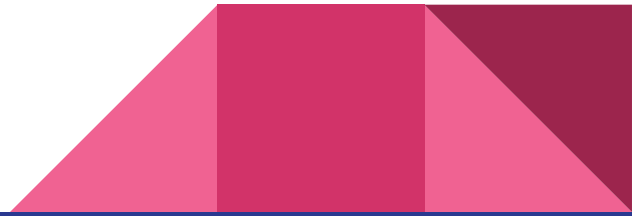
Properties of Inconsistency Measures for Databases

Hernan Razo
Samantha Berg
Christopher Anthony

Introduction

- The study focuses on the inconsistency in databases.
- The Measures used to quantify inconsistencies in databases.
- Properties of inconsistency measures.
- Proposes a new measure of inconsistency.
- Verified the new measure using Case study of HoloClean System.



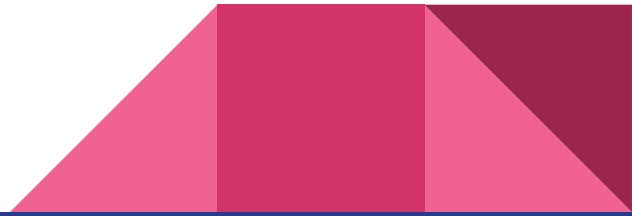
Integrity Constraints and Inconsistent Databases

Integrity Constraints:

- The columns in databases have certain set of rules that must be followed. Simple example: A column for contact number should not have characters.

Inconsistent Databases

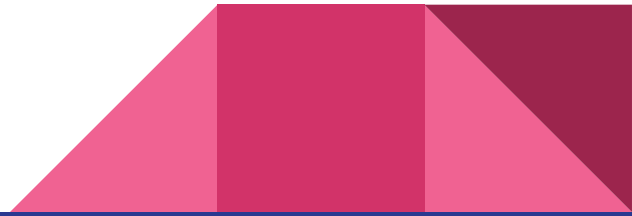
- A database that violates the integrity constraints is said to be inconsistent.
- Same data exists in different formats in different tables.



Why Inconsistency in Databases?

There can be various reasons for inconsistency in a database:

- Data Generation
- Data Integration
- Imprecise Data
- Difference in integrity constraints across databases.



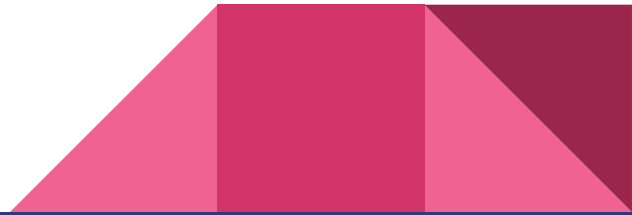
Measuring Inconsistency of a Database

The Inconsistency Measures are as below:

- I_d: Drastic Measure
- I_p: Problematic Tuples
- I_{MI}: Minimal Inconsistent Subsets
- I_{MC}: Maximum consistent subsets
- I_R: Minimal tuples to remove to satisfy constraints

Properties of Inconsistency Measures

- Positivity: Positive only if the database is inconsistent
- Monotonicity: Strict constraints cannot reduce inconsistency
- Continuity: Single operation has limited impact on inconsistency
- Progression: We can always find an operation that reduces inconsistency
- Computational Complexity of the Measure



Proposed Measure

- The new measure is a linear relaxation of the Measure I_R (the minimal number of tuples to be removed to satisfy the minimum constraints)
- The ILP of I_R is used with a change in the second equation.
- In the first equation, we ensure that for every Minimal inconsistent subset that violate the constraint we remove at least one tuple to resolve the violation.
- The result of the ILP is the value of the measure I_R.

$$\text{Minimize}: \sum_{i \in ids(D)} x_i \cdot \kappa(\langle -i \rangle(\cdot), D) \text{ subj. to:}$$

$$\forall E \in \text{MI}_\Sigma(D): \sum_{i \in ids(E)} x_i \geq 1 \quad (1)$$

$$\forall i \in ids(D): x_i \in \{0, 1\} \quad (2)$$

Proposed Measure

- In the second equation x_i is 0 or 1 based on which the tuple is removed or kept: $x_i = 1$: remove the tuple $x_i = 0$: keep the tuple.
- The second equation is changed to $0 <= x_i <= 1$
- In the New Measure the second condition is replaced with a new condition "$\forall i \in \text{ids}(D) : 0 \leq xi \leq 1$" it allows to use any value between 0 and 1 for the variable.

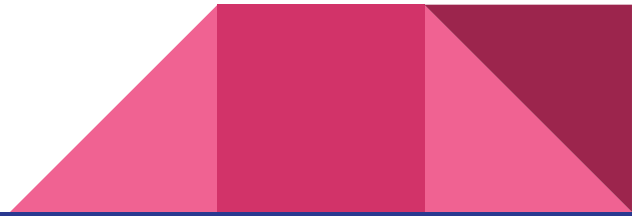
Experimental Study

8 Datasets are used for the experiment and random noise is added to it using one of the two algorithms:

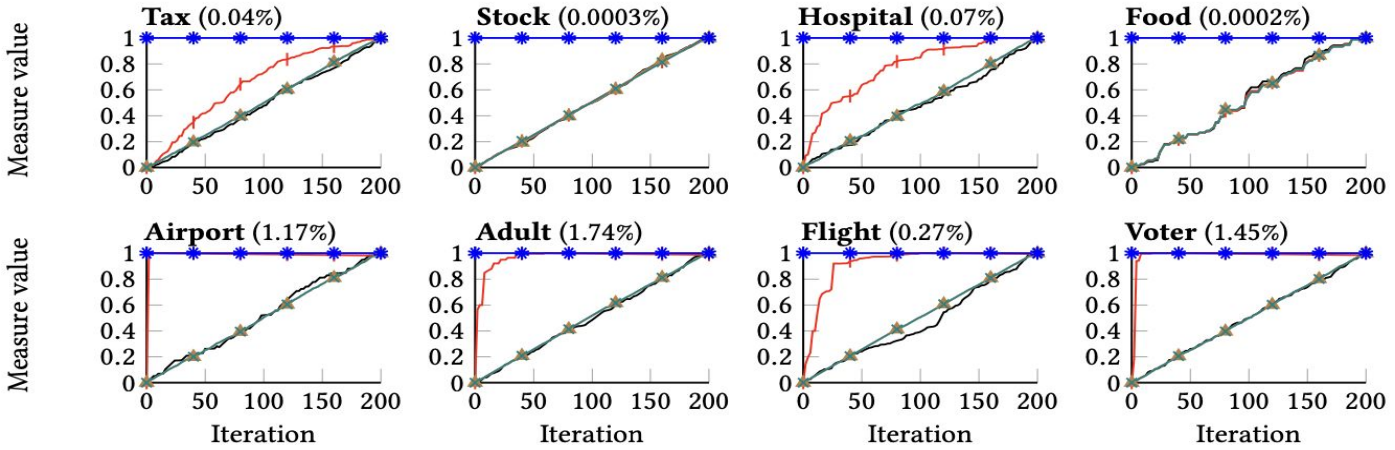
- Constraint Oriented Noise (CONoise)
  - Randomly Select a Constraint
  - Randomly Select 2 tuples from the Database
  - Modify value in one of the tuples to violate constraints
- Random Noise (RNoise)
  - Randomly select a database cell.
  - Change its value to another value from the active domain or to a typo.

Experimental Study

- In each iteration one tuple or cell is changed.
- After modifying the databases the Measure values are then computed for the databases.

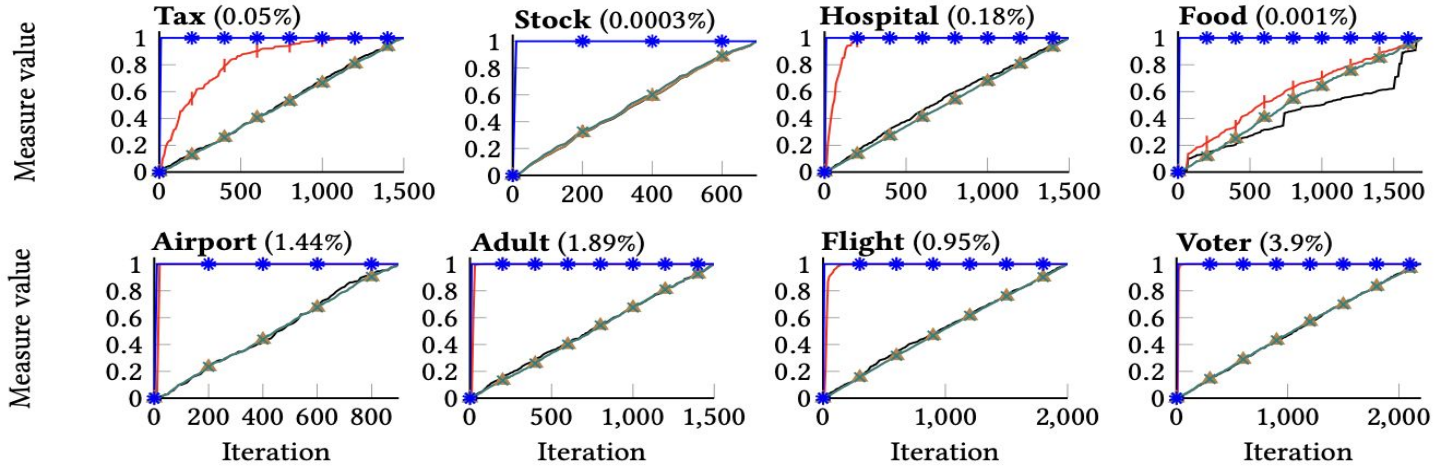


Comparing Measure Values for Noise added with CONoise



(a) **Noise added with** CONoise.

Comparing Measure Values for Noise added with RNoise



(b) Noise added with RNOise ($\alpha = 0.01$ **and** $\beta = 0$**).**

Case Study HoloClean

- The HoloClean system is a cleaning system that the authors treated as a Blackbox to further strengthen their findings.
- The HoloClean system features one-shot automatic. The system was simulated by providing 1 constraint at a time.
- The measure was computed after every step.
- It was observed that I_d and I_P fall short of effectively indicating progress. Contrarily, the other measures, particularly I_R and $\mathcal{I}_{\mathcal{R}}^{\mathrm{lin}}$ are able to capture the reduction in the inconsistency level, and show an almost linear decay as desired.