Probabilistic $k^m$-anonymity
(Efficient Anonymization of Large Set-valued Datasets)

Gergely Acs (INRIA)
gergly.acs@inria.fr

Jagdish Achara (INRIA)
jagdish.achara@inria.fr

Claude Castelluccia (INRIA)
claude.castelluccia@inria.fr
Overview

- Motivation
- Background: $k^m$-anonymity
- Why $k^m$-anonymity is impractical?
- Relaxation of $k^m$-anonymity: Probabilistic $k^m$-anonymity
- How to anonymize to have probabilistic $k^m$-anonymity?
- Performance evaluation
- Conclusions
De-identification

- **Personal data** is any information relating to an identified or identifiable individual (EU Directive 95/46/EC)

- **De-identification** breaks links between individuals’ identity and their data (records)

- Regulations apply only to **personal data**! 
  **De-identified data is non-personal data** and hence out of the regulation

- NOTE: de-identification does NOT include the control of (sensitive) attribute inference
No direct Personal ID in the dataset (e.g., phone numbers)

Each user has a subset of items (e.g., visited locations, watched movies, purchased items, etc.)

High-dimensional and sparse data!


Privacy test: Location uniqueness

- Derived from Call Data Records
- 4,427,486 users
- 1303 towers (i.e., locations)
- Mean tower # per user: 11.42 (std.dev: 17.23)
- Max. tower # user: 422
Privacy test: Location uniqueness

- If the adversary knows \( m \) towers of a user, what is the probability that the user is the only one who have these towers in the dataset?

- **Similar study:**
  
Background: $k^m$-anonymity

- For ANY $m$ items, there are at least $k$ users who have these items
  - if $m$ equals the maximum item number per user, then $k^m$ is equivalent to $k$-anonymity
  - However, $k$-anonymity suffers from the curse of dimensionality\cite{Aggarwal2005} (i.e., very bad utility for high-dimensional, sparse data)

- Rationale of $k^m$-anonymity: adversary is unlikely to know all the items of a user

- Allows larger utility by applying fewer generalizations (aggregations)

\cite{Aggarwal2005} C. C. Aggarwal, *On K-anonymity and the Curse of Dimensionality*, VLDB, 2005
### Example: $k$ vs. $k^m$-anonymity

<table>
<thead>
<tr>
<th>Rec#</th>
<th>Original Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{LA}</td>
</tr>
<tr>
<td>2</td>
<td>{LA, Seattle}</td>
</tr>
<tr>
<td>3</td>
<td>{New York, Boston}</td>
</tr>
<tr>
<td>4</td>
<td>{New York, Boston}</td>
</tr>
<tr>
<td>5</td>
<td>{LA, Seattle, New York}</td>
</tr>
<tr>
<td>6</td>
<td>{LA, Seattle, New York}</td>
</tr>
<tr>
<td>7</td>
<td>{LA, Seattle, New York, Boston}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rec#</th>
<th>2-anonymity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{East US}</td>
</tr>
<tr>
<td>2</td>
<td>{East US}</td>
</tr>
<tr>
<td>3</td>
<td>{West US}</td>
</tr>
<tr>
<td>4</td>
<td>{West US}</td>
</tr>
<tr>
<td>5</td>
<td>{LA, Seattle, West US}</td>
</tr>
<tr>
<td>6</td>
<td>{LA, Seattle, West US}</td>
</tr>
<tr>
<td>7</td>
<td>{LA, Seattle, West US}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rec#</th>
<th>$2^2$-anonymity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{LA}</td>
</tr>
<tr>
<td>2</td>
<td>{LA, Seattle}</td>
</tr>
<tr>
<td>3</td>
<td>{West US}</td>
</tr>
<tr>
<td>4</td>
<td>{West US}</td>
</tr>
<tr>
<td>5</td>
<td>{LA, Seattle, West US}</td>
</tr>
<tr>
<td>6</td>
<td>{LA, Seattle, West US}</td>
</tr>
<tr>
<td>7</td>
<td>{LA, Seattle, West US}</td>
</tr>
</tbody>
</table>

![Diagram of US regions with LA, Seattle, New York, Boston, and All (US) connections]
Problem of $k^m$-anonymity

- Verifying $k^m$-anonymity can have exponential complexity in $m$ \[^1\] if $m$ is large (typically when $m \geq 5$)

- The exact speed depends on the structure of the generalization hierarchy and the dataset itself\[^1\]

\[ \Rightarrow \text{DOES NOT WORK FOR MANY REAL-WORLD DATASETS!} \]

Probabilistic $k^m$-anonymity

- For **ANY** $m$ items, there are at least $k$ users who have these items with probability at least $p$
  - where $p > 0.9$, and typically should be around 0.99 or 0.999

- Intuition: instead of checking all possible $m$ items, we select **randomly** some of them from the dataset, and check $k$-anonymity of **only** these samples!
  - we have $k$-anonymity for **ANY randomly** selected $m$ items with large probability (based on sampling theorems)!

- How to sample these $m$ items?
- How many samples are needed?
How to sample $m$-itemsets?

- **Naïve approach:**
  1. Sample a record
  2. Sample $m$ items from this record

  Biased towards selecting more popular itemsets!
  (e.g., popular places in location data)

- However, adversary may learn unpopular items easily
  e.g., home address is not necessarily popular...

- **Our approach** is more general:
  Select among all $m$-itemsets uniformly at random using a fast-mixing Markov chain

  Adversary can learn any $m$-itemset with equal probability!
How many samples?

- From the Chernoff-Hoeffding bound:
  \[ N = O \left( (1 - p)^{-2} \ln \left( \frac{1}{1 - p} \right) \right) \]
  
to have \( k^m \)-anonymity with probability \( p \)

- \textit{Independent from} \( m \), the dataset size, and the number of all items!

<table>
<thead>
<tr>
<th>( p )</th>
<th>( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>( \approx 60 \text{ K} )</td>
</tr>
<tr>
<td>0.999</td>
<td>( \approx 5 \text{ M} )</td>
</tr>
<tr>
<td>1</td>
<td>( \infty )</td>
</tr>
</tbody>
</table>
Anonymization

**INPUT**: $p$ – probability, $k, m$ – privacy parameters, $D$ – dataset

1. **SAMPLING**: Pick (uniformly at random) a single $m$-itemset from $D$ using MCMC sampling

2. **IF** the sample does NOT satisfy $k$-anonymity
   **GENERALIZE** an item in the sample such that generalization error is minimized (e.g., average cell size in location data)

3. **REPEAT** the above steps until $O \left( (1 - p)^{-2} \ln \left( \frac{1}{1 - p} \right) \right)$ consecutive samples satisfy $k$-anonymity

**AMPLIFY UTILITY**: Execute the above algorithm multiple times and select the one which has the least generalization error
Running complexity

- The required number of samples which must satisfy k-anon. is
  \[ N = O \left( (1 - p)^{-2} \ln \left( \frac{1}{1 - p} \right) \right) \]

- For each sample, the Markov chain sampling runs in
  \[ O(m^2 |D|) \]

- The maximum number of generalizations is the number of possible items which is \( O(|\mathbb{I}|) \)

- Hence, the total complexity is
  \[ O \left( m^2 |D| |\mathbb{I}| (1 - p)^{-2} \ln \left( \frac{1}{1 - p} \right) \right) \]
  \( \Rightarrow \) polynomial in the number of records \(|D|\), number of possible items \(|\mathbb{I}|\), \(m\), and probability \(p\)
Performance evaluation: Privacy guarantee

RECALL: a user has fewer than 11 visited towers on average

- We can have different privacy guarantee (i.e., $k$, $p$) for different $m$!
- In the evaluation:
  - when $m \leq 4$: $k$ is 10 or 20, $p = 1$ (rationale: too easy to learn fewer than 4 locations)
  - when $m \geq 5$: $k$ is 10 or 20, $p$ is 0.99 or 0.999 or 0 (no guarantee)
- Execution time: couple of hours in all cases (dominated by $p = 1$)
Performance evaluation

Privacy GOAL 1:

- if $1 \leq m \leq 4$: $20^m$-anonymity with prob. 1
- if $m = 5$, $20^m$-anonymity with prob. $p$
- if $m \geq 5$, $p = 0$ (no guarantee)

Original:
Performance evaluation

Privacy GOAL 2:

• if $1 \leq m \leq 4$: $20^m$-anonymity with prob. 1
• if $5 \leq m \leq 11$, $20^m$-anonymity with prob. $p$

Original:
Average partition size

- Average territory of the aggregated cells

\[ p = .99 \]

\[ p = .999 \]
Conclusions

- $k^m$-anonymity is guaranteed with certain confidence
  - Adversarial knowledge is limited to any $m$ items
  - Probabilistic relaxation improves scalability and utility

- Proposed anonymization to achieve this guarantee
  - Running time is polynomial in $m$, dataset size, and universe size

- Is it enough? If so, how to choose $k$, $m$, $p$?
  - Perform Privacy Risk Analysis
Thank You!

Q (&A)
MCMC for sampling m-itemsets

Start with any existing m-items in the dataset.

REPEAT

1. PROPOSAL:
   1.1 sample a user uniformly at random
   1.2 select m items C from this user also uniformly at random

2. PROBABILISTIC ACCEPTANCE:
   2.1 accept it (i.e., S=C) with a probability, which is
       \[ \min(1, \frac{\text{Pr["S is proposed"]}}{\text{Pr["C is proposed"]}}) \]

UNTIL Convergence
personal data is any information relating to an identified or identifiable individual

- can be used to identify him or her, and to know his/her habits

- account must be taken of all the means available [...] to determine whether a person is identifiable

- any processing of any personal data must be (1) transparent (to the individual), (2) for specified explicit purpose(s), (3) relevant and not excessive in relation to these purposes

- Legally nonbinding: all member states have enacted their own data protection legislation

- Anonymized data is considered to be non-personal data, and as such, the directive does not apply to that