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

BDA 2022

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

[1]LIRIS, Univ Lyon, INSA Lyon, UCBL, CNRS, Villeurbanne
[2]Companie Nationale du Rhône

October 2022

Functional dependencies and domain knowledge

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

- Data from a hydropower turbine:
  - incoming flow F $(\mathrm{m}^3 \cdot \mathrm{s}^{-1})$
  - elevation E of the waterfall (m)
  - power P produced (MW)
- Domain knowledge:
  - P is *determined* by E and F, i.e. $P = f(E, F)$

Question. Is domain knowledge supported by data?

- *function* $P = f(E, F) \Leftrightarrow$ *functional dependency* $EF \rightarrow P$ *holds*
- $EF \rightarrow P$ does not hold: (t_3, t_6) is a *(unique) counterexample*

Drawbacks of functional dependencies

- A functional dependency (FD) $X \rightarrow A$ holds in a relation r, written $r \models X \rightarrow A$, if

$$\forall t_1, t_2 \in r, t_1[X] = t_2[X] \implies t_1[A] = t_2[A]$$

- Real-life problems:
  - ✗ mathematical equality is *too restrictive*
  - ✗ may not hold on the *whole dataset*
- Theoretical solutions:
  - ✓ use *predicates* instead of equality
  - ✓ use a *coverage measure* to estimate the partial validity of $X \rightarrow A$

Predicates to relax equality

- Each attribute A is equipped with a *binary predicate* comparing every two values in the *domain* ($\texttt{dom}$) of A:

$$\phi_A : \texttt{dom}(A) \times \texttt{dom}(A) \rightarrow \{\texttt{true}, \texttt{false}\}$$

- e.g.: distance, similarity, order, ... [Caruccio et al., 2021, Song et al., 2020]

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

- A FD $X \rightarrow A$ holds in a relation r w.r.t. (R, Φ), written $r \models_\Phi X \rightarrow A$, if

$$\forall t_1, t_2 \in r, \bigwedge_{B \in X} \phi_B(t_1[B], t_2[B]) \implies \phi_A(t_1[A], t_2[A])$$

The g_3-error and the error validation problem

- *g_3-error* coverage measure introduced in [Kivinen, Mannila, 1995]:
  - for classical FDs and equality,
  - *minimal proportion* of tuples to remove from r to satisfy $X \rightarrow A$
- *adapted* to predicates [Faure--Giovagnoli et al., 2022]:

$$g_3^{\Phi}(r, X \rightarrow A) = 1 - \frac{\max(\{|s| \mid s \subseteq r, s \models_{\Phi} X \rightarrow A\})}{|r|}$$

- thus, assessing domain knowledge is solving:

Problem. Error Validation Problem with Predicates (EVPP)

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

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

Back to the example

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

$$\phi_P(x,y) = \begin{cases} \text{true} & \text{if } |x-y| \leqslant 0.1 \\ \text{false} & \text{otherwise} \end{cases}$$

$$\phi_P = \phi_E = \phi_F$$

- (t_3, t_6) no longer an *"erroneous"* *counterexample*
- (t_4, t_6) *"real"* *counterexample*, so $r \not\models_\Phi EF \to P$
- $g_3^\Phi(r, EF \to P) = 0.5$

Situation

- about the complexity of EVPP:
  - *polynomial* for usual FDs with equality [Huhtala et al., 1999],
  - *NP-complete* for specific relaxed FDs (e.g. differential, matching, comparable) [Song et al., 2013, Caruccio et al., 2021]
- *what makes the problem tractable (or not)?*
  - *idea:* study the impact of (common) *predicates properties* on EVPP:

    (ref): $\phi_A(x, x) = \texttt{true}$

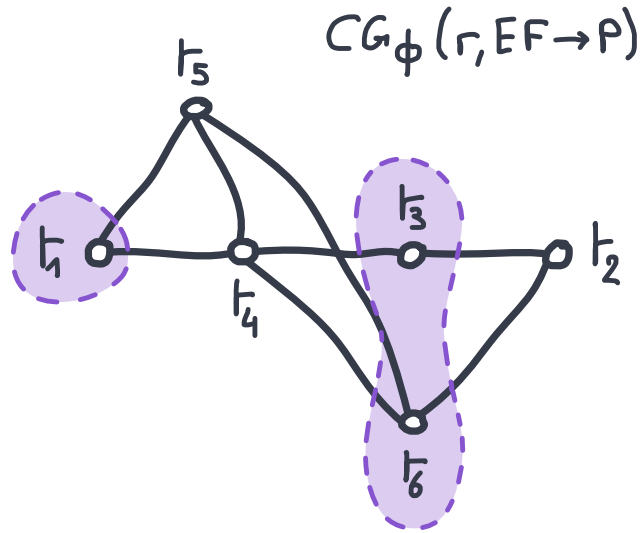
    (sym): $\phi_A(x, y) = \texttt{true}$ implies $\phi_A(y, x) = \texttt{true}$

    (tra): $\phi_A(x, y) = \phi_A(y, z) = \texttt{true}$ implies $\phi_A(x, z) = \texttt{true}$

    (asym): $\phi_A(x, y) = \phi_A(y, x) = \texttt{true}$ implies $x = y$
  - *goal:* a quick-reference map of EVPP complexity

Conflict-graph

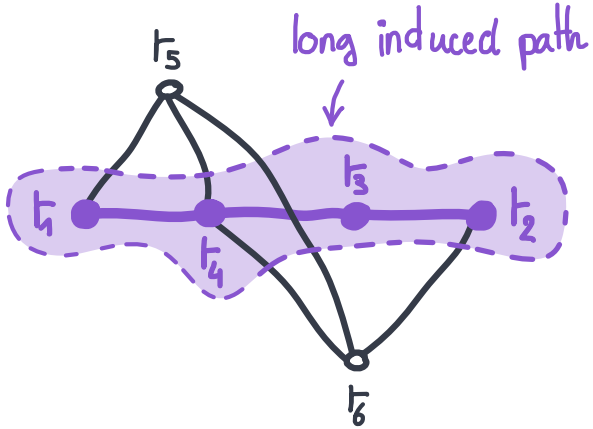
r	F	E	P
r_1	2.5	10.1	22.9
r_2	2.7	10.4	23.2
r_3	2.6	10.3	23.0
r_4	2.5	10.2	23.3
r_5	2.6	10.1	23.1
r_6	2.6	10.3	22.9

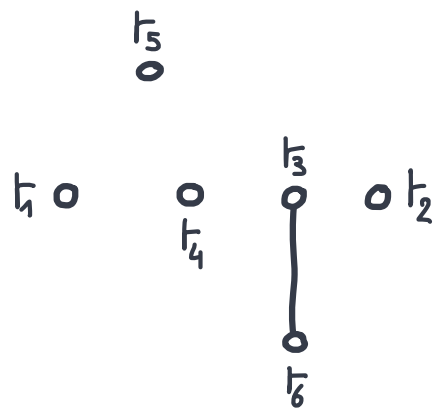


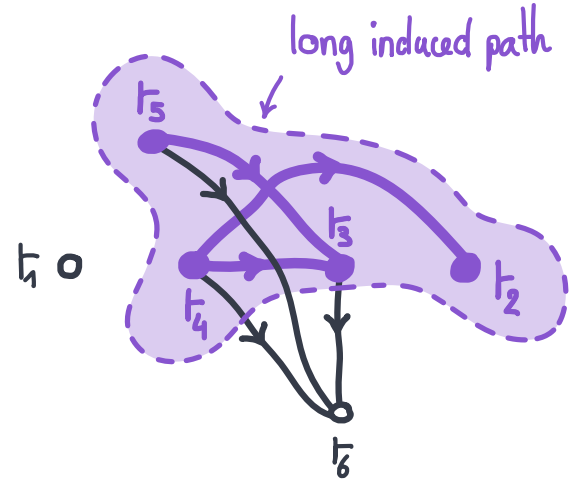
- $CG_\Phi(r, EF \to P)$ *conflict-graph* of $EF \to P$ in r (see [Bertossi, 2011])
- for $s \subseteq r$, $s \models_\Phi EF \to P \Leftrightarrow s$ is an *independent set* of $CG_\Phi(r, EF \to P)$
- solving $\mathrm{EVPP} \Leftrightarrow$ finding the *maximal size of an independent set* in CG_Φ

The structure of conflict-graph

- Finding the maximal size of an independent set is NP-*complete*
- The *properties* of the predicates bound the *structure* of the conflict-graph!



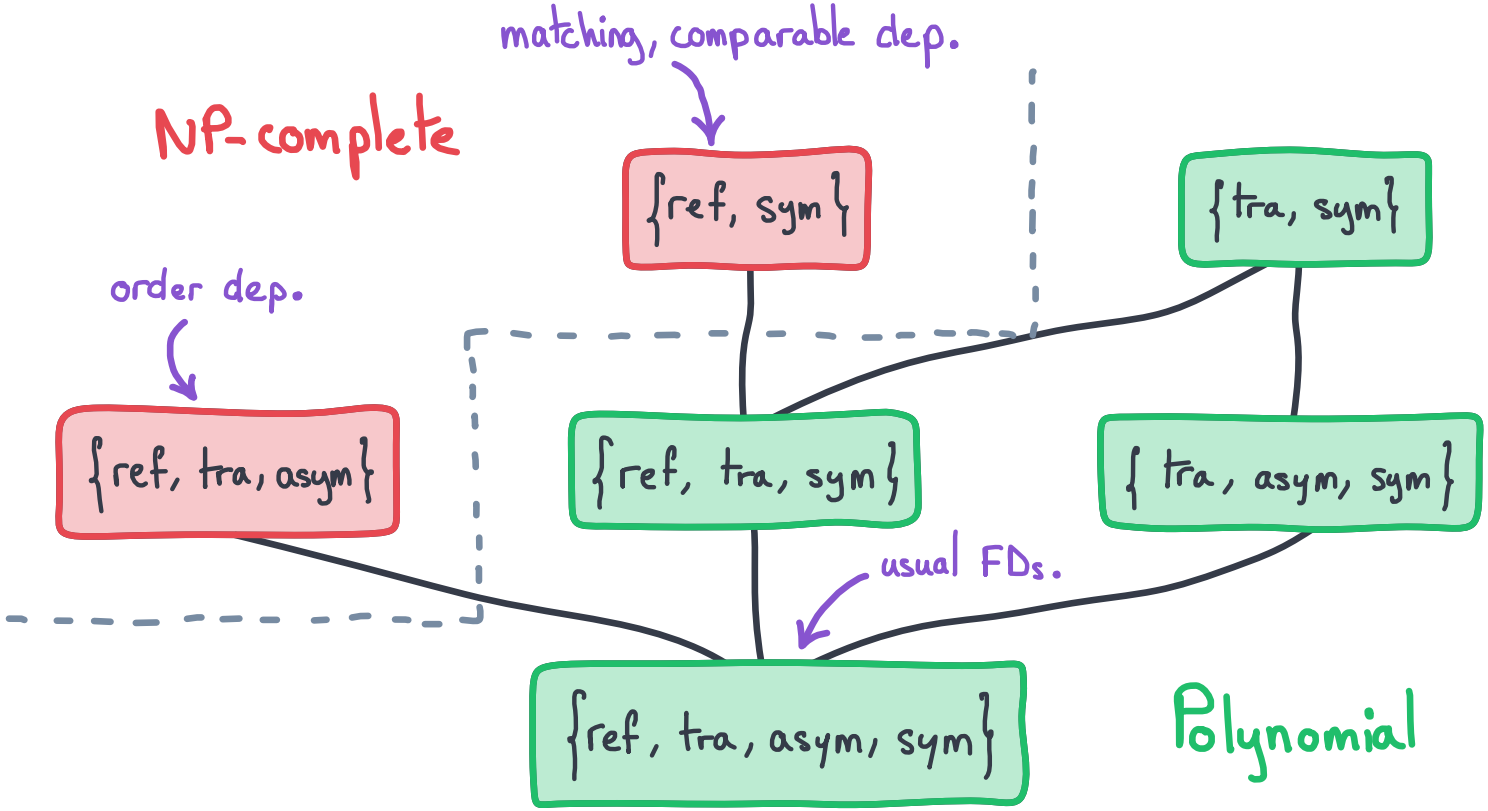




$\phi_p(x,y) = \text{true} \Leftrightarrow |x-y| \leq 0.1$ $\phi_p(x,y) = \text{true} \Leftrightarrow x = y$ $\phi_p(x,y) = \text{true} \Leftrightarrow x \leq y$

$CG_\phi(r, EF \to P)$ with $\phi_p = \phi_E = \phi_F$

The complexity of EVPP



Theorem. [Vilmin et al., 2022] The problem EVPP is :

- NP-*complete* when predicates enjoy `ref` and `sym`
- NP-*complete* when predicates enjoy `ref`, `tra` and `asym`
- *polynomial* when predicates enjoy `tra` and `sym`

Conclusion

- EVPP : estimate the g_3-error of a functional dependency with predicates
  - can be used to confront experts knowledge against data [Faure--Giovagnoli et al., 2022]
  - complexity depends on the *properties* of predicates and the underlying *conflict-graph* [Bertossi, 2011]
- Main results:
  - *having* `sym` and `tra` $\implies$ EVPP *polynomial*
  - *dropping* `sym` or `tra` $\implies$ EVPP NP-*complete*
- Further research:
  - practical algorithms for special cases?
  - connection with repairs for sets of FDs? [Livshits et al., 2017]

References

- L. Bertossi
  Database repairing and consistent query answering.
  Synthesis Lectures on Data Management, vol. 3, p. 1–121, 2011.
- L. Caruccio, V. Deufemia, and G. Giuseppe
  Relaxed functional dependencies–a survey of approaches.
  IEEE Transactions on Knowledge and Data Engineering, vol. 28, p. 147–165, 2015.
- P. Faure--Giovagnoli, J.-M Petit, V.-M Scuturici
  Assessing the Existence of a Function in a Dataset with the g3 Indicator.
  IEEE International Conference on Data Engineering, 2022.
- Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen
  TANE: An efficient algorithm for discovering functional and approximate dependencies.
  The computer journal, vol. 42, p. 100–111, 1999.
- J. Kivinen, H. Mannila
  Approximate inference of functional dependencies from relations.
  Theoretical Computer Science, vol. 149, p. 129–149, 1995.

References

- E. Livshits, B. Kimelfeld, R. Sudeepa
  Computing optimal repairs for functional dependencies
  ACM Transactions on Database Systems, vol. 45, p. 1–46, 2020.
- S. Song, L. Chen, and P. Yu
  Comparable dependencies over heterogeneous data.
  The VLDB journal, vol. 22, p. 253–274, 2013.
- S. Song, F. Gao, R. Huang, and C. Wang
  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
  Relaxed functional dependencies–a survey of approaches.
  IEEE Transactions on Knowledge and Data Engineering, vol. 28, p. 147–165, 2015.