the secret. Thus, this is a more general version of keys and antikeys that we had before. While a key still uniquely determines a row, for every antikey we require not only two but, more generally, l rows having equal values in the antikey columns.

Similarly to the previous notation, a matrix with n columns such that any subset of at least k columns uniquely determines a row and for any subset of at most $k-1$ columns there are at least l rows with equal entries in these columns is said to *represent (realize)* $\alpha^n_{k,l}$. Note that for $l = 2$ we get α^n_k, as before.

Clearly, the properties of keys and antikeys depend only on their equality in the columns, but not on the actual entries of the matrix. Consequently, we can consider a column to be a partition of the set of rows defined by the equality of the corresponding entries. Thus, every column is a partition of $\{1, 2, \ldots, m\}$, and there are n such partitions in the matrix M. In other words, a column is a collection of disjoint sets, called *classes*, such that their union is the whole set $\{1, 2, \ldots, m\}$. A subset $A \subseteq \{1, 2, \ldots, m\}$ is *covered* by a partition if A is contained in one of the classes of the partition.

By representing columns as partitions, the conditions for the matrix M to realize $\alpha^n_{k,l}$ can be formulated as follows:

1. For any $k-1$ partitions there are $k-1$ classes, one from each of the partitions, such that the intersection of these classes has size of at least l.

2. Each 2-element subset of $\{1, 2, \ldots, m\}$ is covered by at most $k-1$ different partitions.

The first condition is needed to guarantee that for every $k-1$ columns there are at least l rows with equal corresponding entries. The second condition ensures that any k columns uniquely determine a row, that is, for every pair of rows there are no more than $k-1$ columns in which these rows have equal values.

Consider the following matrix which illustrates the case $n = 3, k = 3, l = 3$:

1	2	3
0	0	0
0	0	2
0	0	3
0	4	0
0	5	0
6	0	0
7	0	0

Indeed, for any pair of the columns there are three rows having all zeros in these columns. For example, the first three rows have zeros in the first two positions. In addition, all rows are distinct, so three columns uniquely determine a row.

Since the first five rows have zeros in the first column and the last two rows have the values 6 and 7, the partition corresponding to this column contains three classes: $\{1,2,3,4,5\},\{6\},\{7\}$. In the same way, the second partition is $\{\{1,2,3,6,7\},\{4\},\{5\}\}$ and the third partition is $\{\{1,4,5,6,7\},\{2\},\{3\}\}$.

For every two partitions we can choose a class from each of them such that these two classes have three common elements. Furthermore, for three partitions there is no pair of rows that is covered by all of them.

So, instead of using a matrix, we can formulate the problem in terms of the families of partitions. In this case, we have to find a family of n partitions of $\{1,2,\ldots,m\}$ with certain intersection properties.

6.3 Some results

In this section, we consider minimum matrix representation of $\alpha^n_{k,l}$. That is, we want to estimate $s(\alpha^n_{k,l})$.

We start with the following lemma:

Lemma 6.1.

$$\binom{s(\alpha^n_{k,l})}{2} \geq \binom{n}{k-1}\binom{l}{2}.$$

Proof. Let A be a $(k-1)$-element subset of X. So, there exist l rows that are equal in A. That is, there are $\binom{l}{2}$ pairs of rows having equal values in A. In the same way, there are $\binom{l}{2}$ pairs of rows for a $(k-1)$-element subset $B \subseteq X$ different from A.

Suppose that these two sets of pairs intersect, that is, there are some pairs of rows that are equal in A and in B at the same time. Then, these pairs are equal in $A \cup B$ too. This implies that $A \cup B$ is not a key. However, since A and B are different, $|A \cup B| \geq k$. So, this contradicts to the fact that all subsets with at least k elements are keys.

Thus, any $(k-1)$-element subset of X has a unique set of $\binom{l}{2}$ pairs of rows associated with it. There are $\binom{n}{k-1}$ such subsets of columns, and each of them has $\binom{l}{2}$ unique pairs of rows. Since the set of rows must contain enough pairs, we have that $\binom{m}{2} \geq \binom{n}{k-1}\binom{l}{2}$. □

Lemma 6.1 is a generalization of Lemma 5.8. It provides a lower bound for $s(\alpha^n_{k,l})$:

$$\binom{s(\alpha^n_{k,l})}{2} \geq \binom{n}{k-1}\binom{l}{2} \geq \frac{n^{k-1}}{(k-1)^{k-1}} \cdot \frac{l^2}{4}. \tag{6.1}$$

As a result,

$$s(\alpha^n_{k,l}) > \frac{l}{\sqrt{2}}\left(\frac{1}{k-1}\right)^{(k-1)/2} n^{(k-1)/2}. \tag{6.2}$$

Notice that for $l = 2$, (6.2) gives the lower bound from Theorem 5.13.

By generalizing the construction in Lemma 5.2, we can build a matrix representing $\alpha^n_{k,l}$. The first row contains zeros only. Every other row corresponds to some $(k-1)$-element subset of columns so that for any $(k-1)$-element subset of X there is a set of $l-1$ rows associated with it. Thus, the ith $(2 \leq i \leq m)$ row has zeros in the columns from the corresponding $(k-1)$-element subset and i's in all other positions. So, the matrix has $1 + (l-1) \cdot \binom{n}{k-1}$ rows. Any k columns uniquely determine a row, and for any $k-1$ columns we have l rows equal in them. Consequently, we get the following trivial upper bound for $s(\alpha^n_{k,l})$:

$$s(\alpha^n_{k,l}) \leq 1 + (l-1) \cdot \binom{n}{k-1}. \tag{6.3}$$

Now, let us consider $s(\alpha^n_{k,l})$ for concrete values of parameters. Clearly, for $l = 2$, the matrix represents $\binom{X}{k}$, so, in this case, the estimates for the minimum number of rows are as in Section 5.2.

For $k = 2$, consider a Steiner system $S(2, q+1, q^2+q+1)$ (see Appendix B), where q is a power of a prime. Such a system exists, since it is a projective plane of order q. By using this system, we can construct a matrix realizing $\alpha^{q^2+q+1}_{2,q+1}$ and having q^2+q+1 rows. Each column corresponds to one block. When considered as a partition, every column contains a class with zero values corresponding to the block and singleton classes with each of the values equal to the row's ordinal number. Thus, for every column there are exactly $q+1$ rows with all zeros in this column. On the other hand, for every pair of columns no two rows are equal in them, since for every pair of points there is a unique

block containing them, and all entries of a row that are not in some block are unique to this row. So,

$$s(\alpha_{2,q+1}^{q^2+q+1}) \leq q^2+q+1. \tag{6.4}$$

By Lemma 6.1,

$$\binom{s(\alpha_{2,q+1}^{q^2+q+1})}{2} \geq \binom{q^2+q+1}{1}\binom{q+1}{2} = \frac{(q^2+q+1)(q^2+q)}{2} = \binom{q^2+q+1}{2}. \tag{6.5}$$

Together with (6.4), this implies that

$$s(\alpha_{2,q+1}^{q^2+q+1}) = q^2+q+1. \tag{6.6}$$

For example, for $q=2$ ($n=7$, $k=2$, $l=3$), the matrix is the following:

1	2	3	4	5	6	7
0	0	0	1	1	1	1
0	2	2	0	2	2	0
0	3	3	3	0	0	3
4	4	0	4	0	4	0
5	0	5	0	0	5	5
6	0	6	6	6	0	0
7	7	0	0	7	0	7

Let us consider the case $k=2$ and $l=3$. By Lemma 6.1,

$$\binom{s(\alpha_{2,3}^{n})}{2} \geq \binom{n}{1}\binom{3}{2} = 3n. \tag{6.7}$$

For an upper bound, we use a Steiner triple system (see Appendix B). We build a matrix in the following way. Each column is associated with one triple. It contains zero values in the rows with positions corresponding to the triple and the values equal to the row's ordinal number for all other rows. Therefore, every column has exactly three rows with all zero values in this column. In addition, every pair of columns has at most one row with a specified pair of values, by the properties of Steiner triple system. Thus, the matrix represents $\alpha_{2,3}^{n}$. Clearly, the number of triples over the set of m rows must be at least n. So,

$$\frac{m(m-1)}{6} \geq n. \tag{6.8}$$

Since we can construct a Steiner triple system for $m \equiv 1 \pmod 6$ or $m \equiv 3 \pmod 6$, we have to find the smallest value of m satisfying one of these congruences together with (6.8). Consequently,

$$\lceil \sqrt{6}\sqrt{n} \rceil \leq s(\alpha_{2,3}^{n}) \leq \lceil \sqrt{6}\sqrt{n} \rceil + 4. \tag{6.9}$$

The case $k = 3$ is considered in [19] where it is formulated in terms of partitions with certain intersection properties.

The following theorem gives lower and upper estimates of $s(\alpha^n_{k,l})$ for the case $l = 3$:

Theorem 6.2.

$$\frac{3}{\sqrt{2}}\Big(\frac{1}{k-1}\Big)^{(k-1)/2} n^{(k-1)/2} < s(\alpha^n_{k,3}) < 2^{(k+2)\log_3 2} n^{(k-1)\log_3 2}.$$

Proof. We can get the lower bound by substituting $l = 3$ into (6.2).

Let us show the upper bound. The following proof is based on the ideas of the proof of Theorem 5.13.

Assume that p is a prime number. By [25], there exists a set D of size at most $p^{\log_3 2}$ such that for any integer i we can find three elements $d_1, d_2, d_3 \in D$ with the following property:

$$i \equiv d_1 - d_2 \equiv d_2 - d_3 \pmod{p}.$$

Define the set of polynomials $\mathcal{P}$ in the following way:

$$\mathcal{P} = \{c_{k-1}x^{k-1} + c_{k-2}x^{k-2} + \cdots + c_1 x + c_0 \colon\ c_0, c_1, \ldots, c_{k-2} \in D,\ c_{k-1} \in \{0,1,2\}\}.$$

So,

$$|\mathcal{P}| \le 3p^{(k-1)\log_3 2}.$$

By using $\mathcal{P}$, we build a matrix M representing $\alpha^p_{k,3}$. The matrix M contains $|\mathcal{P}|$ rows, each of them corresponding to a polynomial from $\mathcal{P}$, and p columns enumerated from 0 to $p-1$. For the row associated with $z(x) \in \mathcal{P}$, the jth entry is $z(j) \pmod{p}$.

Let us show that this is a valid construction.

First, we demonstrate that any k-element subset of X is a key. Assume, by contradiction, that the rows corresponding to the polynomials $z_1(x)$ and $z_2(x)$ have k equal entries:

$$z_1(t_i) \equiv z_2(t_i) \pmod{p}, \quad 0 \le t_1 < t_2 < \cdots < t_k \le p-1.$$

Therefore, the polynomial $z_1(x) - z_2(x)$ has k different roots. Since the degree of this polynomial is at most $k-1$, this is a contradiction. So, any k-element subset of X is a key.

Second, we show that for any $(k-1)$-element subset of X there are at least three rows with equal entries in the corresponding positions. Assume that the integers $0 \le t_1 <$

$t_2 < \cdots < t_{k-1} \leq p-1$ are chosen arbitrarily. Thus, we need to find three different rows with equal entries in the columns corresponding to $t_1, t_2, \ldots, t_{k-1}$. Define the polynomial $w(x)$ in the following way:

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

By the definition of D, for any a_i there exist $c_i, c_i', c_j'' \in D$ such that $a_i \equiv c_i - c_i' \equiv c_i' - c_i''$ (mod p). Consequently, for

$$\begin{aligned} z(x) &= 2x^{k-1} + c_{k-2}x^{k-2} + \cdots + c_1 x + c_0, \\ z'(x) &= x^{k-1} + c'_{k-2}x^{k-2} + \cdots + c'_1 x + c'_0, \\ z''(x) &= c''_{k-2}x^{k-2} + \cdots + c''_1 x + c''_0 \end{aligned}$$

we have that

$$w(x) = z(x) - z'(x) = z'(x) - z''(x).$$

It is easy to see that $z(x), z'(x)$ and $z''(x)$ are different members of $\mathcal{P}$. Moreover, by the definition of $w(x)$, $z(t_i) \equiv z'(t_i) \equiv z''(t_i) \pmod{p}$ for $i = 1, 2, \ldots, k-1$. As a result, for any $(k-1)$-element subset of X there exist three different rows having equal values in the corresponding columns.

So, the matrix represents $\alpha^p_{k,3}$. Therefore,

$$s(\alpha^p_{k,3}) \leq 3p^{(k-1)\log_3 2}.$$

Chebyshev's theorem states that for any integer n there is a prime number p such that $n \leq p \leq 2n$. We can build a matrix representing $\alpha^p_{k,3}$ and just omit $p - n$ of them. Clearly, the resulting matrix represents $\alpha^n_{k,3}$. Thus,

$$s(\alpha^n_{k,3}) \leq 3p^{(k-1)\log_3 2} \leq 3(2n)^{(k-1)\log_3 2} < 2^{(k+2)\log_3 2} n^{(k-1)\log_3 2},$$

which proves the upper bound.

The proof is complete. □

Chapter 7

Conclusion

In this thesis, we presented extremal combinatorial problems in relational databases. We have shown some of the most important results in this area. By starting with providing essential definitions and concepts, we have demonstrated the results concerning maximum numbers of minimal keys and basic functional dependencies in a relation and minimum matrix representation of closure operations.

We have also introduced a connection between relational models and secret sharing. As a consequence, we have generalized the results for minimum matrix representation of α_k^n for the case of $\alpha_{k,l}^n$, motivated by the considered connection. In particular, we have provided the lower bound of $s(\alpha_{k,l}^n)$ for arbitrary values of n, k, and l. Moreover, we have presented constructions, examples, and estimates for certain values of k and l. Finally, we have given the upper bound of $s(\alpha_{k,l}^n)$ for the case $l = 3$.

On the whole, the relational model gives rise to many interesting extremal combinatorial problems. There are many directions for further research in this area. They include obtaining more precise estimates for some special cases and generalizing the previous results.

Appendix A

Sperner families

Let $\mathcal{F}$ be a collection of subsets of an n-element set X.

Definition A.1. A family of sets $\mathcal{F}$ is called a *Sperner family* (or *Sperner system*) if none of the sets contains another one. That is, if $A, B \in \mathcal{F}$ and $A \neq B$, then $A \not\subseteq B$ and $B \not\subseteq A$.

It is easy to see that $\binom{X}{k}$ for $k = 0, 1, \ldots, n$ is a Sperner family. Indeed, since all of its members have the same cardinality k and are distinct, no set can be contained in another one. Clearly, this family has $\binom{n}{k}$ members.

Therefore, since $k = \lfloor n/2 \rfloor$ gives the maximum value of $\binom{n}{k}$, there exists a Sperner family having $\binom{n}{\lfloor n/2 \rfloor}$ members. The following theorem, due to Sperner [27], states that such a family is the largest:

Theorem A.2 (Sperner's Theorem)**.** *Let $\mathcal{F}$ be a Sperner family over an n-element set. Then*

$$|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}.$$

Sperner's Theorem is implied by the following theorem which is more sharp. It was independently discovered by Lubell [23], Meshalkin [24], and Yamamoto [30]. Even though this is a special case of a result of Bollobás [6], it is more commonly known as the *LYM inequality*:

Theorem A.3 (LYM inequality)**.** *Let $\mathcal{F}$ be a Sperner family over an n-element set. Then*

$$\sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}} \leq 1.$$

Proof. The idea of the following proof, which is a reformulation of a proof due to Lubell [23], is to associate each member of $\mathcal{F}$ with some permutations of X, and then count the number of such permutations.

There are $n!$ permutations of X. We can generate a permutation of X by selecting some set $A \in \mathcal{F}$ and then concatenating a permutation of A with a permutation of $X \setminus A$. In this way, each k-element subset $A \in \mathcal{F}$ is associated with $k!(n-k)!$ permutations of X. Since $\mathcal{F}$ is a Sperner family, every permutation can only be associated with at most one member of $\mathcal{F}$. Otherwise, if two prefixes of a permutation were members of $\mathcal{F}$, then one of them would contain another, contradicting to the fact that $\mathcal{F}$ is a Sperner family. Consequently, the number of permutations generated in this way is

$$\sum_{A \in \mathcal{F}} |A|!(n-|A|)!.$$

Clearly, this number is at most the total number of permutations of X. Thus,

$$\sum_{A \in \mathcal{F}} |A|!(n-|A|)! \leq n!.$$

After dividing this inequality by $n!$, we get that

$$1 \geq \sum_{A \in \mathcal{F}} \frac{|A|!(n-|A|)!}{n!} = \sum_{A \in \mathcal{F}} \frac{1}{\binom{n}{|A|}},$$

which proves the theorem. □

If we denote by σ_k the numbers of k-element subsets in Sperner family $\mathcal{F}$, then the LYM inequality can be written as following:

$$\sum_{k=0}^{n} \frac{\sigma_k}{\binom{n}{k}} \leq 1.$$

Now, we can prove Sperner's Theorem using the LYM inequality. Note that

$$\binom{n}{\lfloor n/2 \rfloor} \geq \binom{n}{k}.$$

That is,

$$\frac{\sigma_k}{\binom{n}{\lfloor n/2 \rfloor}} \leq \frac{\sigma_k}{\binom{n}{k}},$$

from which it follows that

$$\sum_{k=0}^{n} \frac{\sigma_k}{\binom{n}{\lfloor n/2 \rfloor}} \leq \sum_{k=0}^{n} \frac{\sigma_k}{\binom{n}{k}} \leq 1.$$

So,

$$\sum_{k=0}^{n} \sigma_k \leq \binom{n}{\lfloor n/2 \rfloor}.$$

Since

$$|\mathcal{F}| = \sum_{k=0}^{n} \sigma_k,$$

we get that

$$|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor},$$

which proves Sperner's Theorem.

Appendix B

Steiner systems

Let X be an n-element set of *points*.

Definition B.1. A (n, k, λ) *design* over X is a collection of distinct k-element subsets of X, called *blocks*, such that every pair of different points is contained in exactly λ blocks.

If, more generally, for a family of distinct k-element subsets of X we have that every t-element subset of X is contained in exactly λ blocks, then the family is called a $t-(n, k, \lambda)$ *design*.

Definition B.2. A *Steiner system* $S(t, k, n)$ is a $t-(n, k, 1)$ design.

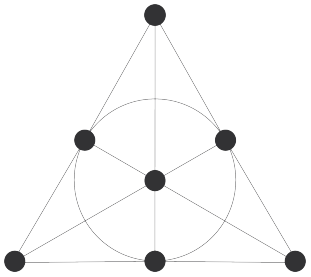
In other words, a Steiner system $S(t, k, n)$ is a collection of k-element subsets of X such that any t different points of X are in exactly one block.

Clearly, $1 \leq t \leq k \leq n$. If $t = 1$, then the family is a partition of X into k-element subsets. If $k = t$, then the family is exactly $\binom{X}{k}$. If $k = n$, then there is one block containing all n points.

Let b be the number of blocks in the collection. There are $\binom{n}{k}$ k-element subsets of X, and each of them has $\binom{k}{t}$ t-elements subsets. By the definition, every t-elements subset is contained in exactly one block. As a result, $\binom{n}{t} = b \cdot \binom{k}{t}$. So, a Steiner system $S(t, k, n)$ contains $b = \frac{\binom{n}{t}}{\binom{k}{t}}$ blocks.

Definition B.3. A *projective plane* of order q consists of a set of $q^2 + q + 1$ *points* and a family of its subsets, called *lines*, such that every line passes though $q + 1$ points and any two distinct points lie on a unique line.

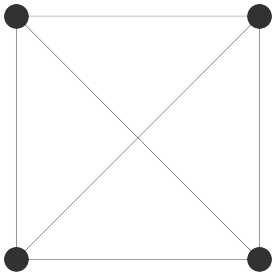
Clearly, a projective plane of order q is an $S(2, q+1, q^2+q+1)$, with lines as blocks. It is known that a projective plane of order q exists when q is a power of a prime. For $q = 1$, the only possible projective plane of order q is a triangle. For $q = 2$, the unique plane of order q is the *Fano plane*:



Definition B.4. An *affine plane* of order q consists of a set of q^2 points and a family of lines such that every line passes through q points and any two distinct points lie on a unique line.

It is easy to see that an affine plane of order q is an $S(2, q, q^2)$. The main difference of affine planes from projective planes is that they may contain "parallel" lines, that is, lines that do not intersect.

We can construct an affine plane of order q from a projective plane of the same order. In order to do this, we have to remove one block together with all of its points from the projective plane. For example, by removing a line from the Fano plane, we can get the following affine plane of order 2:



An $S(2, 3, n)$ is called a *Steiner triple system*. In this case, the blocks are called *triples*. Since $b = \frac{\binom{n}{2}}{\binom{3}{2}} = \frac{\binom{n}{2}}{3} = \frac{n(n-1)}{6}$, the family has $\frac{n(n-1)}{6}$ triples. By Kirkman [22], a Steiner triple system of order n exists if and only if $n \equiv 1$ or $3 \pmod 6$.

References

[1] Alfred V. Aho and Jeffrey D. Ullman. *Foundations of Computer Science*. W. H. Freeman & Co., New York, NY, USA, 1st edition, 1994. ISBN 0716782847.

[2] William Ward Armstrong. Dependency structures of data base relationships. In *IFIP Congress*, pages 580–583, 1974.

[3] A. Békéssy, J. Demetrovics, L. Hannk, P. Frankl, and G.O.H. Katona. On the number of maximal dependencies in a data base relation of fixed order. *Discrete Mathematics*, 30(2):83 – 88, 1980. ISSN 0012-365X. doi: 10.1016/0012-365X(80)90108-9. URL http://www.sciencedirect.com/science/article/pii/0012365X80901089.

[4] F.E. Bennett and Lisheng Wu. On minimum matrix representation of closure operations. *Discrete Applied Mathematics*, 26(1):25 – 40, 1990. ISSN 0166-218X. doi: 10.1016/0166-218X(90)90018-8. URL http://www.sciencedirect.com/science/article/pii/0166218X90900188.

[5] G. R. Blakley. Safeguarding cryptographic keys. *Managing Requirements Knowledge, International Workshop on*, 0:313, 1979. doi: http://doi.ieeecomputersociety.org/10.1109/AFIPS.1979.98.

[6] B. Bollobás. On generalized graphs. *Acta Mathematica Hungarica*, 16:447–452, 1965. ISSN 0236-5294. URL http://dx.doi.org/10.1007/BF01904851. 10.1007/BF01904851.

[7] G. Burosch, J. Demetrovics, G. O. H. Katona, D. J. Kleitman, and A. A. Sapozhenko. On the number of databases and closure operations. *Theor. Comput. Sci.*, 78(2):377–381, January 1991. ISSN 0304-3975. doi: 10.1016/0304-3975(91)90359-A. URL http://dx.doi.org/10.1016/0304-3975(91)90359-A.

[8] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387, June 1970. ISSN 0001-0782. doi: 10.1145/362384.362685. URL http://doi.acm.org/10.1145/362384.362685.

[9] J. Demetrovics, Z. Füredi, and G. O. H. Katona. Minimum matrix representation of closure operations. *Discrete Appl. Math.*, 11(2):115–128, June 1985. ISSN 0166-218X. doi: 10.1016/S0166-218X(85)80003-2. URL http://dx.doi.org/10.1016/S0166-218X(85)80003-2.

[10] J. Demetrovics, G.O.H. Katona, and D. Miklós. Recent combinatorial results in the theory of relational databases. *Mathematical and Computer Modelling*, 38(79):763 – 772, 2003. ISSN 0895-7177. doi: 10.1016/S0895-7177(03)90060-4. URL http://www.sciencedirect.com/science/article/pii/S0895717703900604. Hungarian Applied Mathematics.

[11] János Demetrovics. On the equivalence of candidate keys with sperner systems. *Acta Cybern.*, 4:247–252, 1980.

[12] János Demetrovics and Gy. Gyepesi. On the functional dependency and some generalizations of it. *Acta Cybern.*, 5:295–305, 1981.

[13] János Demetrovics and Gy. Gyepesi. A note on minimal matrix representation of closure operations. *Combinatorica*, 3(2):177–179, 1983.

[14] János Demetrovics and Gyula O. H. Katona. Extremal combinatorial problems in relational data base. In *Proceedings of the 1981 International FCT-Conference on Fundamentals of Computation Theory*, FCT '81, pages 110–119, London, UK, UK, 1981. Springer-Verlag. ISBN 3-540-10854-8. URL http://dl.acm.org/citation.cfm?id=647890.739408.

[15] János Demetrovics and Gyula O. H. Katona. Extremal combinatorial problems of database models. In *Proceedings of the 1st Symposium on Mathematical Fundamentals of Database Systems*, MFDBS '87, pages 99–127, London, UK, UK, 1988. Springer-Verlag. ISBN 3-540-19121-6. URL http://dl.acm.org/citation.cfm?id=646395.690980.

[16] János Demetrovics and Gyula O. H. Katona. A survey of some combinatorial results concerning functional dependencies in database relations. *Ann. Math. Artif. Intell.*, 7(1-4):63–82, 1993.

[17] János Demetrovics, Gyula O.H. Katona, and Attila Sali. Design type problems motivated by database theory. *Journal of Statistical Planning and Inference*, 72 (12):149 – 164, 1998. ISSN 0378-3758. doi: 10.1016/S0378-3758(98)00029-9. URL http://www.sciencedirect.com/science/article/pii/S0378375898000299.

[18] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database systems - the complete book (2. ed.)*. Pearson Education, 2009. ISBN 978-0-13-187325-4.

[19] Péter M. Gergely. Partitions with certain intersection properties. *Journal of Combinatorial Designs*, 19(5):345–354, 2011. ISSN 1520-6610. doi: 10.1002/jcd.20290. URL http://dx.doi.org/10.1002/jcd.20290.

[20] Stasys Jukna. *Extremal Combinatorics: With Applications in Computer Science*. Springer Publishing Company, Incorporated, 1st edition, 2010. ISBN 3642085598, 9783642085598.

[21] Gyula O. H. Katona. Combinatorial and algebraic results for database relations. In *Proceedings of the 4th International Conference on Database Theory*, ICDT '92, pages 1–20, London, UK, UK, 1992. Springer-Verlag. ISBN 3-540-56039-4. URL http://dl.acm.org/citation.cfm?id=645500.655897.

[22] T. P. Kirkman. On a problem in combinatorics. *Cambridge Dublin Math.*, 2:191–204, 1847.

[23] D. Lubell. A short proof of sperner's lemma. *Journal of Combinatorial Theory*, 1(2):299 –, 1966. ISSN 0021-9800. doi: 10.1016/S0021-9800(66)80035-2. URL http://www.sciencedirect.com/science/article/pii/S0021980066800352.

[24] L. D. Meshalkin. Generalization of sperner's theorem on the number of subsets of a finite set. *Theory of Probability and its Applications*, 8(2):203–204, 1963. doi: 10.1137/1108023. URL http://link.aip.org/link/?TPR/8/203/1.

[25] Imre Z. Ruzsa. Sumsets. Preprint, 2012.

[26] Adi Shamir. How to share a secret. *Commun. ACM*, 22(11):612–613, November 1979. ISSN 0001-0782. doi: 10.1145/359168.359176. URL http://doi.acm.org/10.1145/359168.359176.

[27] Emanuel Sperner. Ein satz ber untermengen einer endlichen menge. *Mathematische Zeitschrift*, 27:544–548, 1928. ISSN 0025-5874. URL http://dx.doi.org/10.1007/BF01171114. 10.1007/BF01171114.

[28] Bernhard Thalheim. On the number of keys in relational databases. In *FCT*, pages 448–455, 1987.

[29] Jeffrey D. Ullman. *Principles of Database Systems, 2nd Edition*. Computer Science Press, 1982. ISBN 0-914894-36-6.

[30] Koichi Yamamoto. Logarithmic order of free distributive lattice. *J. Math. Soc. Japan*, 6:343–353, 1954. doi: 10.2969/jmsj/00630343.