The Parameterized Complexity of Dependency Detection in Relational Databases [BFS17]

Filip Rydzi

Fixed-Parameter Algorithms and Complexity

February 11, 2018

Motivation

- Let's assume we have a database and we want to optimize the queries, normalize our data, or do some data cleaning...
- This includes finding **unique column combination**, **functional dependencies** or **inclusion dependencies** in our relations (all of them are NP-complete)
- Algorithms for these problems running well in practice, but guarantee no theoretical performance
- **Goal**: Exploit some properties and find better algorithms

Notations

Notations

R, S ... relational schemata
X, Y ... set of columns
A, B ... a single column
r, s ... instances of R, S
r_i, r_j ... tuples/rows
$r_i[X]$ ... tuple containing only columns in X

Definitions - UNIQUE

Unique column combination

Input: instance r of R, k.
Problem: Does there exits a subset $X \subseteq R$ of size at most k, s.t. for any two distinct tuples r_i and r_j from r holds: $r_i[X] \neq r_j[X]$. The size of Unique column combination equals $|X|$.

Definitions - FD

Functional dependency (FD)

Input: instance r of R, k.
Problem: Does there exist a subset $X \subseteq R$ of size at most k and attribute $A \in R$, s. t. for any pair of tuples from schema R, which agree on X, also agree on A. The expression $X \rightarrow A$ is called functional dependency. FD is non-trivial if $A \notin X$. The size of FD equals $|X|$. X is called *LHS* and A is called *RHS* of FD.

FD_{fixed}

To decide for a given attribute A, whether there exists an FD X .

Recap HITTING SET

HITTING SET

Input: ground set U, $\zeta \subseteq P(U)$, k
Questions: Does there exist a Hitting set H, s.t. $H \subseteq U$ and for all $Z \in \zeta, H \cap Z \neq \emptyset$ and $|H| \leq k$.
NP-complete and W[2]-complete

Reductions from HITTING SET

Lemma 1

HITTING SET $\leq_{FPT}$ UNIQUE $\leq_{FPT}$ FD_{fixed} $\leq_{FPT}$ FD

Figure: Proof (sketch)

	A	B	C	D	E
$U = \{a, b, c, d, e\}$	0	0	0	0	0
$Z_1 = \{a, b, c\}$	1	1	1	0	0
$Z_2 = \{a, d, e\}$	2	0	0	2	2
$Z_3 = \{b, d, e\}$	0	3	0	3	3
$Z_4 = \{b, c\}$	0	4	4	0	0

(a)

	A	B	C	D	
r_0	0	2	1	1	r
	1	1	1	1	r
	2	0	0	1	r
	1	2	2	0	r
	0	1	2	0	r

	A	B	C	D	
r_0	0	2	1	1	r
	1	1	1	1	r
	2	0	0	1	r
	1	2	2	0	r
	0	1	2	0	r
r_B	0	–	1	1	$r' \setminus r$
r_C	0	2	–	1	$r' \setminus r$
r_D	0	2	1	–	$r' \setminus r$

(b)

Reduction from FD to CNF-formula

Lemma 2

FD $\leq_{FPT}$ CNF

Towards the reduction

Given a relation r, we derive a propositional formula that has a satisfying truth assignment of weight k+1 iff there is a non-trivial FD of size k that holds in r.

Construction

- $Var_R = \{x_A | A \in R\}$ - if $A \in LHS$ of FD, set $x_A = TRUE$, otherwise $x_A = FALSE$
- $Var'_R = \{x'_A | A \in R\}$ - if $A \in RHS$ of FD, set $x'_A = TRUE$, otherwise $x'_A = FALSE$

Reduction from FD to CNF-formula

Construction RHS

- $c_R = \bigvee_{x'_A \in Var'_R} x'_A$
- $c_{A,B} = \neg x'_A \vee \neg x'_B \ (A \neq B)$
- $c_A = \neg x'_A \vee \neg x_A$, for every $A \in R$
- $\Phi_{RHS} = c_R \bigwedge_{A,B \in R \wedge A \neq B} c_{A,B} \bigwedge_{A \in R} c_A$
- *Any satisfying assignment chooses exactly one variable from Var'_R, while the corresponding variable in Var_R is not chosen.*

Reduction from FD to CNF-formula

Construction LHS

- $c_{A,r_i,r_j} = \neg x'_A \vee \bigvee_{B \in R \setminus A \wedge r_i[B] \neq r_j[B]} x_B$
- $\Phi_A = \bigwedge_{r_i, r_j \in r \wedge r_i[A] \neq r_j[A]} c_{A,r_i,r_j}$
- $\Phi_{LHS} = \bigwedge_{A \in R} \Phi_A$
- *If A is the RHS of non-trivial FD, then LHS has to contain at least one of the attributes $B \neq A$, s.t. $r_i[B] \neq r_j[B]$ and this has to hold for each attribute A and each pair of tuples r_i, r_j.*

Reduction from FD to CNF-formula

Result of the reduction:

$\Phi_{FD} = \Phi_{LHS} \wedge \Phi_{RHS}$

Recap - Towards the reduction

Given a relation r, we derive a propositional formula that has a satisfying truth assignment of weight k+1 iff there is a non-trivial FD of size k that holds in r.

Reduction from FD to CNF-formula

Proving the correctness of the reduction

We are given a satisfying assignment Φ_{FD} of weight k+1 and we want to derive a non-trivial FD of size k.

- We have a satisfying assignment for Φ_{RHS}
- Exactly one variable $x'_A \in Var'_R$ is set to TRUE, which determines the attribute on RHS
- $B \in X$ iff x_B is TRUE in Var_R, $A \notin X$ because c_A

Reduction from FD to CNF-formula

Proving the correctness of the reduction

Assume $X \rightarrow A$ holds in r then we find a satisfying assignment for Φ_{LHS}.

- x'_A is set to TRUE in Var'_R, others are set to FALSE. This implies that all clauses c_{B,r_i,r_j} with $B \neq A$ are satisfied.
- Since $X \rightarrow A$ holds, X includes, for every pair of tuples r_i, r_j an attribute B, s.t. $r_i[B] \neq r_j[B]$, which satisfies the clause c_{A,r_i,r_j}

Assume $X \rightarrow A$ fails in r.

- Then there is a pair of tuples $r_i, r_j \in r$, s.t. $r_i[A] \neq r_j[A]$, but $r_i[X] = r_j[X]$, consequently c_{A,r_i,r_j} doesn't contain any variables x_B s.t. $B \in X$
- Thus all literals in c_{A,r_i,r_j} from Var_R evaluate to FALSE and $\neg x'_A$ is also FALSE, because A is in RHS.

Reduction from FD to CNF-formula

Lemma 1

HITTING SET $\leq_{FPT}$ UNIQUE $\leq_{FPT}$ FD_{fixed} $\leq_{FPT}$ FD

Lemma 2

FD $\leq_{FPT}$ CNF

Theorem (Theorem 3)

Since HITTING SET and CNF are both W[2]-complete and the reductions in Lemma 1 and Lemma 2 are correct. The problems UNIQUE, FD_{fixed} and FD are **W[2]-complete**.

Definitions - IND

Inclusion dependency (IND)

Input: r,s instances of R, S; k
Problem: Decide if there is IND (X, σ) of size at least k, s.t. $X \subseteq R$ and $\sigma : X \to S$. Where for each $r_i \in X$ there exist a $s_j \in S$, s.t. $r_i[A] = s_j[\sigma(A)]$ for every $A \in X$. The size of IND is $|X|$.

Explanation - Weighted Antimonotone 3-normalized Satisfiability (WA3NS)

Example of WA3NS

$((\neg a \wedge \neg b) \vee (\neg c \wedge \neg d)) \wedge ((\neg a \wedge \neg c) \vee (\neg b \wedge \neg d))$

IND is in W[3]

Assumption

IND is in class W[3]. We will show this by providing the following reduction: $IND \leq_{FPT} WA3NS$.

IND is in W[3]

Theorem 5

$IND \leq_{FPT} WA3NS$.

We construct from two relations R, S an antimonotone formula which has a weight k satisfying assignment iff the relations have an inclusion dependency of size k.

IND is in W[3]

Towards the reduction

- One $A \in X$ cannot be mapped to multiple $B \in S$ by σ and one B cannot be an output of multiple different $\sigma(A)$ (Boolean formula Φ_{map})
- Assume relations r, s contains only a single tuple r_i, s_j each, then a pair (A, B) is forbidden for r_i, s_j if $r_i[A] \neq s_j[B](B = \sigma(A))$
- For each tuple r_i in r there is a tuple s_j in s, s.t. r_i, s_j is not forbidden, i.e. is an IND (Boolean formula Φ)
- $\Phi \wedge \Phi_{map}$ is an instance of WA3NS computed in polynomial time

Corollary 6

IND and IND_{fixed} is in class W[3]

IND is in W[3]-hard

Theorem 7

It can be shown that IND is W[3]-hard and since it's in the W[3] class, it's W[3]-complete.

Conclusion

- Unique
  - NP-complete problem
  - we can easily construct an algorithm running in $2^{|R|}$
  - parametrized version is W[2]-complete
- Functional dependency(FD)
  - restricted variant FD_{fixed} is a NP-complete problem
  - parametrized version is W[2]-complete
- Inclusion dependency
  - a NP-complete problem
  - parametrized version is W[3]-complete

References I

This presentation has been done based on work by Blsius Thomas, Friedrich Tobias and Schirneck Martin as a part of the lecture Fixed-Parameter Algorithms and Complexity at TU Vienna.

Thomas Bläsius, Tobias Friedrich, and Martin Schirneck, *The Parameterized Complexity of Dependency Detection in Relational Databases*, 11th International Symposium on Parameterized and Exact Computation (IPEC 2016) (Dagstuhl, Germany) (Jiong Guo and Danny Hermelin, eds.), Leibniz International Proceedings in Informatics (LIPIcs), vol. 63, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 6:1–6:13.