Data Anonymization -

Generalization Algorithms

Li Xiong, Slawek Goryczka

CS573 Data Privacy and Anonymity
Generalization and Suppression

- **Generalization**: Replace the value with a less specific but semantically consistent value.

  - Z2 = \{410**\}
  - Z1 = \{4107*, 4109*\}
  - Z0 = \{41075, 41076, 41095, 41099\}

- **Suppression**: Do not release a value at all.

  - S0 = \{Male, Female\}
  - S1 = \{Person\}

<table>
<thead>
<tr>
<th>#</th>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41076</td>
<td>&lt; 40</td>
<td>*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>2</td>
<td>48202</td>
<td>&lt; 40</td>
<td>*</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>3</td>
<td>41076</td>
<td>&lt; 40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>4</td>
<td>48202</td>
<td>&lt; 40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
Complexity

Search Space:

• Number of generalizations = \( \prod_{\text{attrib } i} \) (Max level of generalization for attribute i + 1)

If we allow generalization to a different level for each value of an attribute:

• Number of generalizations = \( \prod_{\text{attrib } i} \) (Max level of generalization for attribute i + 1) #tuples
Hardness result

Given some data set $R$ and a QI $Q$, does $R$ satisfy $k$-anonymity over $Q$?

- Easy to tell in polynomial time, NP!

Finding an *optimal* anonymization is not easy

- NP-hard: reduction from $k$-dimensional perfect matching
- A polynomial solution implies $P = NP$

---

Anonymization Strategies

- **Local suppression**
  - Delete individual attribute values
  - e.g. <Age=50, Gender=M, State=CA>

- **Global attribute generalization**
  - Replace specific values with more general ones for an attribute
  - Numeric data: partitioning of the attribute domain into intervals, e.g., Age = {[1-10], ..., [91-100]}
  - Categorical data: generalization hierarchy supplied by users, e.g., Gender = {M, F}
\(k\)-Anonymization with Suppression

- \(k\)-Anonymization with suppression
  - Global attribute generalization with local suppression of outlier tuples.

- Terminologies
  - Dataset: D
  - Anonymization: \(\{a_1, \ldots, a_m\}\)
  - Equivalent classes: E
Finding Optimal Anonymization

- Optimal anonymization determined by a cost metric
- Cost metrics
  - Discernability metric: penalty for non-suppressed tuples and suppressed tuples
    \[
    C_{DM}(g, k) = \sum_{\forall E \text{ s.t. } |E| \geq k} |E|^2 + \sum_{\forall E \text{s.t. } |E| < k} |D||E|
    \]
- Classification metric

Modeling Anonymizations

- Assume a total order over the set of all attribute domains
- Set representation for anonymization
  - e.g., Age: <[10-29], [30-49]>, Gender: <[M or F]>, Marital Status: <[Married], [Widowed or Divorced], [Never Married]>
  - \{1, 2, 4, 6, 7, 9\} -> \{2, 7, 9\}
- Power set representation for entire anonymization space
  - Power set of \{2, 3, 5, 7, 8, 9\} - order of \(2^n\)!
  - \{\} – most general anonymization
  - \{2,3,5,7,8,9\} – most specific anonymization
Optimal Anonymization Problem

- **Goal**
  - Find the best anonymization in the powerset with the lowest cost

- **Algorithm**
  - Set enumeration search through tree expansion - size $2^n$
  - Top-down depth first search

- **Heuristics**
  - Cost-based pruning
  - Dynamic tree rearrangement

Set enumeration tree over powerset of \{1,2,3,4\}
Node Pruning through Cost Bounding

- **Intuitive idea**
  - prune a node $H$ if none of its descendents can be optimal

- **Cost lower-bound of subtree of $H$**
  - Cost of suppressed tuples bounded by $H$
  - Cost of non-suppressed tuples bounded by $A$

\[
LB_{DM}(H, A) = \sum_{\forall t \in D} \begin{cases} |D| & \text{when } t \text{ is suppressed by } H, \\ \max(|E_A, t|, k) & \text{otherwise}. \end{cases}
\]
Useless Value Pruning

- Intuitive idea
  - Prune useless values that have no hope of improving cost

- Useless values
  - Only split equivalence classes into suppressed equivalence classes (size < k)
Tree Rearrangement

- Intuitive idea
  - Dynamically reorder tree to increase pruning opportunities

- Heuristics
  - Sort the values based on the number of equivalence classes induced
Comments

- Interesting things to think about
  - Domains without hierarchy or total order restrictions
  - Other cost metrics
  - Global generalization vs. local generalization
Taxonomy of Generalization Algorithms

- Top-down specialization vs. bottom-up generalization
- Global (single dimensional) vs. local (multi-dimensional)
- Complete (optimal) vs. greedy (approximate)
- Hierarchy-based (user defined) vs. partition-based (automatic)

K. LeFerve, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient Full-Domain $k$-Anonymity. In SIGMOD 05
Generalization algorithms

- Early systems
  - μ-Argus, Hundpool, 1996 - Global, bottom-up, greedy
  - Datafly, Sweeney, 1997 - Global, bottom-up, greedy

- k-Anonymity algorithms
  - AllMin, Samarati, 2001 - Global, bottom-up, complete, impractical
  - MinGen, Sweeney, 2002 - Global, bottom-up, complete, impractical
  - Bottom-up generalization, Wang, 2004 – Global, bottom-up, greedy
  - TDS (Top-Down Specialization), Fung, 2005 - Global, top-down, greedy
  - K-OPTIMIZE, Bayardo, 2005 – Global, top-down, partition-based, complete
  - Incognito, LeFevre, 2005 – Global, bottom-up, hierarchy-based, complete
  - Mondrian, LeFevre, 2006 – Local, top-down, partition-based, greedy
Mondrian

- Top-down partitioning
- Greedy
- Local (multidimensional) – tuple/cell level
Global Recoding

- Mapping domains of quasi-identifiers to generalized or altered values using a *single* function

- Notation
  - $D_{x_i}$ is the domain of attribute $X_i$ in table T

- Single Dimensional
  - $\phi_i : D_{x_i} \rightarrow D'$ for each attribute $X_i$ of the quasi-id
  - $\phi_i$ applied to values of $X_i$ in tuple of T
Local Recoding

- **Multi-Dimensional**
  - Recode domain of value vectors from a set of quasi-identifier attributes
  - $\phi : D_{x1} \times \ldots \times D_{xn} \rightarrow D'$
  - $\phi$ applied to vector of quasi-identifier attributes in each tuple in $T$
Partitioning

- Single Dimensional
  - For each $X_i$, define non-overlapping single dimensional intervals that covers $D_{x_i}$
  - Use $\phi_i$ to map $x \in D_x$ to a summary stat

- Strict Multi-Dimensional
  - Define non-overlapping multi-dimensional intervals that covers $D_{x_1} \ldots D_{x_d}$
  - Use $\phi$ to map $(x_{x_1} \ldots x_{x_d}) \in D_{x_1} \ldots D_{x_d}$ to a summary stat for its region
Global Recoding Example

**k = 2**

**Quasi Identifiers**
Age, Sex, Zipcode

**Partitions**

**Single Dimensional**

**Partitions**
- Age : {[25-28]}
- Sex: {Male, Female}
- Zip : {[53710-53711], 53712}

**Multi-Dimensional**

**Partitions**
- {Age: [25-26], Sex: Male, Zip: 53711}
- {Age: [25-27], Sex: Female, Zip: 53712}
- {Age: [27-28], Sex: Male, Zip: [53710-53711]}
Global Recoding Example 2

$k = 2$
Quasi Identifiers
Age, Zipcode

Patient Data
Single Dimensional
Multi-Dimensional
Greedy Partitioning Algorithm

Problem
- Need an algorithm to find multi-dimensional partitions
- Optimal $k$-anonymous strict multi-dimensional partitioning is NP-hard

Solution
- Use a greedy algorithm
- Based on k-d trees
- Complexity $O(n \log n)$
Anonymize(partition)
    if (no allowable multidimensional cut for partition)
        return φ : partition → summary
    else
        dim ← choose_dimension()
        fs ← frequency_set(partition, dim)
        splitVal ← find_median(fs)
        lhs ← \{t ∈ partition : t.dim ≤ splitVal\}
        rhs ← \{t ∈ partition : t.dim > splitVal\}
        return Anonymize(rhs) ∪ Anonymize(lhs)
Algorithm Example

- $k = 2$
- Dimension determined heuristically
- Quasi-identifiers
  - Zipcode
  - Age

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>25</td>
<td>Female</td>
<td>53712</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>26</td>
<td>Male</td>
<td>53711</td>
<td>Brochitis</td>
</tr>
<tr>
<td>27</td>
<td>Male</td>
<td>53710</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>27</td>
<td>Female</td>
<td>53712</td>
<td>AIDS</td>
</tr>
<tr>
<td>28</td>
<td>Male</td>
<td>53711</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>Zipcode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-26</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>25-27</td>
<td>Female</td>
<td>53712</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>25-26</td>
<td>Male</td>
<td>53711</td>
<td>Brochitis</td>
</tr>
<tr>
<td>27-28</td>
<td>Male</td>
<td>[53710-53711]</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>25-27</td>
<td>Female</td>
<td>53712</td>
<td>AIDS</td>
</tr>
<tr>
<td>27-28</td>
<td>Male</td>
<td>[53710-53711]</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

Patient Data  Anonymized Data
Algorithm Example

Iteration # 1 (full table)

**partition**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>25</td>
<td>Female</td>
<td>53712</td>
<td>Hepatitis</td>
</tr>
<tr>
<td>26</td>
<td>Male</td>
<td>53711</td>
<td>Bronchitis</td>
</tr>
<tr>
<td>27</td>
<td>Male</td>
<td>53710</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>27</td>
<td>Female</td>
<td>53712</td>
<td>AIDS</td>
</tr>
<tr>
<td>28</td>
<td>Male</td>
<td>53711</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

**LHS**

**RHS**

- **dim** = Zipcode
- **fs**
  - **splitVal** = 53711

<table>
<thead>
<tr>
<th>Zipcode</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>53710</td>
<td>1</td>
</tr>
<tr>
<td>53711</td>
<td>3</td>
</tr>
<tr>
<td>53712</td>
<td>2</td>
</tr>
</tbody>
</table>
Algorithm Example continued

Iteration # 2 (LHS from iteration # 1)

**partition**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>26</td>
<td>Male</td>
<td>53711</td>
<td>Bronchitus</td>
</tr>
<tr>
<td>27</td>
<td>Male</td>
<td>53710</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>28</td>
<td>Male</td>
<td>53711</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

**LHS**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>26</td>
<td>Male</td>
<td>53711</td>
<td>Bronchitus</td>
</tr>
</tbody>
</table>

**RHS**

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Male</td>
<td>53710</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>28</td>
<td>Male</td>
<td>53711</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

**fs**

<table>
<thead>
<tr>
<th>Age</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
</tr>
</tbody>
</table>

**dim = Age**

**splitVal = 26**
Algorithm Example

Iteration # 3 (LHS from iteration # 2)

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>53711</td>
<td>Flu</td>
</tr>
<tr>
<td>26</td>
<td>Male</td>
<td>53711</td>
<td>Bronchitis</td>
</tr>
</tbody>
</table>

No Allowable Cut

Summary: Age = [25-26] Zip= [53711]

Iteration # 4 (RHS from iteration # 2)

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>ZipCode</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>Male</td>
<td>53710</td>
<td>Broken Arm</td>
</tr>
<tr>
<td>28</td>
<td>Male</td>
<td>53711</td>
<td>Hang Nail</td>
</tr>
</tbody>
</table>

No Allowable Cut

Summary: Age = [27-28] Zip= [53710 - 53711]
Algorithm Example  continued

Iteration # 5 (RHS from iteration # 1)

**partition**

Summary: Age = [25-27] Zip= [53712]
Experiment

- Adult dataset
- Data quality metric (cost metric)
  - Discernability Metric ($C_{DM}$)
    - $C_{DM} = \sum_{\text{EquivalentClasses } E} |E|^2$
    - Assign a penalty to each tuple
  - Normalized Avg. Eqiv. Class Size Metric ($C_{AVG}$)
    - $C_{AVG} = (\text{total}\_\text{records}/\text{total}\_\text{equiv}\_\text{classes})/k$
Comparison results

- Full-domain method: Incognito
- Single-dimensional method: K-OPTIMIZE

Figure 10. Quality comparison for Adults database using discernability metric
Data partitioning comparison

(a) Optimal single-dimensional partitioning

(b) Greedy strict multidimensional partitioning
Mondrian

Piet Mondrian [1872-1944]
Distributed Anonymization

anonymize-and-aggregate

aggregate-and-anonymize
Privacy is defined as $k$-anonymity ($k = 2$).

<table>
<thead>
<tr>
<th>Provider</th>
<th>Name</th>
<th>Age</th>
<th>Zip</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Alice</td>
<td>[20-30]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Emily</td>
<td>[20-30]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Sara</td>
<td>[20-30]</td>
<td>*****</td>
<td>Epilepsy</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Bob</td>
<td>[31-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>John</td>
<td>[31-35]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Olga</td>
<td>[31-35]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Frank</td>
<td>[31-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Dorothy</td>
<td>[36-40]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Mark</td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Cecilia</td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
</tbody>
</table>
Anonymization Example (attack)

Privacy is defined as $k$-anonymity ($k = 2$).

<table>
<thead>
<tr>
<th>Provider</th>
<th>Name</th>
<th>$T_a^*$</th>
<th>Age</th>
<th>Zip</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Alice</td>
<td></td>
<td>[20-30]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Emily</td>
<td></td>
<td>[20-30]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Sara</td>
<td></td>
<td>[20-30]</td>
<td>*****</td>
<td>Epilepsy</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Bob</td>
<td></td>
<td>[21-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>John</td>
<td></td>
<td>[31-35]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Olga</td>
<td></td>
<td>[31-35]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Frank</td>
<td></td>
<td>[31-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Dorothy</td>
<td></td>
<td>[36-40]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Mark</td>
<td></td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Cecilia</td>
<td></td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
</tbody>
</table>
Privacy is defined as $k$-anonymity ($k = 2$).

<table>
<thead>
<tr>
<th>Provider</th>
<th>Name</th>
<th>Age</th>
<th>Zip</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Alice</td>
<td>[20-30]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Emily</td>
<td>[20-30]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Sara</td>
<td>[20-30]</td>
<td>*****</td>
<td>Epilepsy</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Bob</td>
<td>[31-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>John</td>
<td>[31-35]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Olga</td>
<td>[31-35]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Frank</td>
<td>[31-35]</td>
<td>*****</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Dorothy</td>
<td>[36-40]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Mark</td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Cecilia</td>
<td>[36-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
</tbody>
</table>
$m$-Privacy

A set of anonymized records is $m$-private with respect to a privacy constraint $C$, e.g., $k$-anonymity, if any coalition of $m$ parties ($m$-adversary) is not able to breach privacy of remaining records.
**m-Anonymization Example**

- An attacker is a single data provider (1-privacy)

<table>
<thead>
<tr>
<th>Provider</th>
<th>Name</th>
<th>Age</th>
<th>Zip</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>Alice</td>
<td>[20-40]</td>
<td>*****</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Mark</td>
<td>[20-40]</td>
<td>*****</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Sara</td>
<td>[20-40]</td>
<td>*****</td>
<td>Epilepsy</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Emily</td>
<td>[20-40]</td>
<td>987**</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>Dorothy</td>
<td>[20-40]</td>
<td>987**</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_3$</td>
<td>Cecilia</td>
<td>[20-40]</td>
<td>987**</td>
<td>Flu</td>
</tr>
<tr>
<td>$P_1$</td>
<td>Bob</td>
<td>[20-40]</td>
<td>123**</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Olga</td>
<td>[20-40]</td>
<td>123**</td>
<td>Cancer</td>
</tr>
<tr>
<td>$P_4$</td>
<td>Frank</td>
<td>[20-40]</td>
<td>123**</td>
<td>Asthma</td>
</tr>
<tr>
<td>$P_2$</td>
<td>John</td>
<td>[20-40]</td>
<td>123**</td>
<td>Flu</td>
</tr>
</tbody>
</table>
Parameters $m$ and $C$

- Number of malicious parties: $m$
  - $m = 0$ (0-privacy) is when the coalition of parties is empty, but each data recipient can be malicious
  - $m = n-1$ means that no party trusts any other (anonymize-and-aggregate)

- Privacy constraint $C$:
  - $m$-privacy is orthogonal to $C$ and inherits all its advantages and drawbacks
$m$-Adversary Modeling

- If a coalition of attackers cannot breach privacy of records, then any its subcoalition will not be able to do so as well.
Equivalence Group Monotonicity

- Adding new records to a private equiv. group will not change the privacy fulfillment!
- To verify $m$-privacy it is enough to determine privacy fulfillment only for $m$-adversaries,
- EG monotonic privacy constraints: $k$-anonymity, simple $l$-diversity, ...
- Not EG monotonic constraints: $t$-closeness, ...
Pruning Strategies

- Number of coalitions to verify: exponential to number of providers, but with efficient pruning strategies should be OK!
Verification Algorithms

- top-down algorithm,
- bottom-up algorithm,
- binary algorithm.
Anonymizer for $m$-Privacy

To multidimensional data add one more attribute – data provider, which can be used as any other attribute in anonymization.
Anonymizer for $m$-Privacy

- To multidimensional data add one more attribute – data provider, which can be used as any other attribute in anonymization.
Anonymizer for $m$-Privacy

To multidimensional data add one more attribute – data provider, which can be used as any other attribute in anonymization.
$m$-Anonymizer (diagram)
Experiments Setup

- Dataset: the *Adult* dataset, Census database.
- Attributes: age, workclass, education, marital-status, race, gender, native-country, occupation (sensitive attribute with 14 possible values).
- Privacy defined as a conjunction of $k$-anonymity and $l$-diversity.
- Metrics:
  - Runtime
  - Query error – compares results of random queries issued over original and anonymized data
Experiments

- $m$-Privacy verification runtime for different algorithms vs. $m$

Average number of records per provider = 10

Average number of records per provider = 50
Experiments

- \( m \)-Anonymizer runtime and query error for different anonymizers vs. size of attacking coalitions \( m \)
Experiments

- $m$-Anonymizer runtime and query error for different anonymizers vs. number of data records
Thank you!