TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

DANCE: Data Cleaning with Constraints and Experts

Ahmad Assadi **Tova Milo** **Slava Novgorodov**

Motivation

- Data cleaning is a long standing problem that has attracted much research interest in DB community
- There are some automatic solutions based on "consistency rules" which are using (1) minimal repair or (2) preferences
- However, such repairs usually doesn't represent the **ground truth**
- **Our solution**: **Using both constraints and human experts**

Example of the data and constraints

Sample rules: Based on official UEFA regulations:

- $Games(x_1, x_2, x_3, x_4, x_5) \wedge x_5 = "GroupStage" \wedge Teams(x_1, y_1) \wedge Teams(x_2, y_2) \rightarrow y_1 \neq y_2$
- $Countries(x_1, x_2) \wedge x_2 > 0 \rightarrow Teams(y_1, x_1)$

Database:

GAMES

Team1	Team2	Team1_Goals	Team2_Goals	Stage
Celtic	ManCity	3	3	Group Stage
…	…	…	…	…

TEAMS

Name	Country
Celtic	UK
ManCity	UK
…	…

COUNTRIES

Name	Num_of_Teams
UK	5
…	…

✓ Correct tuple ✗ Wrong tuple

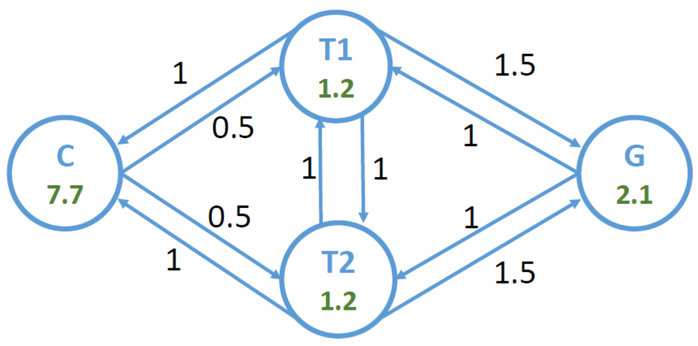
Violation:
{Games(Celtic, Manchester City, 3, 3, Group Stage),
Teams(Celtic, UK),
Teams(Manchester City, UK)}

Tuples Graph

Intuition: Build a weighted graph based on violations in order to find the tuples that have the maximal potential to fix the inconsistency

Relation Error Probability (β): Per each relation, there is a probability of a tuple from the relation being wrong. This mainly depend of the data source (e.g. data from official web-site can be more trusted then data aggregated from user generated content.

Graph Construction:
Graph Vertices: Suspicions tuples
Graph Edges: There are an edge from v to u if fixing/validating u may remove v from being suspicious
Edge weights: The weight of (u, v) is the potential of question about u to remove v from being suspicious multiplied by β
Vertex weights: Calculated using PageRank-style algorithm, based on edge weights



DANCE algorithm vs. Minimal repair

Minimal repair: (1) - Remove the game (Celtic, Manchester City).

GAMES

Team1	Team2	Team1_Goals	Team2_Goals	Stage
~~Celtic~~	~~ManCity~~	~~3~~	~~3~~	~~Group Stage~~
…	…	…	…	…

TEAMS

Name	Country
Celtic	UK
ManCity	UK
…	…

COUNTRIES

Name	Num_of_Teams
UK	5
…	…

DANCE Algorithm: (1) – Remove Countries (UK, 5)

GAMES

Team1	Team2	Team1_Goals	Team2_Goals	Stage
Celtic	ManCity	3	3	Group Stage
…	…	…	…	…

TEAMS

Name	Country
(2) Celtic	Scotland
(4) ManCity	England
…	…

COUNTRIES

Name	Num_of_Teams
(3) Scotland	1
(5) England	4
…	…