Closure operations as models of databases

Gyula O.H. Katona
MTA Rényi Institute, Budapest

Algebra Across the Borders II
Bolyai Institute, Szeged

June 19, 2012

The relational database model

a Family name	**b** Given name	**c** M or F	**d** Year of birth	**e** Month of birth	**f** Day of birth	**g** Age in years	**h** Age in months	**i** Age in days
Rózsa	Péter	F	1905	02	17	107	1288	39204
Lászó	Kalmár	M	1905	03	27			
Pál	Turán	M	1910	08	18			
György	Hajós	M	1912	02	21			
Pál	Erdős	M	1913	03	26			
Béla	Sz-Nagy	M	1913	07	29			
Alfréd	Rényi	M	1921	03	20			
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

The **types of data**, the **columns** in the table are called **attributes**.

The set of attributes is Ω. Here $\Omega = \{a, b, c, d, e, f, g, h, i, j\}$.

A **row** contains the data of a given **individual**.

Observe

$$
\begin{array}{lcl}
\{d\} & \longrightarrow & \{g\} \\
\{d,e\} & \longrightarrow & \{h\} \text{ (and } \{g\}) \\
\{d,e,f\} & \longrightarrow & \{i\} \text{ (and } \{g,h\})
\end{array}
$$

But

$$\{a\} \not\longrightarrow \{b\}$$

In general

Database: an $m \times n$ **matrix**, $|\Omega| = n$. Suppose that the rows are different.

Let $A, B \subseteq \Omega$.

B **functionally depends** on A

if the data in the columns of A **determine** the data of B,

that is there are **no two rows** which **agree** in A but different in b.

In notation: $A \longrightarrow B$.

There are many other types of **dependencies**,

but we will consider only these. Useful for reducing storage size.

Attitudes of ”**logicians**” versus ”**data mining**”.

System of functional dependencies,

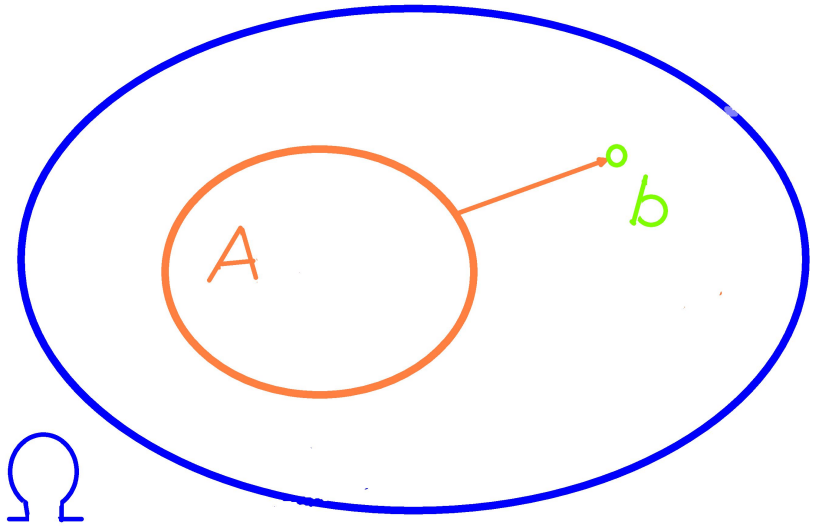
Armstrong axioms.

Our approach:

look at **only** the dependencies $A \longrightarrow b$ where $b \in \Omega$ is only one column.

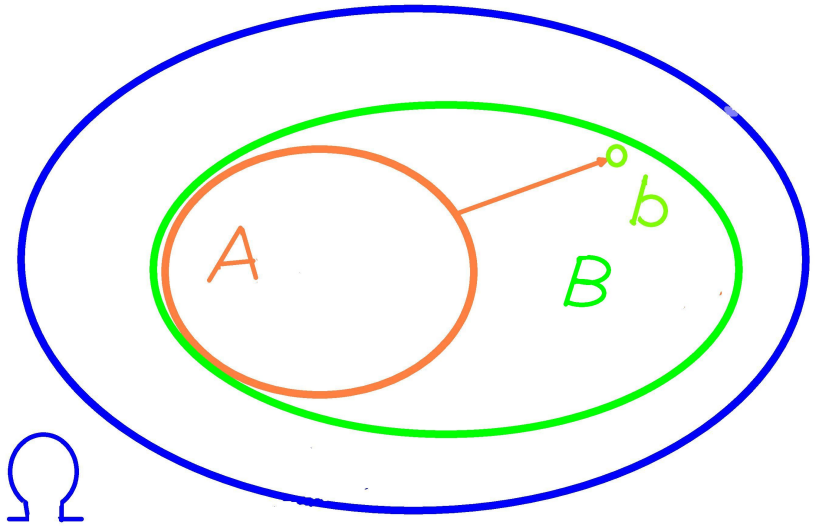
Set function $\mathcal{C}_M : 2^{\Omega} \longrightarrow 2^{\Omega}$ on the subsets of Ω:

$$\mathcal{C} = \mathcal{C}_M(A) = \{b \colon b \in \Omega,\ A \longrightarrow b\}.$$



Set function $\mathcal{C}_M : 2^{\Omega} \longrightarrow 2^{\Omega}$ on the subsets of Ω:

$$\mathcal{C} = \mathcal{C}_M(A) = \{b : b \in \Omega,\ A \longrightarrow b\}.$$



B=$\mathcal{C}_M(A)$

It has **three properties:**

$$
\begin{aligned}
&A \subseteq \mathcal{C}(A),\\
&A \subseteq B \implies \mathcal{C}(A) \subseteq \mathcal{C}(B),\\
&\mathcal{C}(\mathcal{C}(A)) = \mathcal{C}(A).
\end{aligned}
$$

A set function satisfying these properties is called a **closure operation**.

Theorem **(Armstrong, Demetrovics)**

For any closure $\mathcal{C}$ there exists a matrix M such that

$$\mathcal{C}_M = \mathcal{C}.$$

Forgetting about other rules, dependencies

and

the actual content of the database,

this **closure operation** (shortly **closure**)

can be considered as

the model of the database.

Minimum matrix representation of a closure:

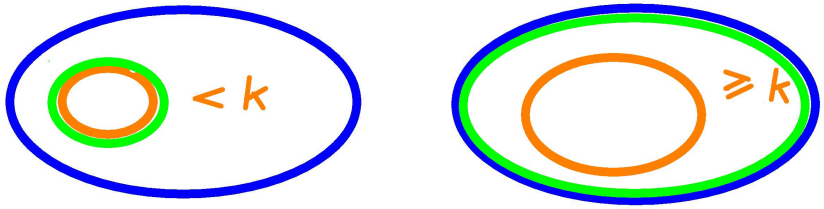
$$s(\mathcal{C}) = \min_{M\,:\,\mathcal{C}_M = \mathcal{C}} \{\text{number of rows in } M\}.$$

Minimum matrix representation of a closure:

$$s(\mathcal{C}) = \min_{M\,:\,\mathcal{C}_M=\mathcal{C}} \{\text{number of rows in } M\}.$$

Too difficult in general! **Special case:**

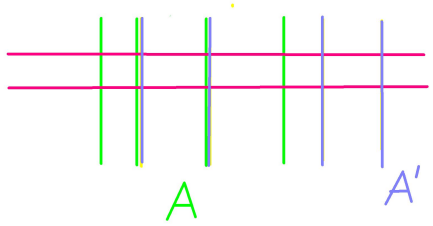
$$\mathcal{C}_n^k(A) = \begin{cases} A & \text{if } |A| < k \\ \Omega & \text{otherwise.} \end{cases}$$



Lemma Demetrovics-K, 1981)

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

Proof M represents $\mathcal{C}_n^k$ and has $s(\mathcal{C}_n^k)$ rows. $|A| = |A'| = k-1$ distinct subsets of Ω and they determine the same pair of rows



$|A \cup A'| \geq k$, the two rows are equal here, but not everywhere,

a contradiction.

Theorems

$s(\mathcal{C}_n^1) = 2$ **(trivial) (DK, 1981)**,

$s(\mathcal{C}_n^2) = \left\lceil \frac{1+\sqrt{1+8n}}{2} \right\rceil$ **(easy) (DK, 1981)**,

$s(\mathcal{C}_n^{n-1}) = n$ **(easy) (DK, 1981)**,

$s(\mathcal{C}_n^n) = n+1$ **(easy) (DK, 1981)**,

$s(\mathcal{C}_n^3) = n$ **(difficult) (D-Füredi-K, 1985;**

Bennett-Wu, 1990; Ganter-Gronau, 1991)

(**Design theory** is used.)

$c_1(k)n^{\frac{k-1}{2}} < s(\mathcal{C}_n^k) < c_2(k)n^{\frac{k-1}{2}}$ **(difficult) (D-Füredi-K, 1985)**

Closely related to **Shamir's secret sharing** in **cryptology**.

Difficult results of **Krisztián Tichler** for the case when $\mathcal{C} \subset \binom{\Omega}{2}$.

A model of changing databases

Adding new rows, it might **destroy** functional dependencies.

A model of changing databases

Adding new rows, it might **destroy** functional dependencies.

$$A \longrightarrow \{b\}$$

A					b
0	1	0	1	$1\ldots$	0
0	1	0	1	$1\ldots$	0
0	1	1	0	$1\ldots$	1
0	1	1	0	$1\ldots$	1

A model of changing databases

Adding new rows, it might **destroy** functional dependencies.

$$A \not\longrightarrow \{b\}$$

A					b
0	1	0	1	1 . . .	0
0	1	0	1	1 . . .	0
0	1	1	0	1 . . .	1
0	1	1	0	1 . . .	1
0	1	1	0	1 . . .	**0**

A model of changing databases

Adding new rows, it might **destroy** functional dependencies.

$$A \not\longrightarrow \{b\}$$

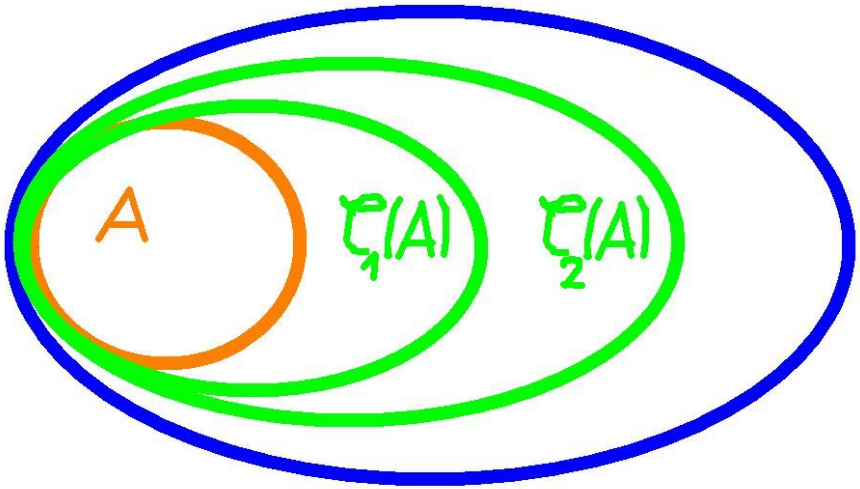
A					b
0	1	0	1	1 . . .	0
0	1	0	1	1 . . .	0
0	1	1	0	1 . . .	1
0	1	1	0	1 . . .	1
0	1	1	0	1 . . .	**0**

Deleting rows, it might **create** new dependencies.

A model of changing databases

Consider only closures satisfying $\mathcal{C}(\emptyset) = \emptyset$.

$\mathcal{C}_1$ is **reacher** than $\mathcal{C}_2$ if $\mathcal{C}_1(A) \subseteq \mathcal{C}_2(A)$ holds for every $A \subseteq \Omega$.



A model of changing databases

Consider only closures satisfying $\mathcal{C}(\emptyset) = \emptyset$.

$\mathcal{C}_1$ is **reacher** than $\mathcal{C}_2$ if $\mathcal{C}_1(A) \subseteq \mathcal{C}_2(A)$ holds for every $A \subseteq \Omega$.

This is a **partially ordered set** $\mathbb{P}$. (Transitive!)

A model of changing databases

Consider only closures satisfying $\mathcal{C}(\emptyset) = \emptyset$.

$\mathcal{C}_1$ is **reacher** than $\mathcal{C}_2$ if $\mathcal{C}_1(A) \subseteq \mathcal{C}_2(A)$ holds for every $A \subseteq \Omega$.

This is a **partially ordered set** $\mathbb{P}$. (Transitive!)

The **richest** one is in which $\mathcal{C}(A) = A$ holds for every $A \subseteq \Omega$.

The **least rich** one satisfies $\mathcal{C}(A) = \Omega$ for every $A \subseteq \Omega$.

A model of changing databases

A is **closed** if $\mathcal{C}(A) = A$.

The **family of closed sets** is denoted by $\mathcal{Z} = \mathcal{Z}(\mathcal{C})$.

It satisfies

(i) $\emptyset \in \mathcal{Z}$,

(ii) $A, B \in \mathcal{Z}$ implies $A \cap B \in \mathcal{Z}$.

A model of changing databases

A is **closed** if $\mathcal{C}(A) = A$.

The **family of closed sets** is denoted by $\mathcal{Z} = \mathcal{Z}(\mathcal{C})$.

It satisfies

(i) $\emptyset \in \mathcal{Z}$,

(ii) $A, B \in \mathcal{Z}$ implies $A \cap B \in \mathcal{Z}$.

It is easy to see that for any such $\mathcal{Z}$

there is a **unique closure** $\mathcal{C}$ with $\mathcal{Z}(\mathcal{C}) = \mathcal{Z}$.

A model of changing databases

A is **closed** if $\mathcal{C}(A) = A$.

The **family of closed sets** is denoted by $\mathcal{Z} = \mathcal{Z}(\mathcal{C})$.

It satisfies

(i) $\emptyset \in \mathcal{Z}$,

(ii) $A, B \in \mathcal{Z}$ implies $A \cap B \in \mathcal{Z}$.

It is easy to see that for any such $\mathcal{Z}$

there is a **unique closure** $\mathcal{C}$ with $\mathcal{Z}(\mathcal{C}) = \mathcal{Z}$.

Lemma $\mathcal{C}_1$ is **reacher** than $\mathcal{C}_2$ iff $\mathcal{Z}(\mathcal{C}_1) \supseteq \mathcal{Z}(\mathcal{C}_2)$

A model of changing databases

A is **closed** if $\mathcal{C}(A) = A$.

The **family of closed sets** is denoted by $\mathcal{Z} = \mathcal{Z}(\mathcal{C})$.

It satisfies

(i) $\emptyset \in \mathcal{Z}$,

(ii) $A, B \in \mathcal{Z}$ implies $A \cap B \in \mathcal{Z}$.

It is easy to see that for any such $\mathcal{Z}$

there is a **unique closure** $\mathcal{C}$ with $\mathcal{Z}(\mathcal{C}) = \mathcal{Z}$.

Lemma $\mathcal{C}_1$ is **reacher** than $\mathcal{C}_2$ iff $\mathcal{Z}(\mathcal{C}_1) \supseteq \mathcal{Z}(\mathcal{C}_2)$

This gives an **equivalent definition** of the poset $\mathbb{P}$.

A model of changing databases

The **rank function** r of a poset associates

a **non-negative integer** with every element of the poset

with the following properties.

$r(a) = 0$ for some element.

If $a < b$ in the poset and there is no c satisfying $a < c < b$ then $r(b) = r(a)+1$.

A model of changing databases

The **rank function** r of a poset associates

a **non-negative integer** with every element of the poset

with the following properties.

$r(a) = 0$ for some element.

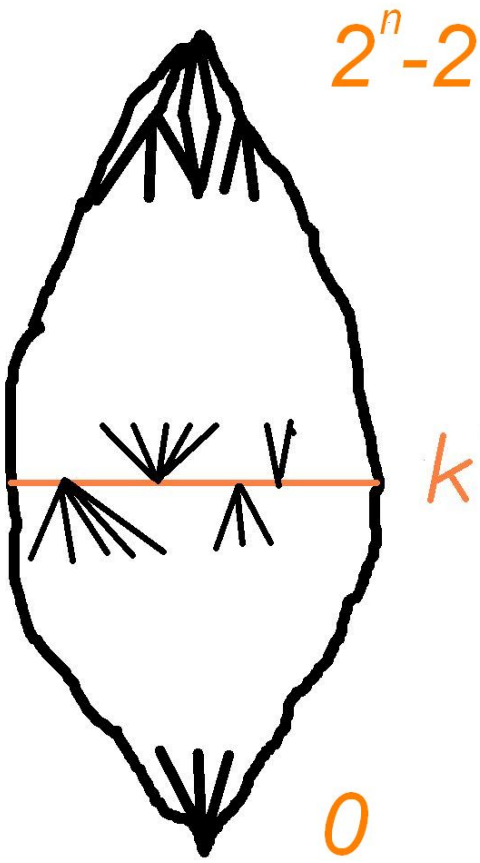
If $a < b$ in the poset and there is no c satisfying $a < c < b$ then $r(b) = r(a)+1$.

The **poset** $\mathbb{P}$ of closures has a rank function:

$$r(\mathcal{C}) = r(\mathcal{Z}(\mathcal{C})) = |\mathcal{Z}| - 2.$$

Minimum rank =0, **maximum** rank = $2^n - 2$.

A model of changing databases



A model of changing databases

Proposition

If some rows are added to the (database) matrix then the closure $\mathcal{C}_M$ will move up in the poset,

conversely, if rows are omitted then the closure will move down in the poset.

Therefore poset $\mathbb{P}$ can be considered as a model of a changing database.

Theorem (Burosch-D-K-Kleitman-Sapozhenko, 1991)

The number $\alpha(n)$ of closures on an n-element set Ω is

$$2^{\binom{n}{\lfloor n/2 \rfloor}} < \alpha(n) < 2^{2\sqrt{2}\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}.$$

Theorem (Burosch-D-K-Kleitman-Sapozhenko, 1991)

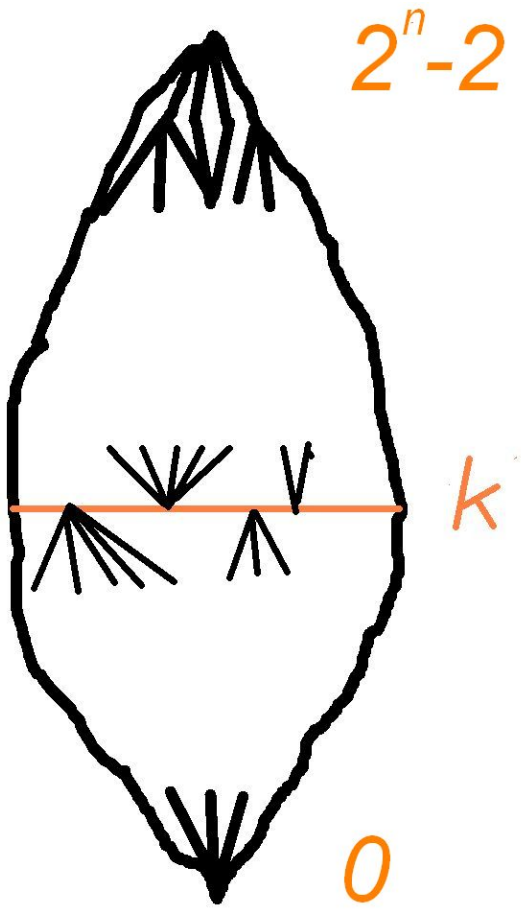
The number $\alpha(n)$ of closures on an n-element set Ω is

$$2^{\binom{n}{\lfloor n/2 \rfloor}} < \alpha(n) < 2^{2\sqrt{2}\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}.$$

Theorem (Alekseev, 1992)

$$\alpha(n) = 2^{\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}.$$

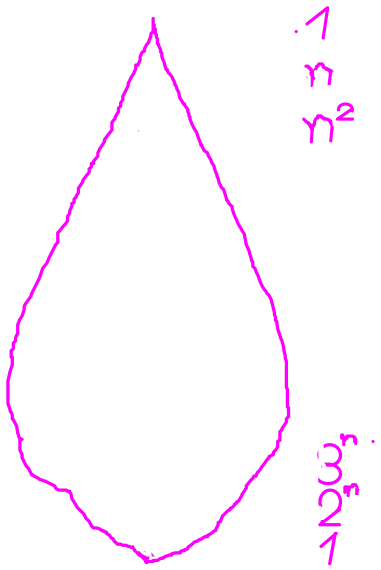
$\alpha(n, k)$ is the number of closures on the kth level.



Theorem (Burosch-D-K-Kleitman-Sapozhenko, 1993)

$$\alpha(n,k) \sim \eta(k)(k+1)^n,$$

$$\alpha(n, 2^n - k - 2) \sim \theta(k) n^k.$$



Open

Which is the **largest level** in $\mathbb{P}$?

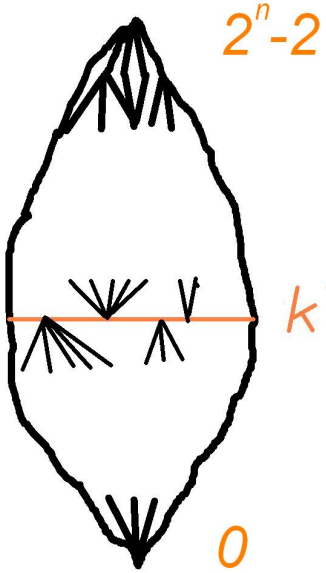
Determine it at least asymptotically.

$$f_1(n,k) = \max\{\deg_{\text{up}}(\mathcal{C}) : \ r(\mathcal{C}) = k\},$$

$$f_2(n,k) = \min\{\deg_{\text{up}}(\mathcal{C}) : \ r(\mathcal{C}) = k\},$$

$$f_3(n,k) = \max\{\deg_{\text{down}}(\mathcal{C}) : \ r(\mathcal{C}) = k\},$$

$$f_4(n,k) = \min\{\deg_{\text{down}}(\mathcal{C}) : \ r(\mathcal{C}) = k\}.$$



Theorem (Burosch-D-K, 1987)

$f_1(n,k) = 2^n - k - 2,$

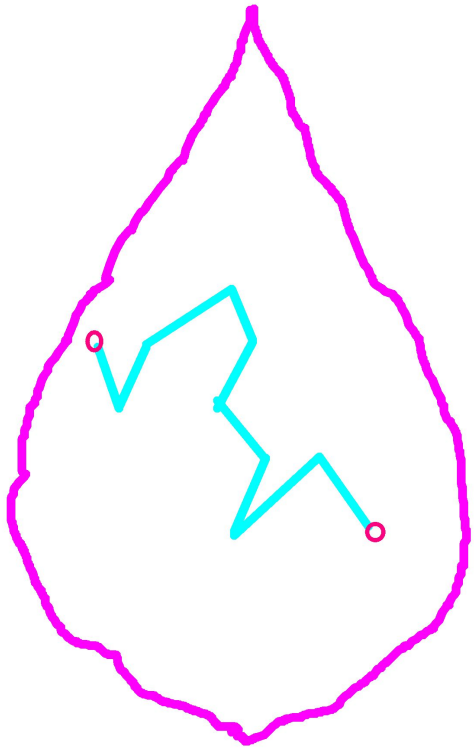
$f_2(n,k) = 0$ or 1 or 2, exact, but complicated conditions determined.

$f_4(n,k)$ is nearly exactly determined as $\log_2(k+1)$

Open

Nothing is known about $f_3(n,k)$.

Distance of two databases is defined as the **shortest path** in the Hasse diagram in this partially ordered set $\mathbb{P}$.



K-Sali (2010) determined the **maximum distance**

between databases for given n.

Open What is the maximum distance between the kth and ℓth levels?

What is a **random closure**?

One **possible** definition:

choose the entries of the **“generating matrix”** randomly.

Random databases

All entries in the matrix are chosen **totally independently**

with probabilities $q_1, q_2, \ldots, q_d$.

$H_2(q_1, q_2, \ldots, q_d) = -\log(q_1^2 + q_2^2 + \ldots + q_d^2)$ is the

Rényi entropy of second order.

What are the typical sizes for $|A|$ for which

$A \longrightarrow b$ holds with high probability?

Theorem (D-K-Miklós-Seleznev-Thalheim, 1998)

Let M be a random $m \times n$ matrix.

If A is a set of columns and $|A|$ is **somewhat larger** than

$$\frac{2 \log_2 n}{H_2(q_1, q_2, \ldots, q_d)}$$

then $A \longrightarrow b$ holds with high probability for every column b.

If it is **somewhat smaller** then the functional dependency does not hold.

Theorem (D-K-Miklós-Seleznev-Thalheim, 1998)

Let M be a random $m \times n$ matrix.

If A is a set of columns and $|A|$ is **somewhat larger** than

$$\frac{2 \log_2 n}{H_2(q_1, q_2, \ldots, q_d)}$$

then $A \longrightarrow b$ holds with high probability for every column $b \notin A$.

If it is **somewhat smaller** then the functional dependency does not hold.

Generalized (K, 2010) for the case when the columns

have **different** probability distributions.

If $q_1 = q_2 = \frac{1}{2}$ then $H_2 = 1$.

$b \in \mathcal{C}(A)(b \notin A)$ holds with high probability iff $|A| \geq 2 \log n$.

If $q_1 = q_2 = \frac{1}{2}$ then $H_2 = 1$.

$b \in \mathcal{C}(A)(b \notin A)$ holds with high probability iff $|A| \geq 2 \log n$.

More work is needed for

average size of $\mathcal{C}(A)$, given $|A|$,

average size of **closed sets**,

distribution of sizes of **closed sets**,

etc.

Error-correcting keys

A **key** K in a database is a set of attributes (columns)

uniquely **determining** the individual (row).

In terms of the closure: $\mathcal{C}(K) = \Omega$.

It is a **minimal key** if no proper subset is a key,

that is, if $K' \subset K, K' \neq K$ then $\mathcal{C}(K) \neq \Omega$.

Error-correcting keys

There are **wrong** data.

Suppose there is **at most one error** in the data of one individual.

$$\text{correct data} \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0$$

Error-correcting keys

There are **wrong** data.

Suppose there is **at most one error** in the data of one individual.

$$\text{correct data} \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0$$

$$\text{erronous data} \quad 0 \quad 1 \quad \mathbf{1} \quad 0 \quad 1 \quad 0$$

Error-correcting keys

There are **wrong** data.

Suppose there is **at most one error** in the data of one individual.

$$\text{correct data} \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0$$

$$\text{erronous data} \quad 0 \quad 1 \quad \mathbf{1} \quad 0 \quad 1 \quad 0$$

The key K **might not** determine the row.

But a larger set $K \subset L$ **might**.

This is called a **minimal 1-error correcting key**.

Error-correcting keys

How much larger can/must be C than K?

Theorem (D-K-Miklós, 2000)

If the sizes of the minimal keys are at most k

then the sizes of the minimal error correcting keys

cannot exceed $c_2(k)k^3$,

and there is an example

when one of them has size at least $c_1(k)k^3$.

Closure $\mathcal{C}$ is given.

Functional dependency $A \longrightarrow B$ holds iff $B \subset \mathcal{C}(A)$.

Closure $\mathcal{C}$ is given.

Functional dependency $A \longrightarrow B$ holds iff $B \subset \mathcal{C}(A)$.

$A_1 \longrightarrow B_1, A_2 \longrightarrow B_2, \ldots A_m \longrightarrow B_m$ are called **independent**,

if non of them **can be deduced** from the other $m-1$ ones.

Closure $\mathcal{C}$ is given.

Functional dependency $A \longrightarrow B$ holds iff $B \subset \mathcal{C}(A)$.

$A_1 \longrightarrow B_1, A_2 \longrightarrow B_2, \ldots A_m \longrightarrow B_m$ are called **independent**,

if non of them **can be deduced** from the other $m - 1$ ones.

Problem Find the maximum number

of independent functional dependencies.

Have a guess!

Closure $\mathcal{C}$ is given.

Functional dependency $A \longrightarrow B$ holds iff $B \subset \mathcal{C}(A)$.

$A_1 \longrightarrow B_1, A_2 \longrightarrow B_2, \ldots A_m \longrightarrow B_m$ are called **independent**,

if non of them **can be deduced** from the other $m-1$ ones.

Problem Find the maximum number

of independent functional dependencies.

Have a guess!

You will see the answer in the lecture of **Dezső Miklós**

Thank you for

listening my talk

in this weather!