Practical Issues in Anonymity

Chris Clifton, Shawn Merrill (Purdue University)
Keith Merrill (Brandeis University)

Problems with Anonymization

RE-IDENTIFICATION! Many anonymization schemes keeps being “broken”, eventually people find sufficient data to link/re-identify (e.g. $k$-anonymity)

But, there is still a use case: Private use under a data use agreement

• Want to provide protection against accidental (or low resource) re-identification
• Contractual data use agreement to “pull back” data if linking datasets found
Problems with Anonymization

Even if we aren't concerned about re-identification
- Anonymization algorithm impacts practical utility more than value of “utility metric” (Nergiz & Clifton 2007)
- Choice of (user-defined) Generalization Hierarchy has even greater impact on utility
- Difficulties with global generalization scaling on large datasets
  - Efficiency
  - Utility
  - Outliers

Need to sanitize the data in a way that preserves its use for the recipient.

Issue: Poor Utility from Bad Generalization Hierarchies

- Example: Issue with poor generalization hierarchy
  - Million-record anonymization of health data
  - Initial hierarchy (straightforward splits): minimum group size of 48, even with k=2
  - Improved hierarchy (data-dependent) showed significantly better granularity
    - Differences each level of k [2, 4, 6, 8, 10, 20]
- Similar issues arise with differential privacy
  - Higher relative noise for small groups, even for histogram
Hierarchy Example: How Anonymization Can Go Awry

A college town will have a different age distribution than a retirement community.

• Given this Hierarchy:

```
+-------------------+
|                  65+ |
|                   |
+--------------------+
|            20-65   |
|                   |
+--------------------+
|      20-24 25-29  |
|                   |
+--------------------+
|    80-84          |
```

• The presence of few (<k) 80-84 year olds forces everything to be generalized to “working age” and “retirement age”

User-Defined Hierarchy: Issues

• Relies on a curator’s knowledge of the data
• Too data-driven causes significant information leaks
  – Similar problems to local recoding, clustered anonymization
• Context-insensitivity can lead to issues like semantic similarity among attributes
  – deFinetti Attack
• Can vary greatly based on the attribute [age vs. zip code vs. car type] and specifics [Lafayette, IN vs. Lafayette, LA]
Hierarchy Example

A better generalization hierarchy:
- Must be made without direct use of the data
  - Minimality attacks
- Can be done with relative frequency of the values in the population

* A case for differential privacy?

Further Issues

- Release of data at multiple levels
  - Potential interactions impact privacy
  - Inter-level consistency can improve results
- Lattice rather than hierarchy
Issue: Scaling

- Challenge: Difficulty in scaling generalization-based anonymization to million record dataset
  - Many techniques fail
  - Few that succeed result in significant record suppression
- Idea: Independently anonymize partitions
  - Potential for different generalizations for different partitions
    - *Will this reduce suppression?*
  - Agnostic to algorithm, privacy definition

Definition: Parallel Composition

We say that a sanitization scheme \( A \) satisfies parallel composition if, given disjoint datasets \( D_1, \ldots, D_n \) with corresponding outputs \( A(D_i), \bigcup_{i=1}^n A(D_i) \) satisfies the privacy guarantee of the original scheme.

- Satisfied by:
  - Differential Privacy (*McSherry SIGMOD'09*)
    - Privacy budget treated independently for each dataset
  - Generalization-based \( k \)-anonymity, \( l \)-diversity with local recording
- Not satisfied by
  - Generalization-based anonymization with global recording
  - \( t \)-closeness
Definition: Partitioned Preprocessing

Choose a random partition \( \{d_i\} \) of \(|D|\) into positive integers, then partition \( D \) into pieces \( D_i \) of size \( d_i \) uniformly at random. We call \( \bigcup_{i=1}^{n} A(D_i) \) a **partitioned preprocessing** dataset.

- Works for parallel composition techniques
- Potentially stronger against some types of attacks on generalization
  - Minimality
  - deFinetti
- Attack resistance arguments hold for non-parallel decomposable techniques
  - E.g., global recoding (and potential utility benefits)

---

**Partitioned Preprocessing: Potential Utility Benefit**

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Zip</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-50</td>
<td>Male</td>
<td>92***</td>
<td>Yes</td>
</tr>
<tr>
<td>40-50</td>
<td>Male</td>
<td>92***</td>
<td>No</td>
</tr>
<tr>
<td>40-50</td>
<td>Male</td>
<td>92***</td>
<td>No</td>
</tr>
<tr>
<td>40-50</td>
<td>Male</td>
<td>92***</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Zip</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>40-60</td>
<td>Male</td>
<td>925**</td>
<td>No</td>
</tr>
<tr>
<td>40-60</td>
<td>Male</td>
<td>925**</td>
<td>No</td>
</tr>
<tr>
<td>40-60</td>
<td>Male</td>
<td>925**</td>
<td>Yes</td>
</tr>
<tr>
<td>40-60</td>
<td>Male</td>
<td>925**</td>
<td>No</td>
</tr>
</tbody>
</table>

- Some benefits of local recoding
  - “Outliers” only force over-generalization in a single partition
- Each partition satisfies global recoding
  - Difficulty identifying which partition an item belongs to provides defense against attacks
Partitioned Preprocessing: Example

Semantic Attacks: Determine likely distribution of sensitive values in an equivalence class

- Individual may belong to many equivalence classes
  - Attack gives information on one equivalence class
- Attack increases $Pr(x. S = S_i)$ by only a (weighted) proportion of the increase in probability for that class

<table>
<thead>
<tr>
<th>k=20</th>
<th>Underlying Partitions</th>
<th>Visible Partitions</th>
<th>Distribution of Partitions</th>
<th>% of Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average 25,000 size</td>
<td>20</td>
<td>6 + Suppressed Class</td>
<td>6, 5, 6, 1, 1, 1</td>
<td>.244, .30, .295, .062, .048, .024 Suppress: .016</td>
</tr>
</tbody>
</table>

Partitioned Preprocessing: Example

- Original Record:

<table>
<thead>
<tr>
<th>ZIP</th>
<th>YOB</th>
<th>GEN</th>
<th>VISIT</th>
<th>HOSPITAL</th>
<th>COMP</th>
<th>CAT</th>
<th>Possible Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>43125</td>
<td>1967</td>
<td>F</td>
<td>2005-08-31</td>
<td>Riverside Methodist</td>
<td>Mosquito Bite</td>
<td>Other</td>
<td>7,916</td>
</tr>
</tbody>
</table>

- Anonymized Versions:

<table>
<thead>
<tr>
<th>ZIP</th>
<th>YOB</th>
<th>Visit Date</th>
<th>Hospital</th>
<th>Matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>43000 - 43240</td>
<td>1940 - 1979</td>
<td>2004-01-01 - 2005-12-31</td>
<td>Riverside Methodist Hospital</td>
<td>2520</td>
</tr>
<tr>
<td>43068 - 43156</td>
<td>1940 - 1979</td>
<td>2004-01-01 - 2005-12-31</td>
<td>Medium &amp; Large Hospitals</td>
<td>3497</td>
</tr>
<tr>
<td>43068 - 43156</td>
<td>1900 - 1992</td>
<td>2004-01-01 - 2005-12-31</td>
<td>Riverside Methodist Hospital</td>
<td>1068</td>
</tr>
<tr>
<td>43119 - 43156</td>
<td>1940 - 1979</td>
<td>2004-01-01 - 2008-02-31</td>
<td>Large Hospitals</td>
<td>421</td>
</tr>
<tr>
<td>43119 - 43156</td>
<td>1900 - 1992</td>
<td>2005-07-01 - 2005-12-31</td>
<td>Medium &amp; Large Hospitals</td>
<td>169</td>
</tr>
<tr>
<td>43068 - 43156</td>
<td>1900 - 1992</td>
<td>2004-01-01 - 2005-12-31</td>
<td>Large Hospitals</td>
<td>241</td>
</tr>
</tbody>
</table>
• Implications of partitioned preprocessing on differential privacy
  – Near-optimal use of privacy budget
    • Use noise from random partitioning to satisfy differential privacy
  – Potential operational value?
  – Amplification of privacy budget through sampling
• Implications of hierarchies on a differentially private census
  – Appropriate hierarchies, top-coding
  – Any “non-histogram” analyses?