Outline

- Biggest Takeaways, Strengths, Weaknesses
- Background
- Statistical Repair Model
- Reasoning over Metric FDs
- Experimental Setup and Results
- Conclusion
Biggest Takeaways

- There has been a lot of attention separately on logical and quantitative cleaning but this work combines the two.
- Metric FDs are very similar to traditional FDs but they allow for variation in the values of attributes.
- The authors present this algorithm as a set-minimal repair solution where the objective is minimizing statistical distortion, which in this case is EMD.
Strengths

- Accuracy study compared to Winsorization, another well-known repair strategy
- Two different approaches to injecting errors
  - Preserve $D_{\text{pres}}$
  - Use malicious $D_{\text{dest}}$
- Two real datasets: CORA and flights

Weaknesses

- A lot of the time, the difference in statistical distortion was small
- Preprocessing assumption
- Cannot combine attributes (i.e., x and y must be separate)
- Some steps/parts in the experimental results seem off (typos?) or don't have a clear purpose
- No real study of runtime or even general comparison against other algorithms
  - Other than for the Unified Repair Model
    - But no runtime analysis on this, just precision
Background

- Metric Schema
- Metric FDs
- Consistent Relation
- Repair
- Set-Minimal Repair
- EMD
- Statistical Distortion Minimal Repair
A metric schema is a set of attributes $R$ where for every attribute $A \in R$, we have metric $m_A$ and threshold $\theta_A \geq 0$. Each metric $m_A$ satisfies standard properties: symmetry, triangle inequality and identity of indiscernibles ($m_A(a, b) = 0$ iff $a = b$). Let $X \subseteq R = \{A_1, \ldots, A_n\}$. For two tuples $s, t$ over $R$, we write $s[X] \approx_{m_X, \theta_X} t[X]$ to mean $s[A_1] \approx_{m_{A_1}, \theta_{A_1}} t[A_1]$, $\ldots$, $s[A_n] \approx_{m_{A_n}, \theta_{A_n}} t[A_n]$. 
Metric FDs

- Generalize traditional FDs so that there is some variation in the values of the attributes that appear in their consequent.
- Good for domains where only large variations in the values indicate any real semantic difference.
- Synonyms allowed because each attribute value has a canonical name.

**Definition 2.2. (metric FD)**

Given a relation $D$ over a metric schema $R$, let $X$ and $Y$ be sets of attributes in $R$, $m_Y$ and $\Theta_Y$ be the metrics and thresholds for all the attributes in $Y$. Then, $X \rightarrow Y$ denotes a **metric FD**. A relation $D$ satisfies $X \rightarrow Y$ ($D \models X \rightarrow Y$), iff for all tuples $s, t \in D$, $s[X] = t[X]$ implies $s[Y] \approx_{m_Y, \Theta_Y} t[Y]$. 
Consistent Relation

- Given that $M$ is a set of metric FDs, a relation $D$ is consistent iff $M$ is true in every structure that $D$ is true.
- Otherwise, $D$ is inconsistent or dirty.
- $D$ is normalized so that attribute values are replaced by canonical names.

**Definition 2.4.** (consistent relation)

Given $M$, a set of metric FDs, a relation $D$ is **consistent** iff $D \models M$. Otherwise, $D$ is inconsistent (or dirty).
Repair

- Repairs done only via modification
  - Values in tuples may be changed but tuples themselves cannot be inserted or deleted
- Could do cardinality-minimal or set-minimal repairs

**Definition 2.5. (repair)**
A repair of an inconsistent relation $D_d$ is a consistent relation $D_r$ that can be created from $D_d$ using a set of value modifications $V$. 
Set-Minimal Repair

- A repair is set-minimal if no subset of the changed values can be reverted to their original values.
- Every repair is necessary so the choice of changing a value depends on the constraints.
- Can have some freedom in how to repair:
  - Selecting what values to use for modification
  - Minimizing statistical distortion (EMD) is the constraint here.

**Definition 2.6. (set-minimal repair)**

A repair $D_r$ created by a set of changes $V$ of an inconsistent relation $D_a$ is set-minimal iff there is no repair $D_{r'}$ that can be created by a strict subset $V' \subseteq V$. 
Other Types of Repair

- **Cardinality-Minimal Repair**: minimize $|\Delta(D, D')|$, where $\Delta(D, D')$ is the cells whose values differ in $D$, $D'$

- **Cardinality-Set-Minimal Repair**: $D'$ is set minimal iff there exists no $D''$ such that $\Delta(D, D'') \subset \Delta(D, D')$
  - This means that you can’t find a repair that changes a subset of the cells
Earth Mover’s Distance (EMD)

- A distance function that measures the dissimilarity of two histograms
- In this paper, this is based on the frequencies of the unique values in the dataset

**Definition 2.8. (Earth Mover’s Distance (EMD))**

Given two relational histograms, let $P = \{(p_1, w_{p_1}), ..., (p_m, w_{p_m})\}$ and $Q = \{(q_1, w_{q_1}), ..., (q_n, w_{q_n})\}$, each having $m$ and $n$ bins respectively. Define a cost matrix $C$, where $c_{i,j}$ is a measure of dissimilarity between $p_i$ and $q_j$, that models the cost of transforming $p_i$ to $q_j$, and a flow matrix $F$ where $f_{i,j}$ indicates the flow capacity between $p_i$ and $q_j$. EMD is defined in terms of an optimal flow that minimizes

$$d(P, Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{i,j} \times c_{i,j}$$

(1)

The EMD is defined as follows:

$$EMD(P, Q) = \min_F d(P, Q)$$

(2)

subject to the following constraints:

- $\forall i \in [1, m], \forall j \in [1, n]: f_{i,j} \geq 0,$
- $\forall i \in [1, m]: \sum_{j=1}^{n} f_{i,j} = w_{p_i},$
- $\forall j \in [1, n]: \sum_{i=1}^{m} f_{i,j} = w_{q_j}.$

(3)
Statistical Distortion Minimal Repair

- A set-minimal repair for which the EMD of two relations is minimal
  - One relation is ideal and the other relation is inconsistent with respect to a set of metric FDs
- Finding a statistical distortion minimal repair is an NP-hard problem

**Definition 2.9.** (statistical-distortion minimal repair)
Given an ideal relation $D_I$ and a relation $D_d$ that is inconsistent with respect to a set of metric FDs $M$, a statistical-distortion (SD) minimal repair is a set-minimal repair $D_r$ for which $EMD(D_r, D_I)$ is minimal.

**Theorem 2.10.** (complexity)
The problem of finding a SD minimal repair is NP-hard.
Algorithm Steps:

1. Identify a set of unresolved data.
2. Compute a set of tuple modifications to create sets of consistent candidate tuple repairs (t_r); each repair candidate is ensured to be minimal.
3. Choose the repair candidate which minimizes the statistical distortion.
4. Repeat for each tuple in the set of unresolved data
Define a set of metric FD’s M which is a minimal cover
Let $M[t_d] = \{m \in M | \{t_d\} \cup D_c \neq m\}$ be the set of dependencies that would be violated if inconsistent tuple $t_d$ were added unmodified to $D_c$
Let $U$ be the union of antecedents of $M[t_d]$ and $V$ be the union of consequents of $M[t_d]$
For each tuple, consider modifying attributes in $V$, then in $U$.
For consequent repairs, consider modifying $t_d[V]$. This means that we repair the tuple to have the same consequent as some other tuple in $D_c$ that shares its antecedent and then generate a set of candidate repairs.
Calculate the resulting statistical distortion for each candidate and choose the smallest. Rinse and repeat for the antecedent set.
If we fail to find a valid antecedent or consequent-repair, then we consider modifying values in Z, where Z is all attributes in M, known as both-repair. However, this leads to a large number of candidate tuple repairs, so the set must be limited in size to those repair tuples which are closest to $t_d$. As a final step, verify that the recommended repairs are set-minimal using a low-cost reversion step.
Rather than use the strict EMD algorithm to compute statistical distortion, the authors choose to use an approximation to reduce complexity.

Additionally, rather than compute EMD over groups of randomly chosen attributes, they define the set of attribute groups based on the closure of attributes used in $M[t_d]$ to take advantage of the natural attribute relationships that exist and have already been defined by the constraints.
To further reduce the complexity, the authors introduce additional parameters, β and ω.

- β captures the ratio of frequency of each bin in a distribution to the maximum bin frequency, and can be used to prune low frequency values from highly skewed distributions before computing EMD.
- ω handles the case where the distribution is relatively uniform. After pruning based on β, ω controls how many of the top bins are used for EMD.
The authors provide an axiomatization for metric FDs, providing a formal framework for reasoning about metric FDs.

**Theorem 4.2.** (soundness & completeness)
These axioms are sound and complete for metric FDs.
1. Identity: $\forall X \subseteq R, X \rightarrow X$
2. Decomposition: If $X \rightarrow YW$, then $X \rightarrow Y$
3. Composition: If $X \rightarrow Y$ and $Z \rightarrow W$ then $XZ \rightarrow YW$
4. Limited Reduce: If $XY \rightarrow Z, X \rightarrow Y$ and $\Theta_Y = 0$ then $X \rightarrow Z$

**Definition 4.3.** (closure $X^+$)
The closure of $X$, denoted $X^+$, with respect to a set of metric FDs $M$, is defined as $X^+ = \{A \mid M \vdash X \rightarrow A\}$.

**Lemma 4.4.** (closure)
$M \vdash X \rightarrow Y$ iff $Y \subseteq X^+$. 

Reasoning over Metric FDs - Metric FD Axiomatization
Reasoning over Metric FDs - Metric FD Inference Procedure

- Differential dependencies place an upper limit on the complexity of inference for metric FDs
- The authors show that their approach to inference for metric FDs remains linear w.r.t length of the dependencies in metric FD set M
Experimental Setup

- Intel Xeon X3470 with eight 3 GHz processors and 23 GB of RAM and Ubuntu 12.04 OS with all algorithms in Java
- UIS Database generator to generate datasets with Zipf distribution and N tuples
- Two real datasets: flight data and CORA
- Optimizations:
  - Compute EMD over attribute closures
  - Two pruning parameters: $\beta$ and $\omega$
- Creating dirty data with two different injections: consequent and antecedent violations
## Experimental Setup - Parameters

<table>
<thead>
<tr>
<th>Sym.</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td># of tuples</td>
<td>100K, 500K, 1M, 2M, 3M</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Zipf distribution</td>
<td>0.01, 0.25, 0.5, <strong>0.75</strong>, 0.99</td>
</tr>
<tr>
<td>$e$</td>
<td>error percentage</td>
<td>3%, <strong>5%</strong>, 7%, <strong>10%</strong></td>
</tr>
<tr>
<td>$F$</td>
<td># of FDs</td>
<td>1-10, default: <strong>3</strong></td>
</tr>
<tr>
<td>$\beta$</td>
<td>candidate pruning</td>
<td>0, <strong>30</strong>, <strong>50</strong>, 70%</td>
</tr>
<tr>
<td>$\varpi$</td>
<td>distribution pruning</td>
<td>20, <strong>50</strong>, 100, 500</td>
</tr>
</tbody>
</table>

Table 5.1: Parameters and defaults (bolded).
Experimental Results - Repair Quality

- **Accuracy Study**
  - Comparison against Winsorization: the presented algorithm does better
  - Another study to figure out distance: change errors in $D_{\text{ideal}}$
    - Changes in $D_{\text{ideal}}$ slightly affect distance
- **Distributions Preserving Data**
  - Inject errors that follow a distribution (two ways: use $D_{\text{preserve}}$ or malicious $D_{\text{destroy}}$)
  - Quality of repair is affected by injection method but only slightly
- **Comparative Study**
  - Against Unified Repair Model
  - The presented algorithm does slightly better (86.4% vs. 83.9%) for CORA
  - The presented algorithm does slightly worse (82.0% vs. 83.2%) for flight data
Experimental Results - Accuracy Study

(a) Vs Winsorization. (b) Num tuples vs time. Figure 5.2: Distance of repair & scalability.
Experimental Results - Scalability and Performance

- Pruning
  - Accuracy and runtime decreases moderately with $\beta$
  - $\omega$ has great effect on performance
  - Need to use one of the two pruning parameters in order for algorithm to finish

- Number of Tuples and Constraints
  - Quadratic because of EMD library

- Varying the Distribution and Error Rate
  - Variable that controls Zipf distribution ($\gamma$) makes runtime and number of iterations increase
  - Number of attributes impacts calculations significantly not as much as the active domain of each attribute
  - Domain has two values $\rightarrow$ EMD is fast; otherwise if domain is big $\rightarrow$ EMD is slow
Experimental Study - $\beta$

Figure 5.4: Effect of lower number of iterations.

(a) Vs time (min).

(b) Vs EMD distance.
Experimental Study - $\omega$

(a) Vs time (min).
(b) Vs EMD distance.

Figure 5.5: Effect of pruning lower frequency candidates.
Experimental Study - $\gamma$

(a) Number of constraints.  
(b) Varying the Zipf.  

Figure 5.6: Varying parameters vs time.
Experimental Study - Percentage Error

(a) Vs time (min).

(b) Vs EMD distance.

Figure 5.7: Percentage error.
Conclusion

- Metric FDs give traditional FDs some extensibility due to allowing variation in values.
- Statistical distortion (namely, EMD) is used as a constraint in the set-minimal repair problem presented here.
- Adding $\beta$ and $\omega$ as tunable parameters that modify the binned attribute sets reduces EMD complexity and preserves underlying statistical properties.
- Number of attributes impacts calculations significantly not as much as the active domain of each attribute.
Thanks!
Any questions?