PRETSA: Event Log Sanitization for Privacy-aware Process Discovery

Stephan A. Fahrenkrog-Petersen, Han van der Aa & Matthias Weidlich
Motivation

Global Process Analytics Market by Organization Size 2018-2024

California Consumer Privacy Act of 2018

General Data Protection Regulation (GDPR)
Related Work

[Sweeney et al., 2002]
[Monreale et al., 2014]

Data Contribution → Information System → Data Extraction → Event Data → Sanitized Event Data → Process Mining → Process Mining Artifact

Privatized Process Mining → Process Mining Artifact

[Mannhardt et al., 2019]
Research Problem

- Use Case: Process Discovery with performance data
- Privacy Issue: Surveillance of individual process workers —> Illegal e.g. in Germany
- Preserve as much utility as possible
Attack Model

• Trace Linkage Attack
  • Link trace with background knowledge
• Identity Disclosure
• Membership Disclosure
• Attribute Disclosure
Background: k-anonymity

(a) Sensitive table

<table>
<thead>
<tr>
<th>ID</th>
<th>QI\textsubscript{1}</th>
<th>QI\textsubscript{2}</th>
<th>S\textsubscript{1}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>15</td>
<td>Flu</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>25</td>
<td>Fever</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>28</td>
<td>Diarrhea</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>15</td>
<td>Fever</td>
</tr>
<tr>
<td>5</td>
<td>22</td>
<td>28</td>
<td>Flu</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>35</td>
<td>Fever</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>32</td>
<td>Flu</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>25</td>
<td>Diarrhea</td>
</tr>
</tbody>
</table>

(b) 2-anonymous Table

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Zip</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0-20</td>
<td>10-30</td>
<td>Flu</td>
</tr>
<tr>
<td>2</td>
<td>0-20</td>
<td>10-30</td>
<td>Fever</td>
</tr>
<tr>
<td>3</td>
<td>20-30</td>
<td>10-30</td>
<td>Diarrhea</td>
</tr>
<tr>
<td>4</td>
<td>20-30</td>
<td>10-30</td>
<td>Fever</td>
</tr>
<tr>
<td>5</td>
<td>20-30</td>
<td>10-30</td>
<td>Flu</td>
</tr>
<tr>
<td>6</td>
<td>30-40</td>
<td>20-40</td>
<td>Fever</td>
</tr>
<tr>
<td>7</td>
<td>30-40</td>
<td>20-40</td>
<td>Flu</td>
</tr>
<tr>
<td>8</td>
<td>30-40</td>
<td>20-40</td>
<td>Diarrhea</td>
</tr>
</tbody>
</table>
Background: t-closeness

- Extension of k-anonymity
- Limiting difference in global and local distribution
- Earth Mover’s Distance as measure
PRETSA: PREfix-Tree based event log SAnitization

Event Log → Event Log with k-anonymity & t-closeness → Process Model with Performance Data

hu-berlin.de/pda
### PRETSA - Walkthrough

<table>
<thead>
<tr>
<th>Sequence variant</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$ create_po,update_po,receive_gd,check_in,pay_in</td>
<td>10</td>
</tr>
<tr>
<td>$\sigma_2$ create_po,update_po,receive_gd,check_in,reject_in</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma_3$ create_po,receive_gd,update_po,check_in,pay_in</td>
<td>7</td>
</tr>
<tr>
<td>$\sigma_4$ create_po,receive_gd,update_po,check_in,reject_in</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma_5$ create_po,receive_gd,update_po,update_po,check_in,pay_in</td>
<td>1</td>
</tr>
</tbody>
</table>

- Example with an Order-to-Cash process
- Assume $k=8$
PRETSA - Prefix tree

- PRETSA generates a prefix tree from an event log
- Each node in the tree is an equivalence class
PRETSA - Walkthrough

- Go through the tree until violation is found
PRETSA - Walkthrough

- PRETSA deleted the branch with violation
- Move the traces into most similar branch
PRETSA - Result

- Resulting tree
Evaluation Setup

- Utility benefit?
- PRETSA vs. Baseline
- Datasets: Traffic fines, Sepsis & CoSeLog
Experimental Setup

• Compare…
  • …generated event logs —> Nr. Variants
  • …fitness/precision of process models
  • …performance annotations relative error
### Utility Evaluation - Baseline

<table>
<thead>
<tr>
<th>Sequence variant</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$ create_po,update_po,receive_gd,check_in,pay_in</td>
<td>10</td>
</tr>
<tr>
<td>$\sigma_2$ create_po,update_po,receive_gd,check_in,reject_in</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma_3$ create_po,receive_gd,update_po,check_in,pay_in</td>
<td>7</td>
</tr>
<tr>
<td>$\sigma_4$ create_po,receive_gd,update_po,check_in,reject_in</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma_5$ create_po,receive_gd,update_po,update_po,check_in,pay_in</td>
<td>1</td>
</tr>
</tbody>
</table>

- Only release variants that fulfill:
  - k-anonymity
  - t-closeness
- Delete all other variants
Evaluation - Event Logs

### Traffic fines

![Traffic fines graph]

### CoSeLog

![CoSeLog graph]

### Sepsis

![Sepsis graph]

Legend:
- **PRETSA**
- **Baseline**
Evaluation - Process Models

Sepsis fines (fitness)

Sepsis fines (precision)

- PRETSA
- Baseline
- Non-sanitized
Evaluation - Performance Annotations
PRETSA...

...ensures privacy (k-anonymity & t-closeness) for event logs

...uses a prefix tree representation of the event log

...provides event logs with high utility for process discovery

...is available on GitHub under MIT license: 
github.com/samadeusfp/PRETSA

Questions? Reach out to fahrenks@hu-berlin.de