Functional Dependencies with Predicates: What Makes the g_3-error Easy to Compute?

Simon Vilmin [1]
Pierre Faure--Giovagnoli [2,3]
Jean-Marc Petit [2]
Vasile-Marian Scuturici [2]

[1] LIS, Aix-Marseille Université
[2] LIRIS, INSA Lyon
[3] Compagnie Nationale du Rhône

Data vs. Domain Knowledge

F	E	P
2.5	10.1	22.9
2.7	10.4	23.2
2.6	10.3	23.0
2.5	10.2	23.3
2.6	10.1	23.1
2.6	10.3	22.9

(Unique) counterexample

Data from a hydropower turbine:

- incoming flow F ($m^3 . s^{-1}$)
- elevation E of the waterfall (m)
- power P produced (MW)

Domain knowledge:

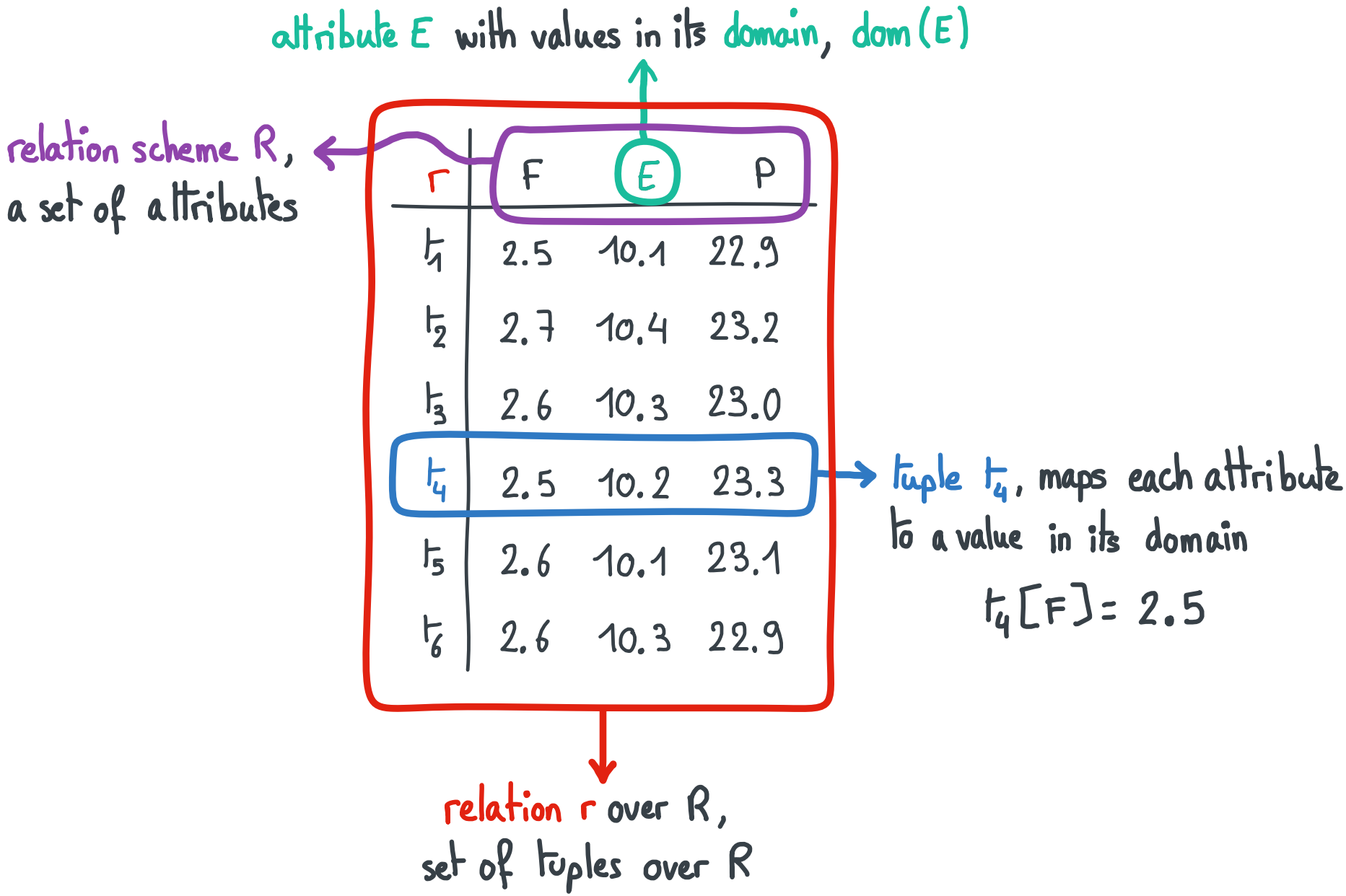
P is determined by E and F,

i.e. $P = f(E, F)$

almost!

Question: is knowledge supported by data?

Some Database Terminology



Domain Knowledge and Functional Dependencies

Question : is knowledge supported by data ?

↓

P determined by E and F

↓

Function $f(E, F) = P$

↓

Functional Dependency (FD) $EF \rightarrow P$

↓

Question : does the FD hold in the relation ?

Functional Dependencies

Syntax

DEF: a Functional Dependency (FD) over R is an expression $X \to A$ where $X \subseteq R$ and $A \in R$.

Semantic

DEF: given r and $X \to A$ over R, $X \to A$ holds in r, $r \models X \to A$, if $\forall t_1, t_2 \in r$, $t_1[X] = t_2[X]$ entails $t_1[A] = t_2[A]$

r_1	B	C	A	D
t_1	0	0	1	a
t_2	0	0	2	b
t_3	0	1	1	c

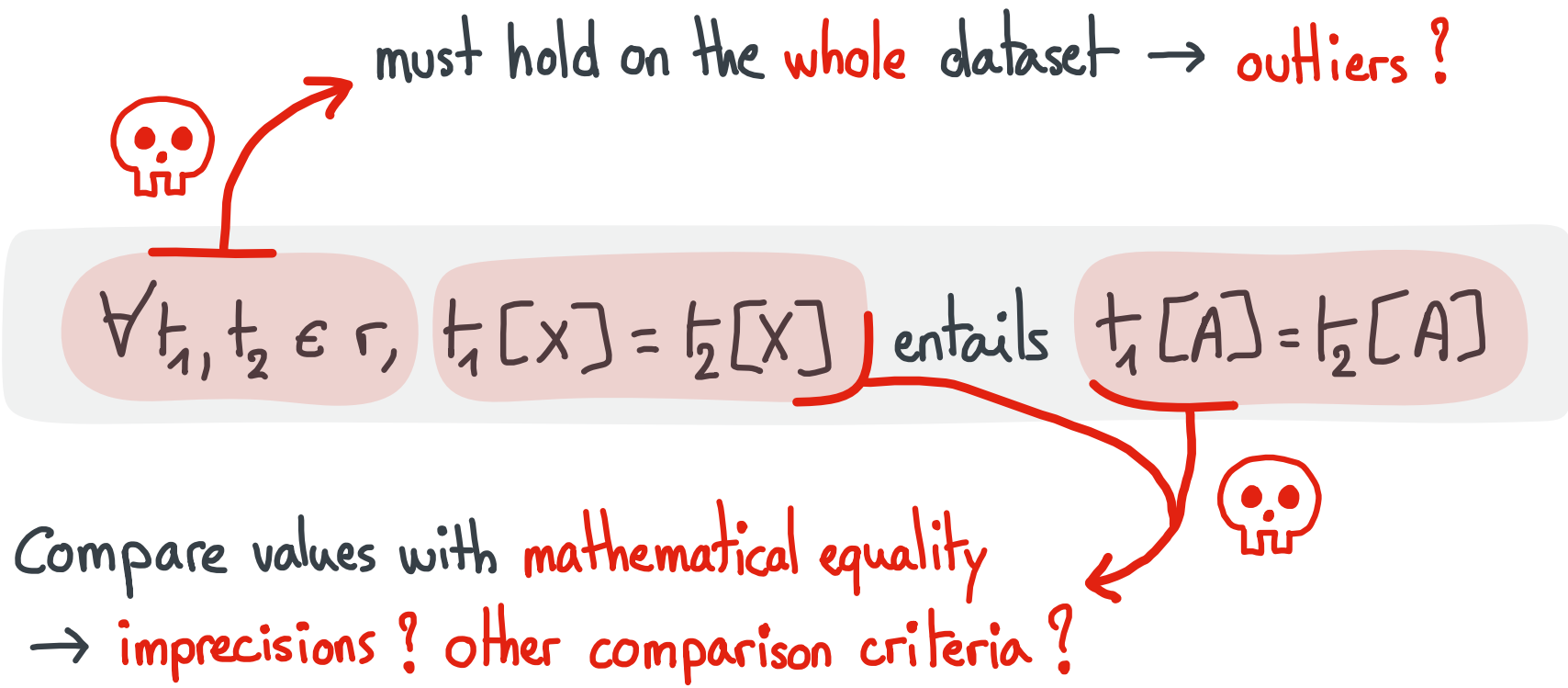
(t_1, t_2) Counterexample

$r_1 \not\models BC \to A$

r_2	B	C	A	D
t_1	0	1	1	a
t_2	0	1	1	b
t_3	1	2	3	c

$r_2 \models BC \to A$

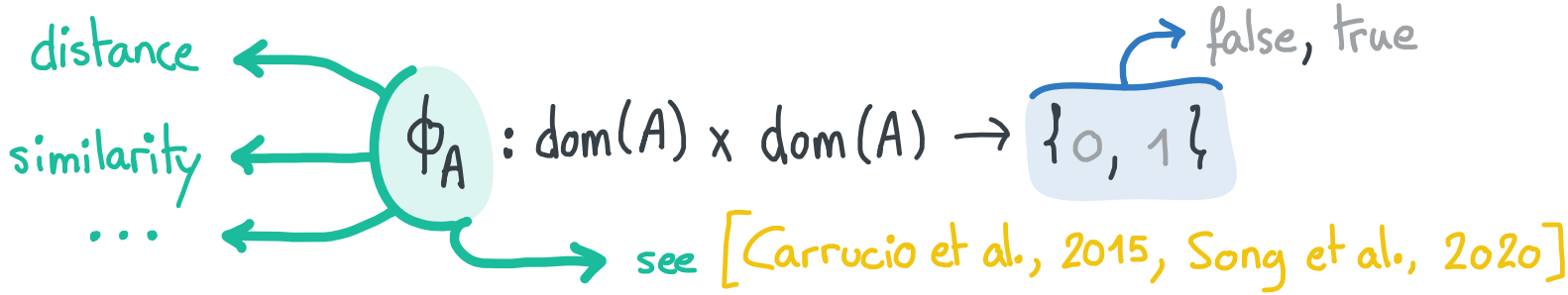
FDs vs. Real Life



Ideas : - replace equality with (binary) predicates
- use a coverage measure

Predicates to relax equality

binary predicate ϕ_A for $A \in R$: predicate to compare values in $dom(A)$



Relation scheme with predicates (R, Φ): Φ set of predicates, one for each $A \in R$

Semantic

DEF: given r and $X \rightarrow A$ over (R, Φ), $X \rightarrow A$ holds in r wrt Φ, written $r \models_\Phi X \rightarrow A$, if $\forall t_1, t_2 \in r$,

$$\bigwedge_{B \in X} \phi_B(t_1[B], t_2[B]) = 1 \text{ implies } \phi_A(t_1[A], t_2[A]) = 1$$

The g_3-error with Predicates

g_3-error : coverage measure for FDs with equality [Kivinen, Mannila, 1995]

minimum proportion of tuples to remove from r to satisfy $X \to A$

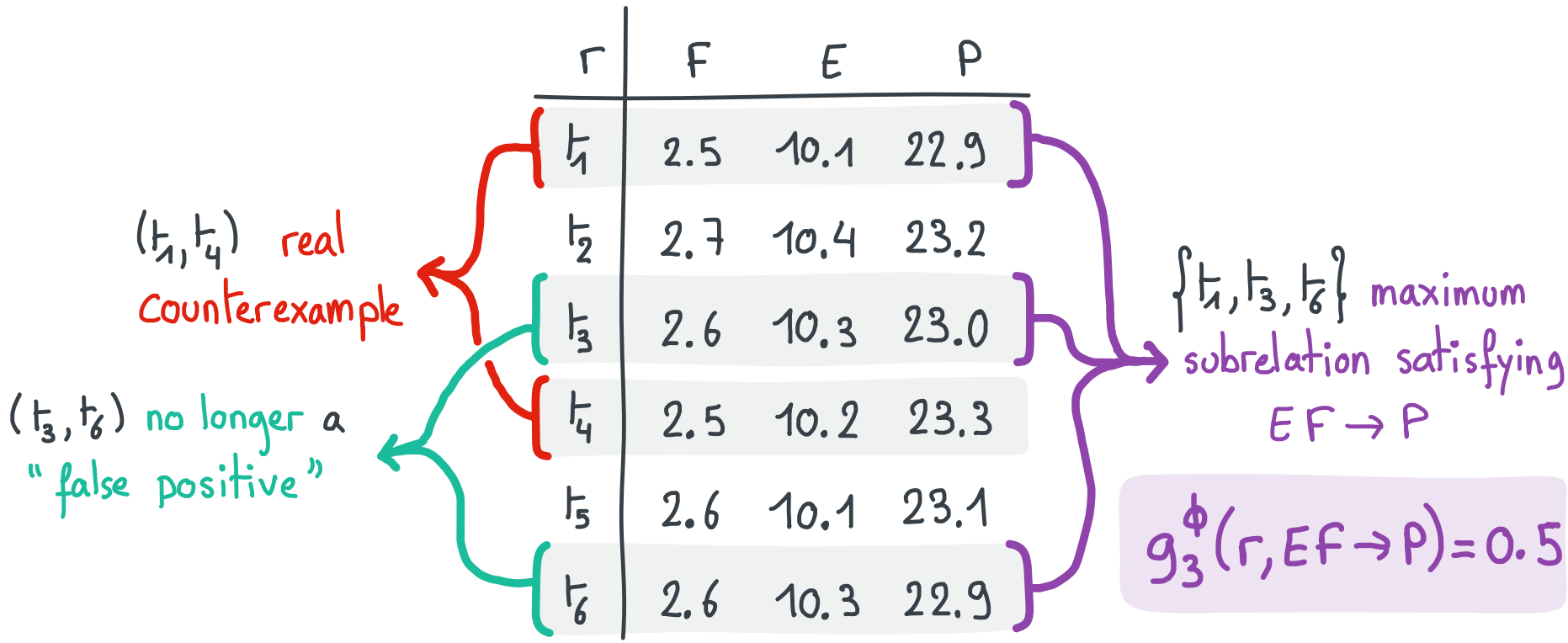
adapted to predicates [Faure--Giovagnoli et al., 2022]

DEF: Let (R, Φ) be a relation scheme with predicates and let r, $X \to A$ be a relation and a FD over (R, Φ). The g_3-error of $r, X \to A$ wrt Φ is:

$$g_3^{\Phi}(r, X \to A) = 1 - \frac{\max(\{|s| : s \subseteq r,\ s \vDash_{\Phi} X \to A\})}{|r|}$$

size of the largest subrelation satisfying $X \to A$

Back to the Example



$$\phi_P = \phi_E = \phi_F \qquad \phi_P(x,y) = 1 \iff |x-y| \leq 0.1$$

$$\Phi = \{\phi_P, \phi_E, \phi_F\}$$

Back to our Problem

Question : is knowledge supported by data ?

↓ function $\leftrightarrow$ FDs

Question : does the FD hold in the relation ?

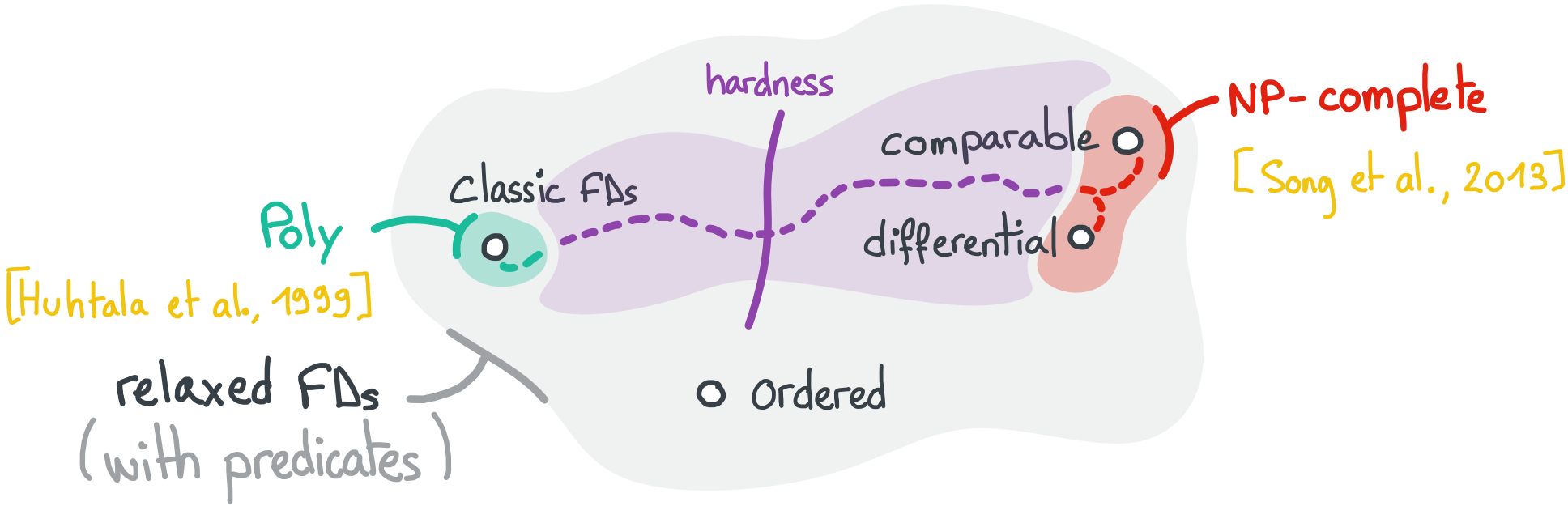
↓ $g_3 + \Phi$

PROB : Error Validation Problem with Predicates (EVPP)

In: a relation scheme with predicates (R, Φ), r and $X \rightarrow A$ over (R, Φ), $k \in \mathbb{R}$

Out: YES if $g_3^{\Phi}(r, X \rightarrow A) \leq k$, NO otherwise

The complexity of EVPP



Question: what makes EVPP tractable or not?

Idea: study predicate properties

(ref) $\phi_A(x,x) = 1$ (sym) $\phi_A(x,y) = 1 \Rightarrow \phi_A(y,x) = 1$

(tra) $\phi_A(x,y) = \phi_A(y,z) = 1 \Rightarrow \phi_A(x,z) = 1$

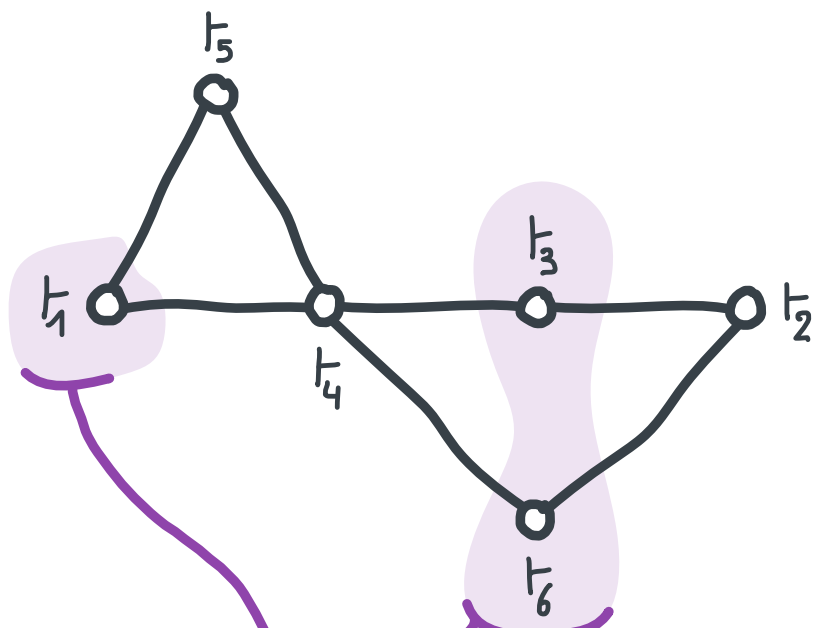
(asym) $\phi_A(x,y) = \phi_A(y,z) = 1 \Rightarrow x = y$

Conflict-graphs

Conflict-graph $CG_{\Phi}(r, EF \to P)$

[Bertossi, 2011]

r	F	E	P
t_1	2.5	10.1	22.9
t_2	2.7	10.4	23.2
t_3	2.6	10.3	23.0
t_4	2.5	10.2	23.3
t_5	2.6	10.1	23.1
t_6	2.6	10.3	22.9



$s \models_{\Phi} EF \to P \iff s$ independent set of $CG_{\Phi}(r, EF \to P)$

EVPP and Maximum Independent Sets

Question: what makes EVPP tractable or not?

↓ CG_{Φ}

Prob: Maximum Independent Set (MIS)

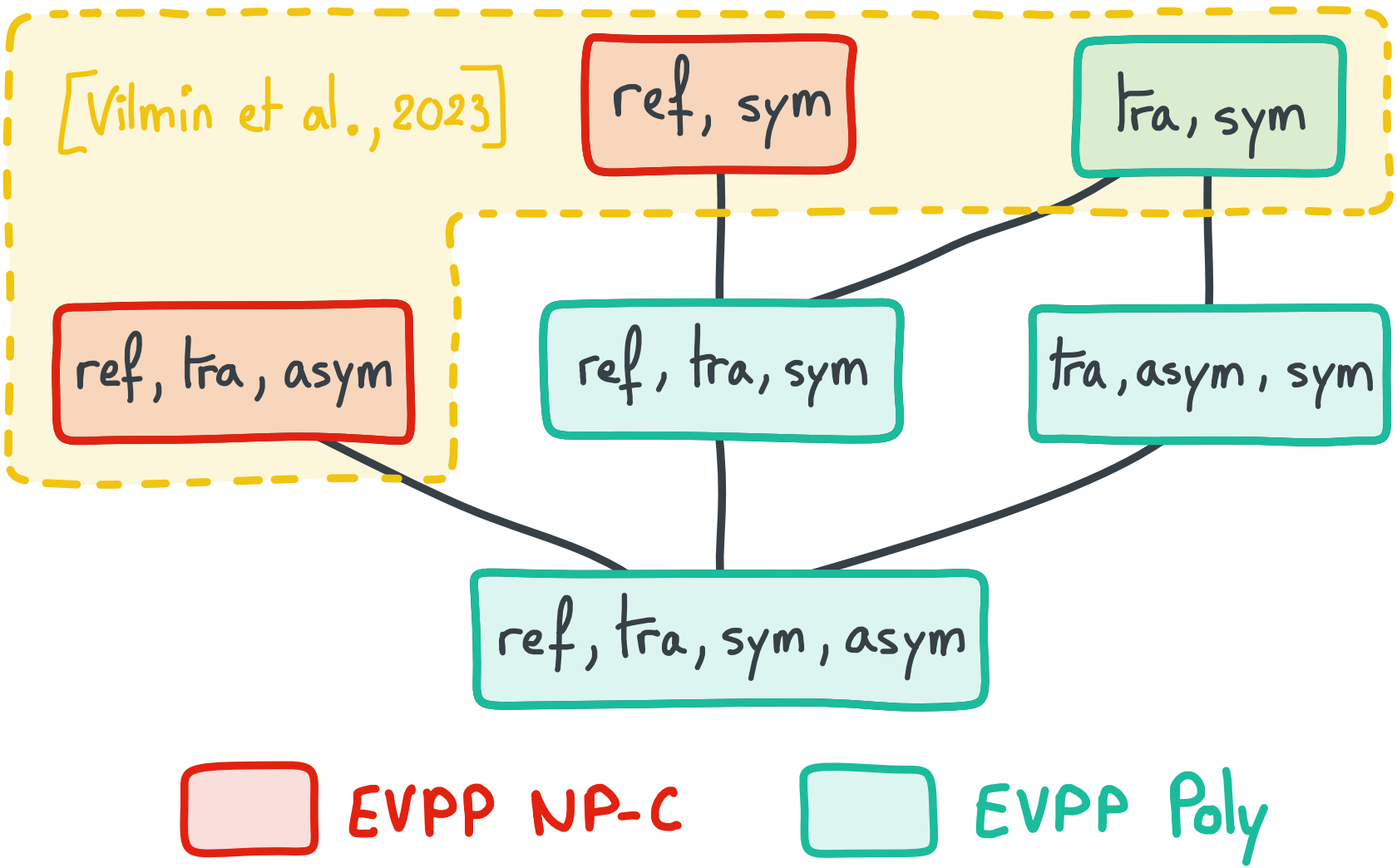
In: a graph $G=(V,E)$, $k \in \mathbb{N}$

Out: YES if there exists an ind. set $I \subseteq V$ of G s.t. $|I| \geq k$, NO otherwise

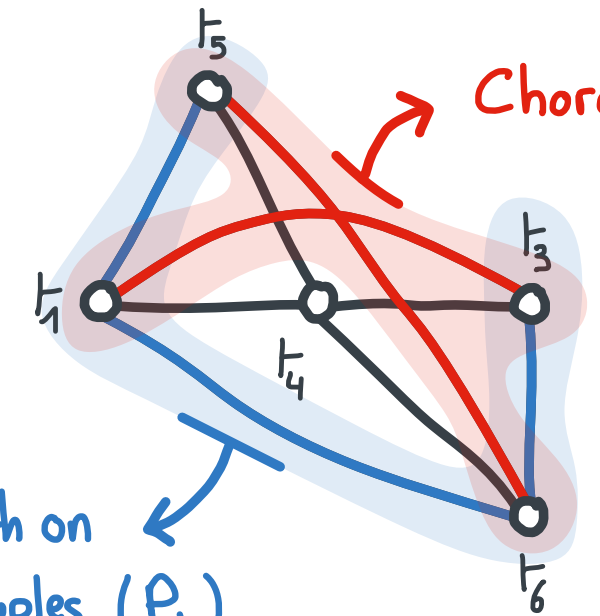
↓

Answer: The structure of CG_{Φ} imposed by Φ

Overview of our Results



r	F	E	P
t_1	2.5	10.1	22.9
t_2	2.7	10.4	23.2
t_3	2.6	10.3	23.0
t_4	2.5	10.2	23.3
t_5	2.6	10.1	23.1
t_6	2.6	10.3	22.9



Chords → P_4 not induced

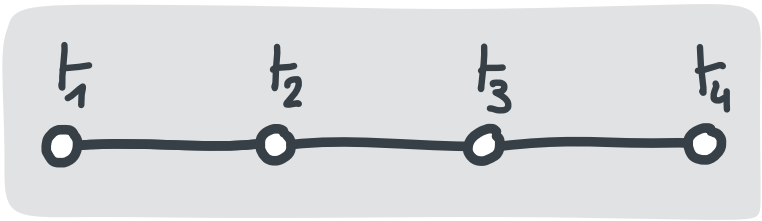
t_2

Path on 4 tuples (P_4)

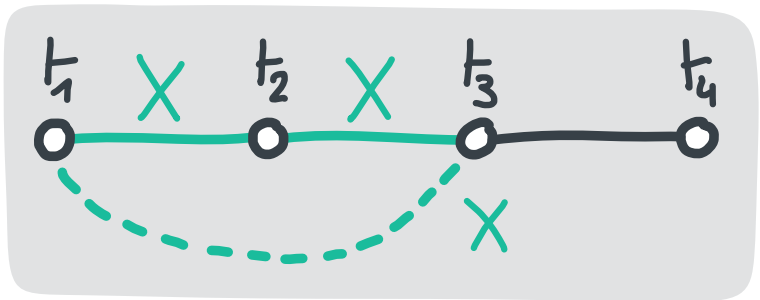
$$\phi_F(x,y) = 1 \iff 2.5 \leqslant x, y \leqslant 2.6$$
$$\phi_E(x,y) = 1 \iff 10.1 \leqslant x, y \leqslant 10.3$$
$$\phi_P(x,y) = 1 \iff 23.0 \leqslant x, y \leqslant 23.2$$

tra and sym, Ideas

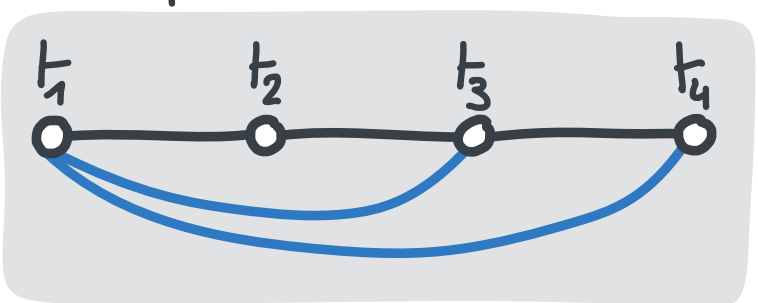
P_4 in $CG_{\mathcal{F}}(r, X \to A)$



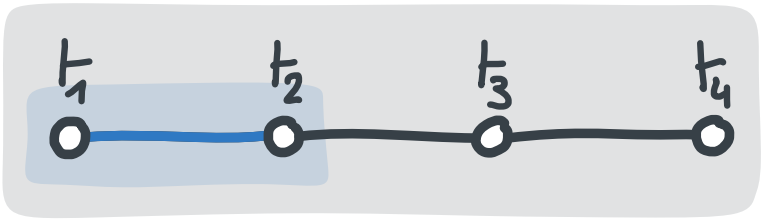
$t_1, \dots, t_4$ agree on X



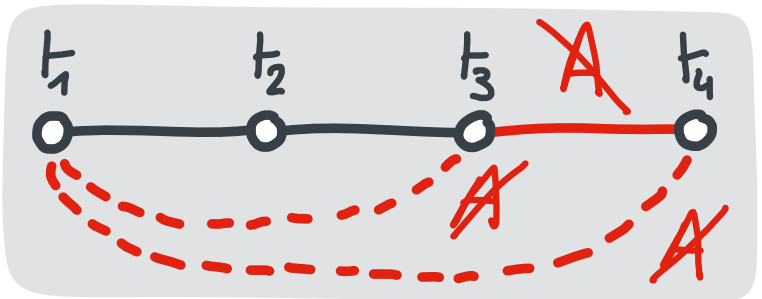
P_4 is not induced



t_1, t_2 agree on X but not on A



t_1, t_3 OR t_1, t_4 disagree on A



Co-graph, MIS poly

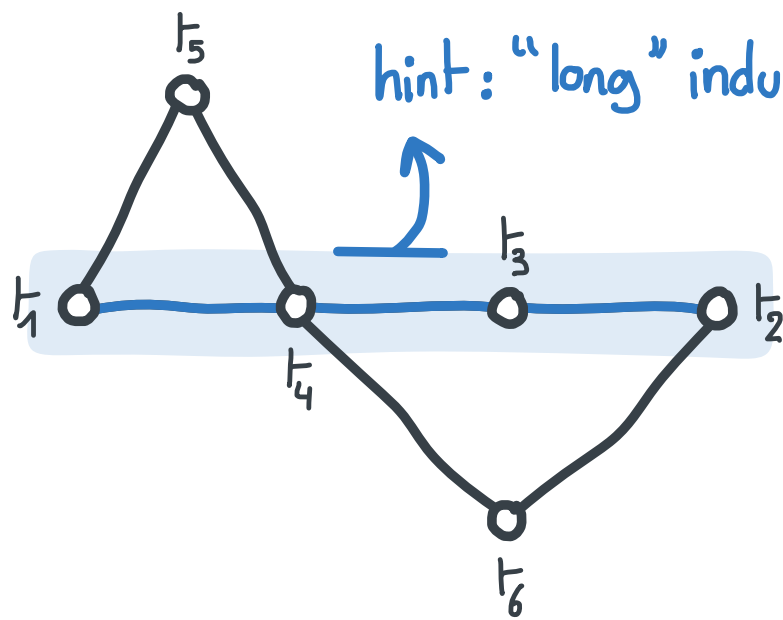
tra, sym: $CG_{\mathcal{F}}$ is P_4-free

→ EVPP poly

ref, sym: CG_Φ can be any graph → EVPP hard

↳ MIS hard

r	F	E	P
r_1	2.5	10.1	22.9
r_2	2.7	10.4	23.2
r_3	2.6	10.3	23.0
r_4	2.5	10.2	23.3
r_5	2.6	10.1	23.1
r_6	2.6	10.3	22.9

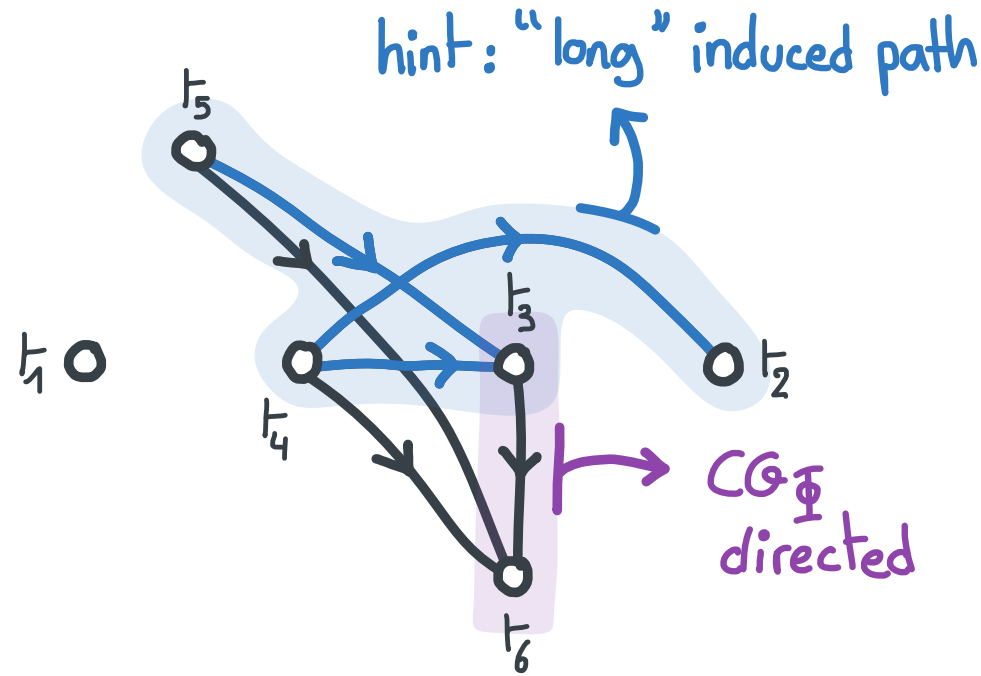


$\phi_P = \phi_E = \phi_F$ $\quad \phi_P(x,y) = 1 \Leftrightarrow |x-y| \leq 0.1$

tra, ref, asym: CG_Φ can be any 2-subdivision graph → EVPP hard

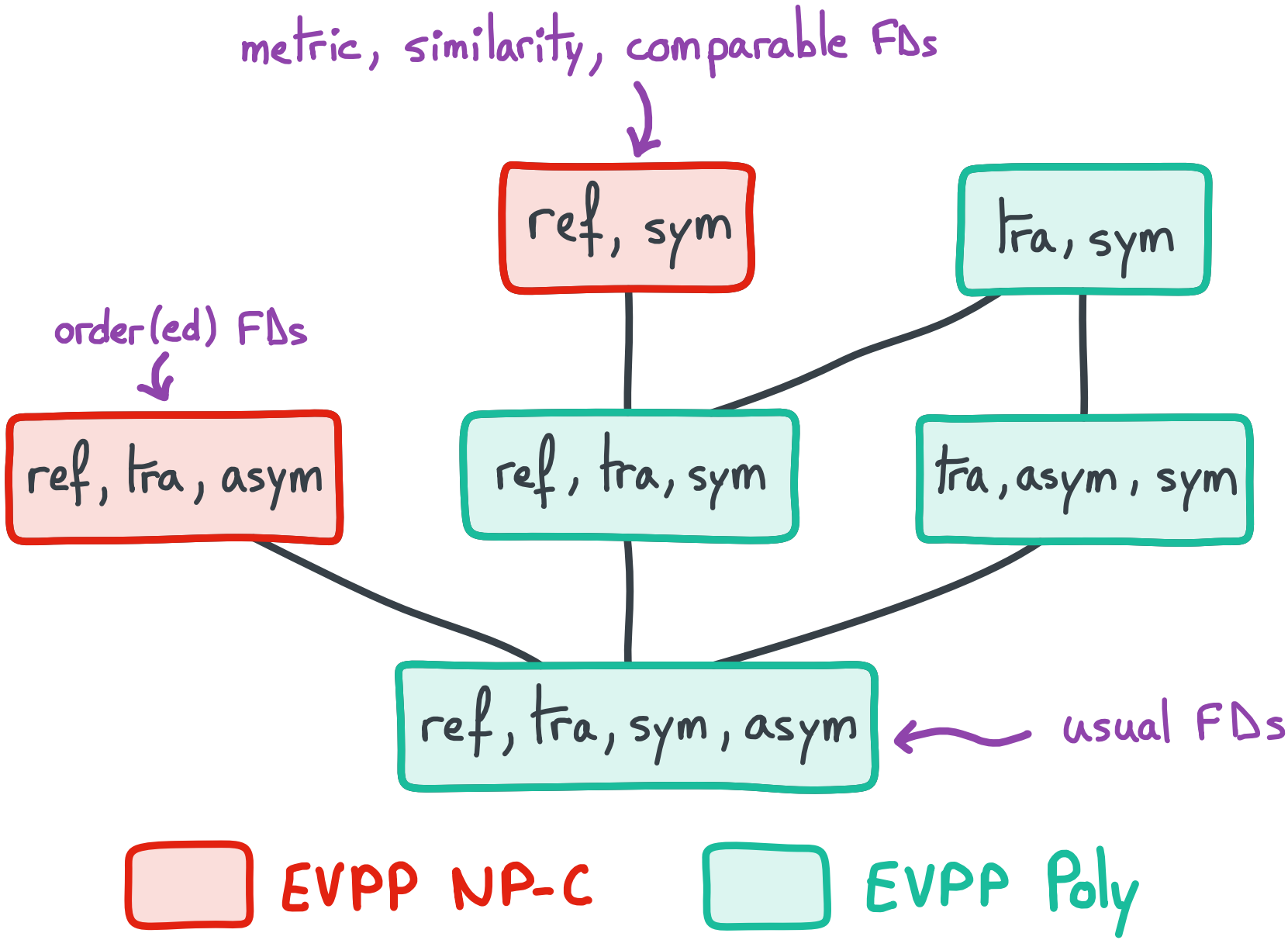
MIS hard

r	F	E	P
t_1	2.5	10.1	22.9
t_2	2.7	10.4	23.2
t_3	2.6	10.3	23.0
t_4	2.5	10.2	23.3
t_5	2.6	10.1	23.1
t_6	2.6	10.3	22.9



$\phi_P = \phi_E = \phi_F$ $\quad \phi_P(x,y) = 1 \Leftrightarrow x \leqslant y$

Hierarchy


metric, similarity, comparable FDs
ref, sym
tra, sym
order(ed) FDs
ref, tra, asym
ref, tra, sym
tra, asym, sym
ref, tra, sym, asym
usual FDs
EVPP NP-C
EVPP Poly

Conclusion

EVPP: estimate the g_3-error of a functional dependency with predicates

- can be used to confront experts knowledge against data [Faure -- Giovagnoli, 2022]
- complexity depends on the properties of predicates and the underlying conflict-graph [Bertossi, 2011]

Main results [Vilmin et al., 2023]

- having sym and tra ⇒ EVPP poly
- dropping sym or tra ⇒ EVPP NP-complete

Further research:

- Practical algorithms for special cases?
- Connections with repairs of sets of FDs? [Livshits et al., 2020]

References

L. Bertossi [Bertossi, 2011]
Database repairing and consistent query answering
Synthesis Lectures on Data Management, 2011

L. Caruccio, V. Deufemia, G. Giuseppe [Caruccio et al., 2015]
Relaxed Functional Dependencies—a survey of approaches
IEEE Transactions on Knowledge and Data Engineering, 2015

P. Faure-- Giovagnoli, J.-M. Petit, V.-M. Scuturici [Faure-- Giovagnoli et al., 2022]
Assessing the Existence of a Function in a Dataset with the g_3 indicator
IEEE International Conference on Data Engineering, 2022

Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen [Huhtala et al., 1999]
TANE: An efficient algorithm for discovering functional and approximate dependencies
The Computer Journal, 1999

J. Kivinen, H. Mannila [Kivinen, Mannila, 1995]
Approximate inference of functional dependencies from relations
Theoretical Computer Science, 1995

References

E. Livshits, B. Kimelfeld, R. Sudeepa [Livshits et al., 2020]
Computing optimal repairs for functional dependencies
ACM Transactions on Database Systems, 2020

S. Song, L. Chen, P. Yu [Song et al., 2013]
Comparable dependencies over heterogeneous data
The VLDB journal, 2013

S. Song, F. Gao, R. Huang, C. Wang [Song et al., 2020]
Data Dependencies over Big Data: A Family Tree
IEEE Transactions on Knowledge and Data Engineering, 2020

S. Vilmin, P. Faure--Giovagnoli, J.-M. Petit, V.-M. Scuturici [Vilmin et al., 2023]
Functional Dependencies with Predicates: What Makes the g_3-error easy to compute?
LNCS, ICCS, 2023