Enumeration Aspects of Databases: Functional Dependencies and Informative Armstrong Relations

Simon Vilmin[1], Jean-Marc Petit[2]

Journées "Graphes et Bases de Données"

March 2023

[1] LORIA, Univ Lorraine, CNRS, INRIA, Nancy
[2] LIRIS, INSA Lyon, UCBL, CNRS, Lyon

Data and their semantics

- A relation r : a collection of tuples t_i over a set $\boxed{R}$ of attributes → Relation schema
- Find knowledge in the data : Find functions between attributes

$$f(X) = A \qquad X \subseteq R,\ A \in R$$

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

$AB \rightarrow D$

$BC \rightarrow D$ No !

Find Knowledge* ⟷ Functional Dependencies (FDs) $X \rightarrow A$

*In our case

Outline

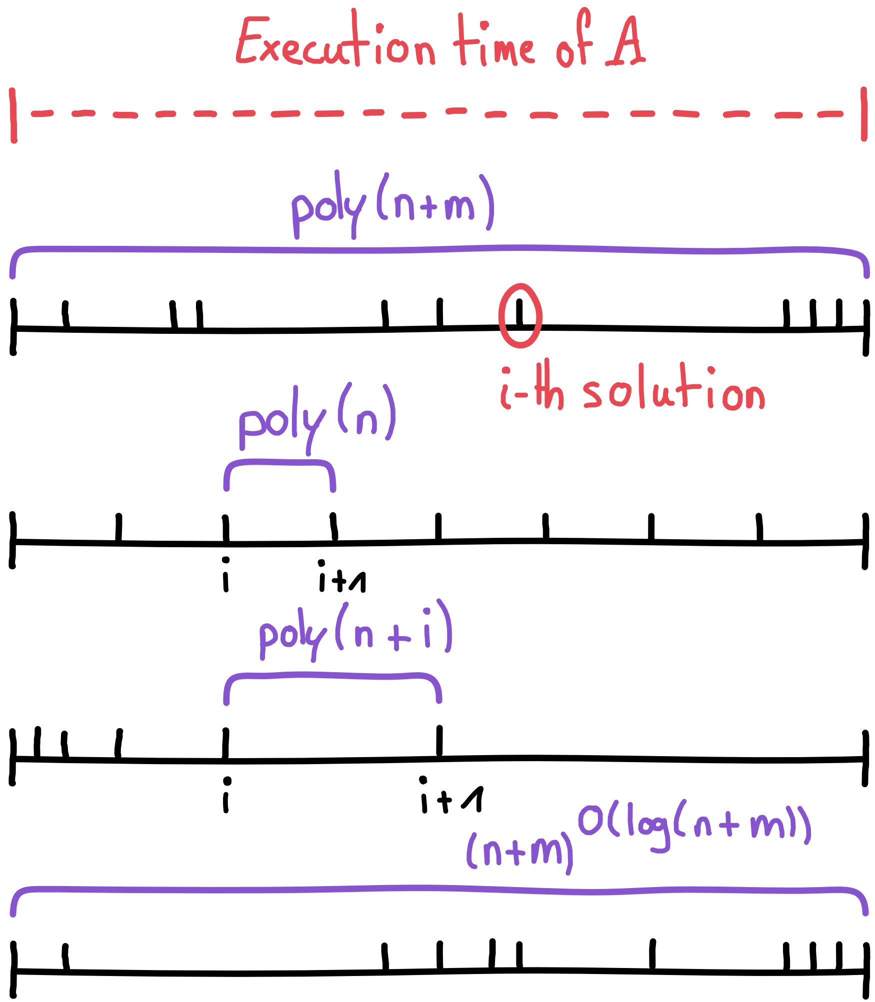
- Objective: understand the FDs holding in the data
- PART I : Find them explicitly
  - What does it mean ?
  - What for ?
  - Complexity ?
- PART II : The data already summarizes the knowledge
  - Informative Armstrong relations (IARs)
  - Preliminary results on enumeration

Enumeration

- Enumeration task: given an input x, list a set of solutions $R(x)$

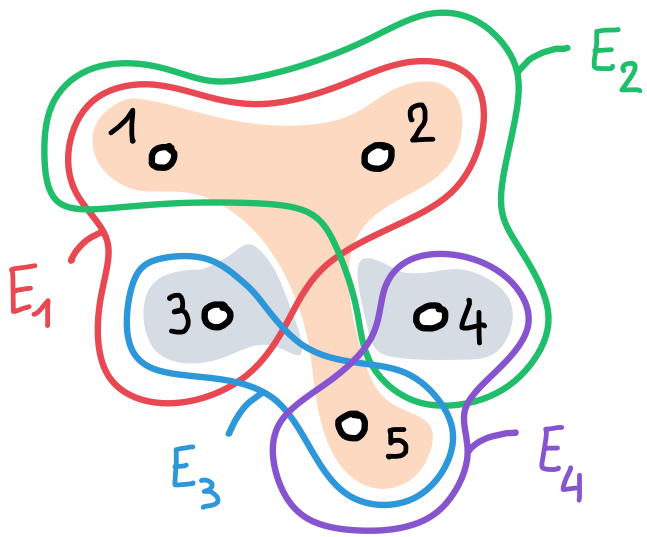
of size poly(x)

Enumeration algorithm A
x of size n, $R(x)$ of size m



- output-polynomial time
- polynomial delay
- incremental polynomial time
- output quasi-polynomial time

Hypergraphs



- $\mathcal{H} = (\mathcal{V} = \{1, \dots, 5\}, \{E_1, E_2, E_3, E_4\})$:

  $E_1 = 123$, $E_2 = 124$, $E_3 = 34$, $E_4 = 45$

- Transversal $T \subseteq \mathcal{V}$: $T \cap E_i \neq \emptyset$ for every E_i

- Independent set $I \subseteq \mathcal{V}$: $E_i \not\subseteq I$ for every E_i

PROB. Enum Minimal Transversals (Enum-MTR) $\quad$ * $E_i \not\subseteq E_j$ $\forall i, j$

Input: a (simple*) hypergraph $\mathcal{H} = (\mathcal{V}, \mathcal{E})$

Task: enumerate the inclusion-wise minimal transversals of $\mathcal{H}$, $MTR(\mathcal{H})$

- Open problem, quasi-poly algorithm [Fredman, Khachiyan, 1996]
- Equivalent to Enum-MIS: listing the maximal independent sets of $\mathcal{H}$, $MIS(\mathcal{H})$

PART I. FINDING FUNCTIONAL DEPENDENCIES

Functional Dependencies (FDs)

DEF. A Functional dependency (FD) over R is an expression $X \to Y$ where $X, Y \subseteq R$.

DEF. Let r be a relation over R and $X \to Y$ a FD over R. The FD $X \to Y$ holds in r, written $r \models X \to Y$, if for every $t_1, t_2 \in r$

$$t_1[X] = t_2[X] \text{ implies } t_1[Y] = t_2[Y].$$

If Σ is a set of FDs, $r \models \Sigma$ means $r \models X \to Y$ for all $X \to Y \in \Sigma$

r	A	B	C	D
t_1	1	1	1	1
t_2	1	1	2	2
t_3	2	1	2	3
t_4	3	2	2	3

- $r \models \{A \to B, D \to C\}$
- $r \not\models C \to B$

Our problem: given r, find the FDs $X \to Y$ s.t. $r \models X \to Y$

Do we want all of them?

- Do we really need all FDs?
  - $r \models X \to Y$ trivially holds if $Y \subseteq X$ — $X \to Y$
  - $r \models \{X \to Y, Y \to Z\}$ entails $r \models X \to Z$ — $X \to Z$
  - $r \models X \to Z$ implies $r \models X \cup Y \to Z$ — $X \cup Y \to Z$

Useless

We can deduce FDs from others, and it does not depend on the choice of r

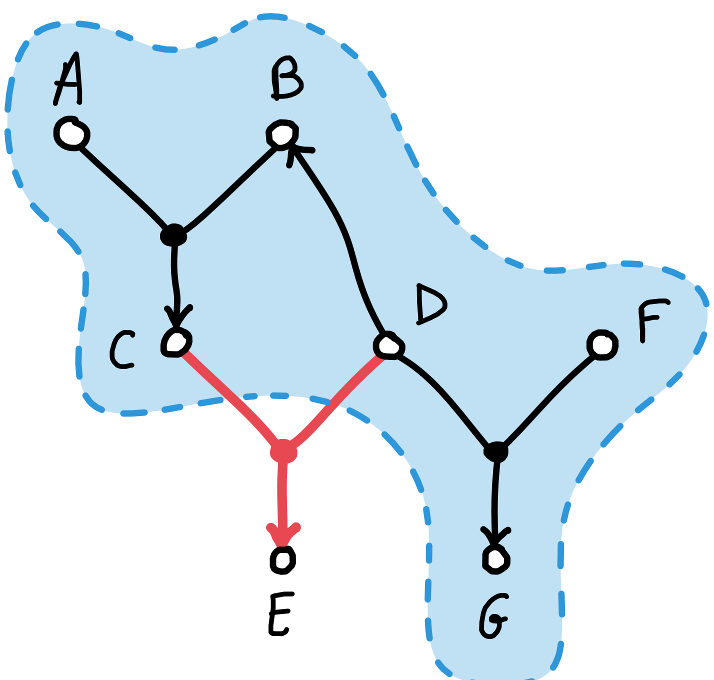
DEF. Let Σ be a set of FDs over R, and let $X \to Y$ be another FD. We say that $X \to Y$ follows from Σ, written $\Sigma \models X \to Y$, if for every relation r over R,

$$r \models \Sigma \text{ implies } r \models X \to Y$$

Closure algorithm

- Deciding $\Sigma \models X \to Y$: implication problem
- To solve it : closure procedure → Forward chaining / Transitive closure
  - takes $X \subseteq R$ as input, returns the closure $\phi(X)$ of X wrt Σ
  - builds $X = X_0, \dots, X_m = \phi(X)$ s.t. $X_i = X_{i-1} \cup \bigcup \{ Y \mid Z \to Y \in \Sigma,\ Z \subseteq X_{i-1} \}$

Prop. $\Sigma \models X \to Y$ iff $Y \subseteq \phi(X)$



- $\Sigma = \{ AB \to C,\ D \to B,\ CD \to E,\ DF \to G \}$

$X = X_0 = ADF$

$X_1 = ADFBG$

$X_2 = ADFBGC$

$X_3 = ADFBGCE$

- Two sets of FDs can be different but equivalent

DEF. Let Σ_1, Σ_2 be sets of FDs over R. We say that Σ_1 follows from Σ_2, written $\Sigma_2 \models \Sigma_1$, if $\Sigma_2 \models X_1 \to Y_1$ for all $X_1 \to Y_1 \in \Sigma_1$. We say that Σ_1 and Σ_2 are equivalent if $\Sigma_1 \models \Sigma_2$ and $\Sigma_2 \models \Sigma_1$.

- Thus, there are sets of FDs "better than others":

  (1) Σ is a nonredundant cover if $\Sigma \setminus X \to Y \not\models \Sigma$ for every $X \to Y \in \Sigma$

  (2) Σ is a minimum cover if it has the least possible number of FDs

  (3) Σ is an optimum cover if $\sum_{X \to Y \in \Sigma} |X| + |Y|$ is minimal among all equiv. Σ'

- (3) $\Rightarrow$ (2) $\Rightarrow$ (1) but (3) hard to optimize, while (1), (2) poly (from Σ)

  [Ausiello et al., 1986]

Back to the problem

PROB. Minimum Cover

Input: a relation r over R

Task: find a minimum cover Σ of the FDs satisfied by r

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

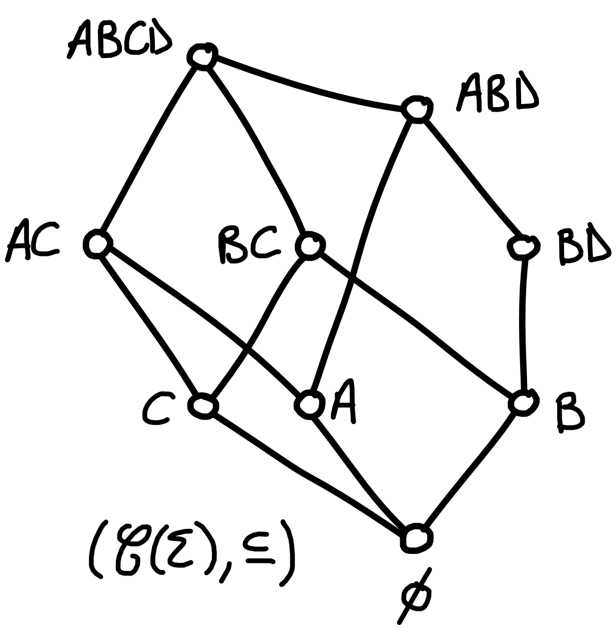
- How do we know we are done?
  - $r \models \{AB \to D, D \to B\}$
  - we also have $r \models CD \to A$

DEF. Let Σ be a set of FDs over R. A relation r over R is an Armstrong relation for Σ if for every FD $X \to Y$ over R

$$r \models X \to Y \text{ iff } \Sigma \models X \to Y$$

Def. A closure system is a pair $(R, \mathcal{C})$ where R is a set and $\mathcal{C} \subseteq 2^R$ s.t. $R \in \mathcal{C}$ and $X_1, X_2 \in \mathcal{C}$ implies $X_1 \cap X_2 \in \mathcal{C}$

- Given Σ, $\mathcal{C}(\Sigma) = \{\phi(X) \mid X \subseteq R\}$ is a closure system (with R)
- Every closure system can be represented by sets of FDs



$\Sigma = \{D \to B,\ CD \to A,\ AB \to D\}$

Prop. For $Z \subseteq R$, $Z \in \mathcal{C}(\Sigma)$ iff $X \subseteq Z$ entails $Y \subseteq Z$ for each FD $X \to Y$ of Σ

Σ represents a closure system
$\to$ an Armstrong relation for Σ represents the same closure system

Def. Let r be a relation over R and $t_1, t_2 \in R$. The agree set of t_1, t_2 is

$$ag(t_1, t_2) = \{A \in R \mid t_1[A] = t_2[A]\}$$

The agree sets of r are denoted $ag(r)$

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

$ag(t_2, t_3) = ABD$

$ag(r) = \{\emptyset, A, B, C, AC, BC, BD, ABCD\}$

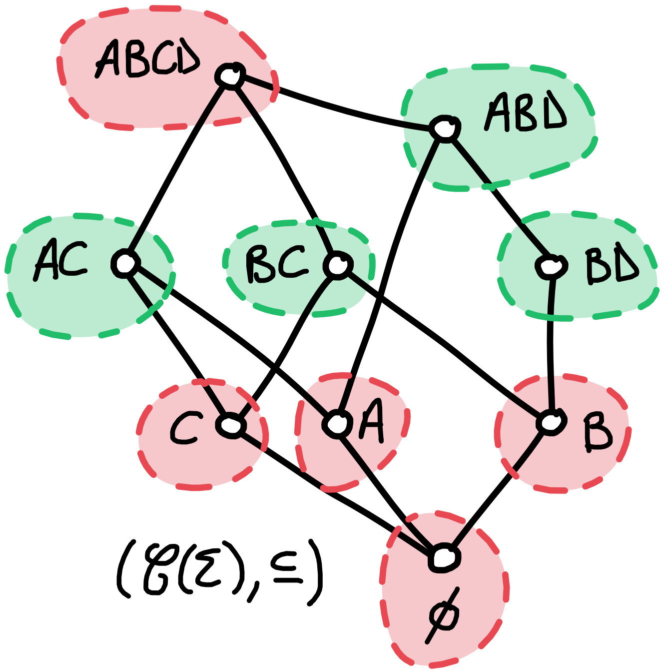
- Rewriting: $r \models X \to Y$ iff for each $Z \in ag(r)$, $X \subseteq Z$ implies $Y \subseteq Z$

  $\to$ every agree set satisfies $X \to Y$

- Going further: $r \models \Sigma$ means that each agree set satisfies each FD of Σ

Prop. If r is an Armstrong relation for Σ, then $ag(r) \subseteq \mathcal{C}(\Sigma)$

What is the minimal amount of information (elements) from $\mathcal{C}(\Sigma)$ we need to store in a relation to obtain an Armstrong relation for Σ?



- A closure system $(R, \mathcal{C})$ is closed under intersection
  - R is trivially in $\mathcal{C}$
  - Some sets are obtained by intersecting others
  - Some are not, they are irreducible

Def. Let $(R, \mathcal{C})$ be a closure system and let $M \in \mathcal{C}$, $M \neq R$. Then,
M is meet-irreducible if $M = X_1 \cap X_2$ implies $M = X_1$ or $M = X_2$ for all $X_1, X_2 \in \mathcal{C}$
$Mi(\mathcal{C})$ is the set of meet-irreducible elements of $(R, \mathcal{C})$

Meet-irreducibles and agree sets

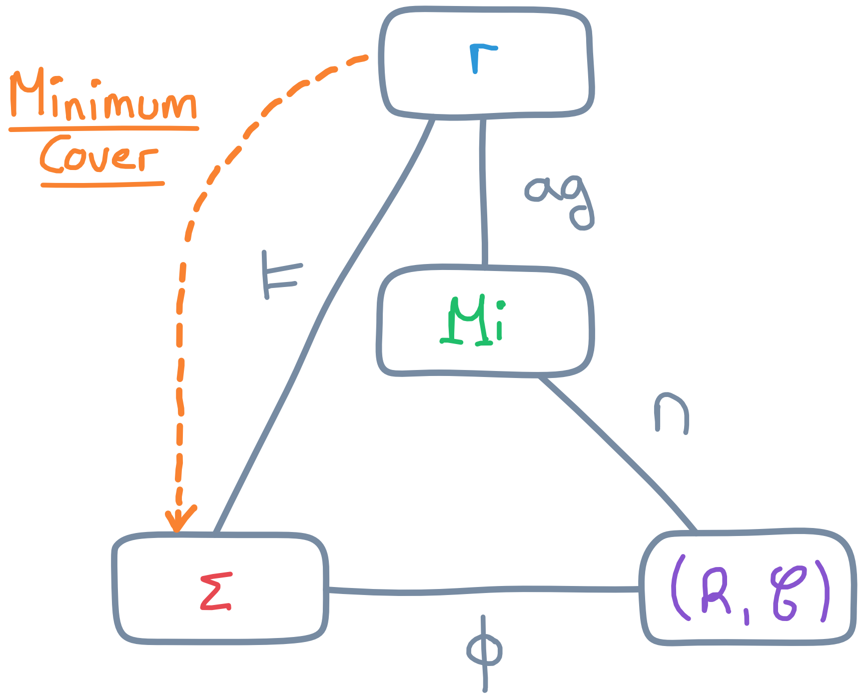
means $\mathcal{M}i(\mathcal{C}(\Sigma))$

Given Σ over R, $\mathcal{M}i(\Sigma)$ is the minimal amount of information needed to reconstruct $\mathcal{C}(\Sigma)$ by intersections

THM. [Beeri et al., 1984] Let Σ be a set of FDs over R, and let r be a relation over R. Then, r is an Armstrong relation for Σ iff

$$\mathcal{M}i(\Sigma) \subseteq ag(r) \subseteq \mathcal{C}(\Sigma)$$

Packing up



ag. A relation r over R defines some meet-irreducible elements Mi

∩. Mi defines a closure system (R, 𝒞)

ϕ. The closure system (R, 𝒞) can be represented by a set Σ of FDs

⊨. Σ represents the FDs of r

Minimum Cover is the problem of finding an alternative representation of a closure system

On the example

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

ag

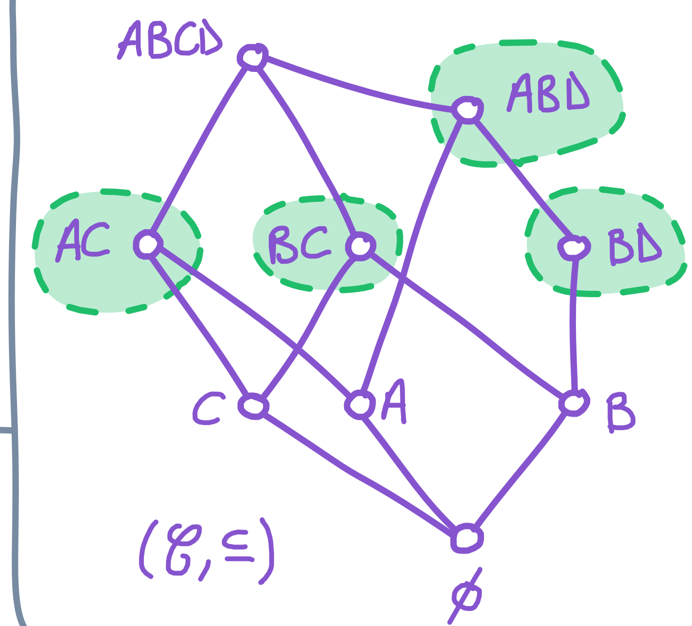
$Mi = \{AC, BC, ABD, BD\}$

$\cap$

$\models$

$\Sigma = \{D \rightarrow B, CD \rightarrow A, AB \rightarrow D\}$

ϕ



Closure systems are ubiquitous

- Closure systems arise from numerous objects / fields
  - Lattice theory
  - Knowledge space theory
  - Pure Horn CNF
  - Formal Concept Analysis
  - Points in $\mathbb{R}^n$
  - matroids
  - graph convexities (geodesic, monophonic)
  - posets (ideals, convex sets)
  - Argumentation Frameworks
  - ...

- Minimum Cover appears in disguise in many fields
- Closure systems coming from special objects may have special interesting properties for Minimum Cover

- What is the size of Σ wrt r in general ?
  - Σ can have size exponential in the size of r
  - r can have size exponential in the size of Σ

- The complexity of some problems depends on the representation

Problem	Σ	r
Enumerating minimal Keys	poly-delay	quasi-poly
Does A belong to a minimal Key	NP-c	poly

(minimal) Key: (minimal) subset K of R which determines everyone, i.e. $K \rightarrow R$ holds

At last, complexity!

Prob. <u>Minimum Cover</u>

Input: a relation r over R

Task: find a minimum cover Σ of the FDs satisfied by r

- Surveys [Bertet et al., 2018], [Wild, 2017]
- Negative side
  - Unknown complexity ...
  - Harder than <u>Enum-MTR</u> [Khardon, 1995]
- Positive side
  - (Exponential) algorithms [Mannila, Räihä, 1992], [Wild, 1995]
  - Tractable cases [Beaudou et al., 2017], [Defrain et al., 2021]

Summary

- Minimum Cover : find a small set of FDs representing the Knowledge in the data
- Goes well beyond databases : it is a matter of representing closure systems
  - appears in Logic, Formal Concept Analysis, Knowledge spaces, ...
  - connections with graphs, posets, matroids, geometries, ...
- But the problem is tough ...
  - unknown complexity (for more than 30 years)
  - harder than Enum-MTR
- The same goes for the dual problem $\Sigma \rightsquigarrow r$!
- Main idea : find particular closure systems
  - graph convexities ?
  - case where Σ has no "cycle" ?

Part II. Informative Armstrong Relations

We love FDs, but...

FDs have drawbacks

- hard to find
- possibly much larger than the data
- not all of them are meaningful

Maybe find another representation ... such as the data itself!

Find a "small" subset of tuples faithfully representing the semantics (FDs) of the data

$\Rightarrow$ informative Armstrong relations

Example

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

$$\Sigma = \{D \to B,\ AB \to D,\ CD \to A\}$$

s_1	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_7	5	1	2	9
t_8	6	3	3	8

$$\Sigma_1 = \{D \to B,\ AB \to D,\ CD \to A\}$$

s_2	A	B	C	D
t_1	3	3	3	3
t_4	3	4	3	4
t_5	7	4	7	4
t_8	6	3	3	8

$$\Sigma_2 = \{D \to B,\ C \to A,\ AB \to D,\ CD \to A\}$$

?

Informative Armstrong Relations

DEF. Let r be a relation over R. A subrelation $s \subseteq r$ is an *Informative Armstrong relations (IAR)* for r if it satisfies exactly the same FDs as r.

Why are they interesting?

- condensed representation of the data
- understanding which FDs are relevant

Previous works are mostly experimental [Bisbal, Grimson, 2001] [De Marchi, Petit, 2007], [Wei, Link, 2018]

First problem, First observations

PROB. Minimum IAR

Input: a relation r over R, $k \in \mathbb{N}$

Question: does r contain an IAR s such that $|s| \leq k$?

Remarks:

Closure system of r

- $ag(t_1, t_2) = \{A \in R \mid t_1[A] = t_2[A]\}$
- s is an Armstrong relation for r iff $Mi(r) \subseteq ag(s) \subseteq \mathcal{C}(r)$
- $s \subseteq r$ implies $ag(s) \subseteq \mathcal{C}(r)$

meet-irreducible elements of r

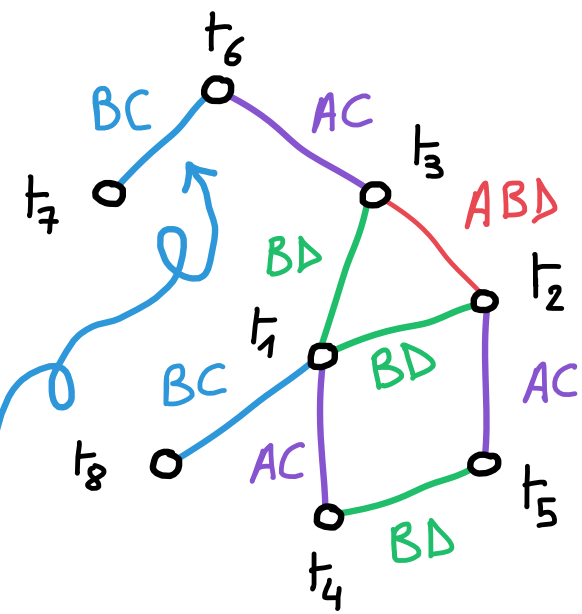
The subrelation s is an IAR for r $\longleftrightarrow$ $Mi(r) \subseteq ag(s)$

A graph of tuples and irreducibles

r	A	B	C	D
t_1	3	3	3	3
t_2	7	3	7	3
t_3	7	3	2	3
t_4	3	4	3	4
t_5	7	4	7	4
t_6	7	1	2	7
t_7	5	1	2	9
t_8	6	3	3	8

$ag(t_6, t_7) = BC$

$BC \in Mi(r)$



$Mi(r) = \{AC, BD, ABD, BC\}$

IARs and graph coloring

Consider the edge-colored graph $G_r = (r, E)$ of the relation r with:

- $(t_1, t_2) \in E$ exactly when $ag(t_1, t_2) \in Mi(r)$
- (t_1, t_2) is given the color $ag(t_1, t_2)$

Colors are exactly $Mi(r)$

$$\boxed{\underline{\text{Minimum IAR}} \longleftrightarrow \text{find a "small" induced subgraph of } G_r \text{ with all the colors!}}$$

Precision:

- For $s \subseteq r$, $G_r[s] = (s, E(s))$ with $E(s) = \{(t_1, t_2) \in E \mid t_1, t_2 \in s\}$
- $G_r[s]$ subgraph of G_r induced by s

Meanwhile, in bioinformatics

PROB. Minimum Rainbow Subgraph (MRS)

Input: a graph $G = (V, E)$ where each edge is given a color in $\{1, \ldots, m\}$, $k \in \mathbb{N}$

Question: is there a subgraph of G with at most k vertices and exactly one edge of each color?

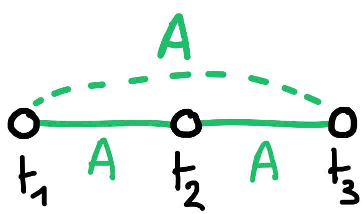
needs not be induced ⚠

Comes from bioinformatics [Bafna et al., 2003], [Catanzaro & Labbé, 2009]

- MRS is **NP**-complete [Camacho et al., 2010]
- most results are approximations [Popa 2014], [Camacho et al., 2010]

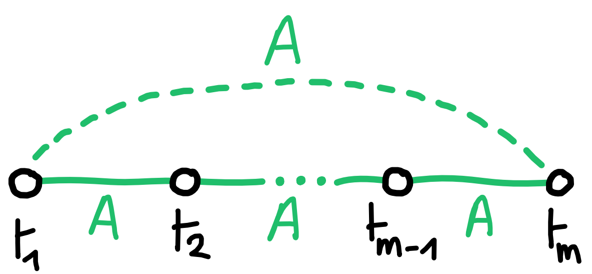
⇝ Minimum IAR particular case of MRS

Properties of IARs



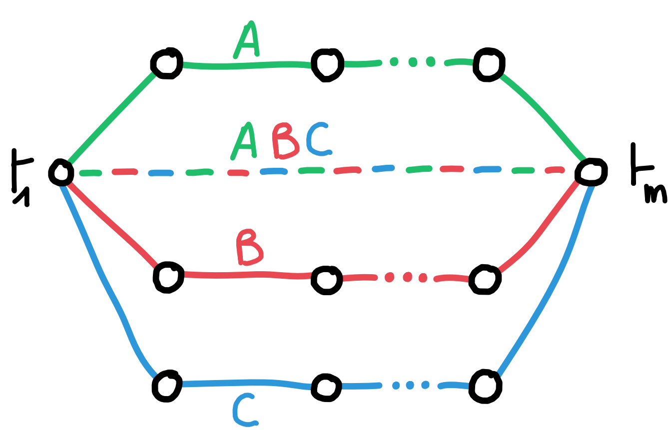
$$t_1[A] = t_2[A] = t_3[A] \Rightarrow t_1[A] = t_3[A]$$

Transitivity of equality



$$t_1[A] = t_2[A] = \cdots = t_m[A] \Rightarrow t_1[A] = t_m[A]$$

Transitivity extends to paths

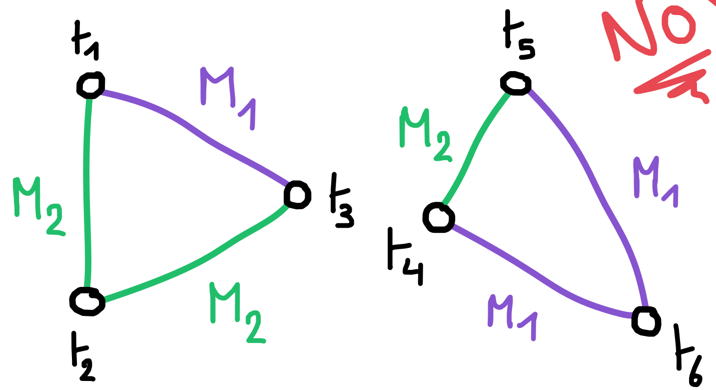


Prop. For every sequence $t_1 = u_1, \ldots, u_m = t_2$ giving a path from t_1 to t_2, we have:

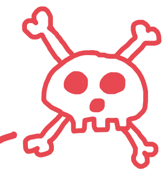
$$\bigcap_{1 \leq i \leq m} ag(u_i, u_{i+1}) \subseteq ag(t_1, t_2)$$

Consequences

The graph G_r has some forbidden patterns



No!

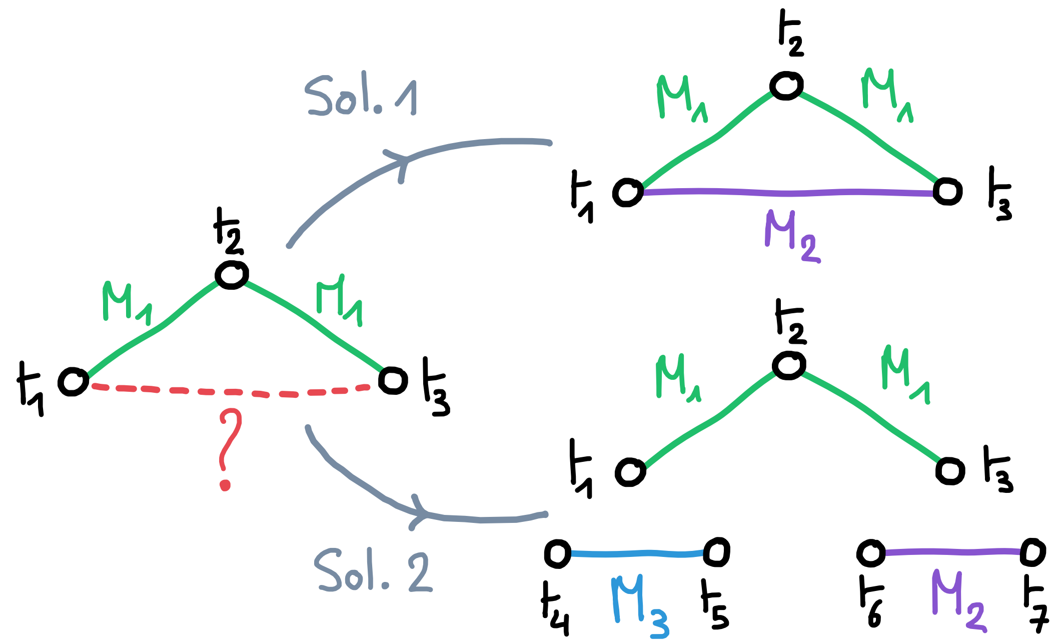


- Hyp : $M_1 \neq M_2$
- Due to t_1, t_2, t_3, $M_1 \leq M_2$ holds
- Due to t_4, t_5, t_6, $M_2 \leq M_1$ holds

- Hyp : $ag(t_1, t_2) = ag(t_2, t_3) = M_1$
- Problem : $ag(t_1, t_3)$?

$\rightsquigarrow$ Sol. 1 : $ag(t_1, t_3) = M_2$

$\rightsquigarrow$ Sol. 2 : $ag(t_1, t_3) = M_2 \cap M_3$



THM. (Petit, V.) The problem Minimum IAR is **NP-complete**.

What about (inclusion-wise) minimal IARs?

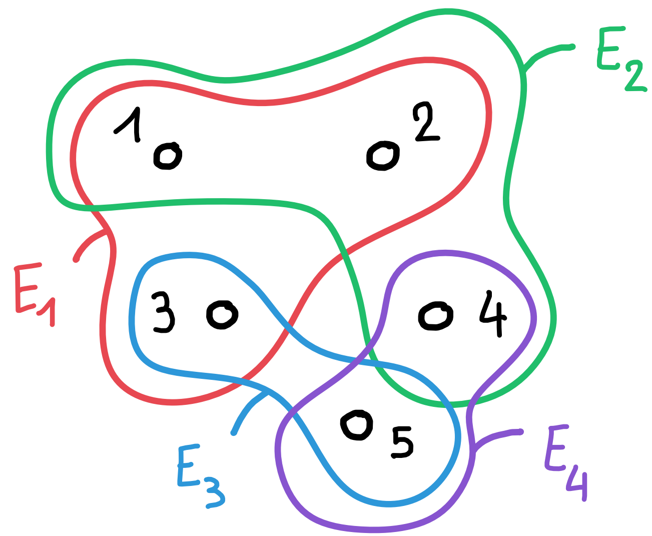
- IARs are closed under taking supersets
- Testing IAR property is easy

⇝ Find a minimal IAR for r: greedy approach

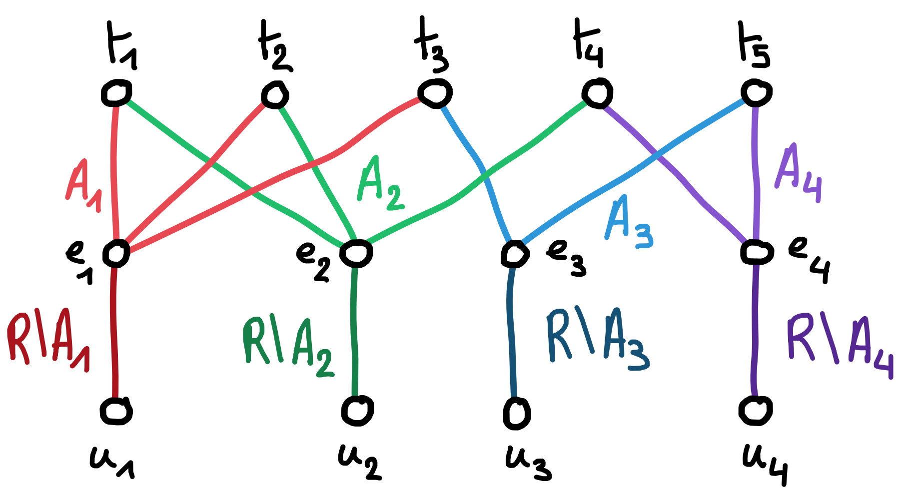
PROB. Enumerating Minimal IAR (Enum-MIAR)

Input: a relation r over R

Task: enumerating the inclusion-wise minimal IARs for r



- $\mathcal{H} = (\mathcal{V} = \{1, \ldots, 5\}, \{E_1, E_2, E_3, E_4\})$:
  $E_1 = 123$, $E_2 = 124$, $E_3 = 34$, $E_4 = 45$



- G_r : incidence bipartite graph of $\mathcal{H}$
  - $R = \{A_1, A_2, A_3, A_4, C\}$
  - $Mi(r) = \{A_1, A_2, A_3, A_4,$
    $R \setminus A_1, R \setminus A_2, R \setminus A_3, R \setminus A_4\}$
  - $t_i[C] = t_j[C]$

$$\boxed{MTR(\mathcal{H}) \longleftrightarrow \min_{\subseteq}(IARs)}$$

THM. (Petit, V.) The problem <u>Enum-MIAR</u> is harder than <u>Enum-MTR</u>

Further remarks on the reduction:

- Bipartite graph
- FDs easy to find

Adapting the reduction to SAT:

THM. (Petit, V.) Let r be a relation over R, and let $t \in r$. It is **NP**-complete to decide whether t belongs to a minimal IAR for r.

- Informative Armstrong relations (IARs) summarize the data

- But their structure seems rather complex
  - hard to find a minimum IAR
  - hard to decide if a tuple belongs to a minimal IAR
  - enumerating minimal IAR is at least quasi-poly

- Perhaps ...
  - restrict the underlying closure system ?
  - restrict the graph of meet-irreducible elements ?

Thank you for your attention!

References Part I

- Ausiello, D'Atri, Saccà [Ausiello et al., 1986]
  Minimal representations of directed hypergraphs
  SIAM Journal on Computing, 1986
- Beaudou, Mary, Nourine [Beaudou et al., 2017]
  Algorithms for k-meet-semidistributive lattices
  Theoretical Computer Science, 2017
- Bertet, Demko, Viaud, Guérin [Bertet et al., 2018]
  Lattices, closure systems and implication bases: A survey of structural aspects and algorithms
  Theoretical Computer Science, 2018
- Defrain, Nourine, Vilmin [Defrain et al., 2021]
  Translating between the representations of a ranked convex geometry
  Discrete Mathematics, 2021.
- Beeri, Dowd, Fagin, Statman [Beeri et al. 1984]
  On the structure of Armstrong relations for functional dependencies
  Journal of the ACM, 1984

References Part I

- Fredman, Khachiyan [Fredman, Khachiyan, 1996]
  On the complexity of dualization of monotone disjunctive normal forms
  Journal of Algorithms, 1996
- Khardon [Khardon, 1995]
  Translating between Horn representations and their Characteristic Models
  Journal of Artificial Intelligence Research, 1995
- Mannila, Räihä [Mannila, Räihä, 1992]
  The Design of Relational Databases
  Addison-Wiley, 1992
- Wild [Wild, 1995]
  Computations with finite closure systems and Implications
  Springer LNCS 959, 1995
- Wild [Wild, 2017]
  The joy of implications, aka pure Horn formulas: mainly a survey
  Theoretical Computer Science, 2017

References Part II

- Bisbal, Grimson [Bisbal, Grimson, 2001]
  Database sampling with functional dependencies
  Information and Software Technology, 2001
- Bafna, Gusfield, Lancia, Shibu [Bafna et al., 2003]
  Haplotyping as perfect phylogeny: A direct approach
  Journal of Computational Biology, 2003
- Catanzaro, Labbé [Catanzaro & Labbé, 2009]
  The pure parsimony haplotyping problem: Overview and computational advances
  International Transactions in Operational Research, 2009
- De Marchi, Petit [De Marchi, Petit, 2007]
  Semantic sampling of existing databases through informative Armstrong relations
  Information Sciences, 2007

References Part II

- Popa [Popa 2014]
  Better lower and upper bounds for the minimum rainbow subgraph problem
  Theoretical Computer Science, 2014
- Camacho, Schiermeyer, Tuza [Camacho et al., 2010]
  Approximation algorithms for the minimum rainbow subgraph problem
  Discrete Mathematics, 2010
- Wei, Link [Wei, Link, 2018]
  DataProf: semantic profiling for iterative data cleansing and business rule acquisition
  Proceedings of the 2018 International Conference on Management on Data, 2018