

Query-Oriented Data Cleaning with Oracles

Moria Bergman

Tova Milo

Slava Novgorodov

Wang-Chiew Tan*

Motivation

- Key decisions are made based on information in databases
- Ideal: complete and correct databases
- Goal: Find **wrong** and **missing** information and correct it
- In practice: Impossible to manually examine each piece of data
- Existing data cleaning tools:
 - provide best effort
 - do not usually address data completeness
- Our solution: Query-directed data cleaning with the help of a crowd of domain experts (oracles)**

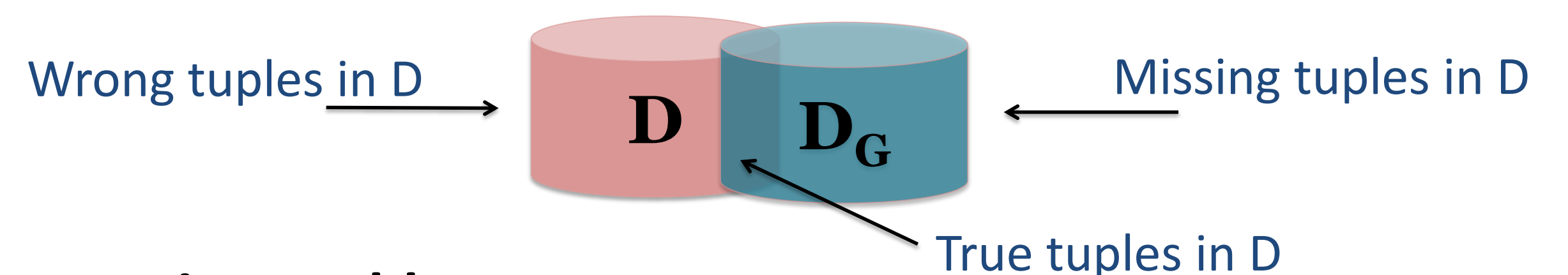
Model & problem definition

Database D:

Truly open-world assumption.

- A fact not in D can be true or false.
- A fact in D can be true or false.

Truth is determined w.r.t. the **conceptual ground truth database D_G** (known to oracles) which contains true tuples and only them.



Edit Generation Problem

Given D, D_G , and Q , interact with the crowd at most k times to derive a sequence e_1, \dots, e_k of edits such that

$$Q(D \oplus e_1 \oplus e_2 \oplus \dots \oplus e_k) = Q(D_G)$$

Removing a wrong answer

Theorem (informally): The edit generation problem is NP-hard even for only removing one wrong query answer

Proof: Reducing from the Hitting-Set Problem

Solution outline: A greedy approach that asks oracles Boolean questions about tuples that appear in many "witness sets".

* Performs well in practice.

Example

Query: Find European teams who won the World-Cup at least twice?

Result(x) : – $Games(d_1, x, y, "Final", u_1)$,
 $Games(d_2, x, z, "Final", u_2)$,
 $Teams(x, "EU"), d_1 \neq d_2$

Result: [Germany], [Spain]

Missing: [Italy]

Database: Games

Year	Winner	Runner-up	Result
2014	GER	ARG	1:0
2010	ESP	NED	1:0
2006	ITA	FRA	5:3
2002	BRA	GER	2:0
1998	ESP	NED	4:2
1994	ESP	NED	3:1
1990	GER	ARG	1:0
1982	ITA	GER	4:1

Teams

Country	Cont
GER	EU
ESP	EU
BRA	EU
ITA	EU

* Wrong
* Missing

Example for the greedy approach

Result(x) : – $Games(d_1, x, y, "Final", u_1)$,
 $Games(d_2, x, z, "Final", u_2)$,
 $Teams(x, "EU"), d_1 \neq d_2$

Result
GER
ESP

Wrong answer

Algorithm:

- ESP has six witnesses
- No singleton witness
- t_3 occurs most frequently
- Ask: "Is t_3 true?" – YES
- Remove from consideration
- Each tuple occurs equally often
- Pick t_5 . (randomly), ask – NO
- Delete t_5 from D
- w_1, w_2, w_3 , are left
- Ask: "Is t_1 true?" – YES
- Remove from consideration
- Singleton sets: $\{t_2\}, \{t_4\}$
- No questions needed - remove t_2 and t_4 from D

	Tuples of the witness
w_1	$t_1 = Games(11.7.10, ESP, NED, Final, 1:0)$ $t_2 = Games(12.7.98, ESP, NED, Final, 4:2)$ $t_3 = Teams(ESP, EU)$
w_2	$t_2 = Games(12.7.98, ESP, NED, Final, 4:2)$ $t_4 = Games(11.7.94, ESP, NED, Final, 3:1)$ $t_3 = Teams(ESP, EU)$
w_3	$t_4 = Games(11.7.94, ESP, NED, Final, 3:1)$ $t_1 = Games(11.7.10, ESP, NED, Final, 1:0)$ $t_3 = Teams(ESP, EU)$
w_4	$t_1 = Games(11.7.10, ESP, NED, Final, 1:0)$ $t_5 = Games(25.06.78, ESP, NED, Final, 1:0)$ $t_3 = Teams(ESP, EU)$
w_5	$t_2 = Games(12.7.98, ESP, NED, Final, 4:2)$ $t_5 = Games(25.06.78, ESP, NED, Final, 1:0)$ $t_3 = Teams(ESP, EU)$
w_6	$t_4 = Games(11.7.94, ESP, NED, Final, 3:1)$ $t_5 = Games(25.06.78, ESP, NED, Final, 1:0)$ $t_3 = Teams(ESP, EU)$

Adding a missing answer

Theorem (informally): The edit generation problem is NP-hard even for only adding one missing query answer

Proof: Reducing from ONE-3SAT Problem

Solution outline:

- Split the query into smaller subqueries
- Exploit to data in database – examine subqueries results
- Ask the crowd to complete subqueries results into a result of Q
- Continue recursively

The general algorithm

- Iteratively treats wrong and missing answers until convergence
- Employs multiple experts to prune errors
- Employs parallelism between components treating wrong and missing answers and whenever possible inside each component

