Disclosure Metrics Born From Statistical Evaluations of Data Utility

Presented by: Devyani Biswal (University of Ottawa)
In collaboration with: Rafal Kulik (University of Ottawa), Luk Arbuckle (Privacy Analytics)
Agenda

• Background
• Privacy Models
• Empirical Results
• Ongoing Research
Context and Motivation

BACKGROUND
**Context**

- **k-anonymity**: Defined for categorical data.
- **Numerical data**: Generalize for k-anonymity.
- **Noise injection**: Applied blindly, won’t achieve k.
Context and Motivation

**Context**
- **k-anonymity**
  Defined for categorical

- **Numerical data**
  Generalize for k-anonymity

- **Noise injection**
  Applied blindly, won’t achieve k

**Motivation**
- **Data utility**
  Improve over existing

- **Privacy bound**
  Use the same k

- **Recipient**
  Produce meaningful statistics

**Research**
Theoretical

PRIVACY MODELS
Categorical Data

\[ \text{Input: } X = (X_1, \ldots, X_n) \]
\[ \text{Output: } X^r = (X_1^r, \ldots, X_n^r) \]

\[ X_{(1)} \rightarrow G_1 = [X_{(1)}, X_{(i)}] \rightarrow G_1^r = [X_1^r, X_i^r] \]
\[ \vdots \]
\[ X_{(n)} \rightarrow G_n = [X_{(j)}, X_{(n)}] \rightarrow G_n^r = [X_{(j)}^r, X_n^r] \]

\[ X_{(i)} \rightarrow X_{(i)}^r \text{ with uniform probability.} \]
Numerical Data

**K-Noise**

**Input:** $X = (X_1, ..., X_n)$

**Output:** $X^r = (X_1^r, ..., X_n^r)$

- $X_1$ → $X_1 + \text{Uni}(-a,a)$ → $X_1^r$
- $X_n$ → $X_n + \text{Uni}(-a,a)$ → $X_n^r$
Experiments

EMPIRICAL RESULTS
Baseline Level of Privacy

Method: k-PRAM

Method: k-noise
Baseline Level of Privacy

Expected number or records within a neighbourhood of \([-2.5,2.5]\) years of each randomized record in X.
## Data Utility Measures

<table>
<thead>
<tr>
<th>Method</th>
<th>Bias</th>
<th>Mse</th>
<th>Rmse</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-PRAM</td>
<td>0.076918689</td>
<td>4.368601</td>
<td>2.090120</td>
</tr>
<tr>
<td>k-Noise</td>
<td>-0.008230827</td>
<td>2.184711</td>
<td>1.478077</td>
</tr>
</tbody>
</table>
Data Utility Visual

The clustered scatter plot represents k-PRAM randomized individuals.

The bias calculated is visibly noticeable when comparing the two methods against the true value.
Data Utility Densities
Ongoing research

**Multidimensional**
Extend to handle correlations, sparsity, etc.

**Adaptive**
Localized noise injection based on empirical distribution

**Distributions**
Impact on data utility from different noise profiles