Diffix: Enabling (Aggregate) Data Markets with Anonymization

Paul Francis

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

MyData 2017 Session:
Roadmap for Personal Data Markets
Markets and Scale

• Market: “a regular gathering of people for the purchase and sale of provisions, livestock, and other commodities.”

• Markets don’t scale without “middlemen” of various sorts

• A “middleman” is antithetical to the MyData concept of individual control over data

• To scale a data market, individual needs to give broad consent to middleman
  • Individual cannot be expected to give consent every transaction, or even track every transaction
Data Market Without Middleman

"Sellers" → Consent → "Buyers"
Data Market Without Middleman

“Sellers”

Consent

“Buyers”
Middleman, but per-transaction consent:
- Still doesn’t scale for seller
MyData Enables and Hinders Data Markets

- **MyData enables** data markets in many ways:
  - Diverse types of data joined together through identity management
  - Diverse data leads to much better analytics
    - Knowledge of diet *and* exercise together more valuable than knowledge of diet or exercise separately
- But MyData concept of individual control of data **hinders** data markets
Anonymization is key enabler

• Most often, data “buyer” is interested in aggregate data, not individual data
• If aggregate data is anonymous, then consent is much simpler:

  Seller  
  “I give consent to sell my data anonymously”  
  Middleman

• Because anonymous data is not “personal” data
• Safe to distribute
Conclusions from GDPR-track session in Tallinn

• People don’t trust claims of anonymity
  • Too many failures in the past

• People don’t even trust DPA “certifications” of anonymity
  • Too much expertise required, too complex
Synopsis

• Buyers in a data market may be interested in only *aggregate data*
• Historically GDPR-level anonymity very hard to achieve
• *Diffix* is a technical breakthrough in anonymity
  • GDPR-level anonymity
  • Minimal data distortion
  • Simple configuration
  • Rich query semantics
• Need transparency to build confidence
Strong anonymization (has been) very hard

• Requires substantial expertise
  • For example, ARX:
    • classify data as identifying, quasi-identifying, sensitive, and insensitive;
    • create masking-based, interval-based, or order-based generalization hierarchies;
    • understand and configure privacy models such as δ-presence, l-diversity, t-closeness, δ-disclosure, k-Anonymity, k-Map, (ε,δ)-differential privacy;
    • risk-based privacy models for prosecutor, journalist and marketer risks

• Re-think for each new use case

• *Often not possible*, even for a single data set with single use case!

• *Definitely not possible* if goal is to join diverse data in a market
Diffix: Breakthrough in anonymity

Deployed as a “box” in front of an unmodified (raw) database
# Payments Table

<table>
<thead>
<tr>
<th>PID</th>
<th>Amount</th>
<th>Date</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$123.78</td>
<td>12.2.02</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>$1229.46</td>
<td>13.4.06</td>
<td>...</td>
</tr>
</tbody>
</table>

# Patient Info Table

<table>
<thead>
<tr>
<th>Patient ID</th>
<th>Name</th>
<th>Address</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bob</td>
<td>2 Elm..</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>14 Pr.</td>
<td>...</td>
</tr>
</tbody>
</table>

# Medical History

<table>
<thead>
<tr>
<th>PID</th>
<th>Diag.</th>
<th>Treatment</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Flu</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2</td>
<td>Cancer</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Diffix Configuration is Simple

- No change to database
- Simply identify and configure key fields that identify users and link user tables
- Example: 30 minutes to configure clinical database with 120 distinct user tables
Much of SQL
```sql
SELECT left(birth_number, 2) AS birth_year, count(*), count_noise(*)
FROM accounts
GROUP BY 1
```

<table>
<thead>
<tr>
<th>birth_year</th>
<th>count</th>
<th>count_noise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>22</td>
<td>1.4</td>
</tr>
<tr>
<td>38</td>
<td>73</td>
<td>1.4</td>
</tr>
<tr>
<td>60</td>
<td>80</td>
<td>1.4</td>
</tr>
<tr>
<td>19</td>
<td>48</td>
<td>1.4</td>
</tr>
<tr>
<td>80</td>
<td>93</td>
<td>1.4</td>
</tr>
<tr>
<td>49</td>
<td>85</td>
<td>1.4</td>
</tr>
<tr>
<td>20</td>
<td>57</td>
<td>1.4</td>
</tr>
<tr>
<td>72</td>
<td>86</td>
<td>1.4</td>
</tr>
</tbody>
</table>
### SELECT operation, count(*), count_noise(*)
FROM transactions
GROUP BY 1

<table>
<thead>
<tr>
<th>operation</th>
<th>count</th>
<th>count_noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREVOD NA UCET</td>
<td>254653</td>
<td>190</td>
</tr>
<tr>
<td>PREVOD Z UCTU</td>
<td>81705</td>
<td>58</td>
</tr>
<tr>
<td>VKLAD</td>
<td>162005</td>
<td>120</td>
</tr>
<tr>
<td>VYBER</td>
<td>517162</td>
<td>210</td>
</tr>
<tr>
<td>VYBER KARTOU</td>
<td>9275</td>
<td>34</td>
</tr>
</tbody>
</table>

Download as CSV | Show chart
Diffix anonymization mechanisms

• Adds noise to answers
• Uses “sticky noise”
  • Prevents averaging attacks
  • No “budget” as with differential privacy
• Adds sticky noise for each data filter
• Filters answers with too few users
  • Noisy threshold
• Removes “low-effect” data filters
SELECT salary, count(*)
FROM table
WHERE gender != 'F' AND dept = 'CS'
GROUP BY salary
SELECT salary, count(*)
FROM table
WHERE gender != 'F' AND dept = 'CS'
GROUP BY salary

Remove condition with minimal effect
SELECT salary, count(*)
FROM table
WHERE gender != 'F' AND dept = 'CS'
GROUP BY salary

Generate sticky noise layers from remaining conditions
Building confidence

• Diffix is complex, not formally proven
• Positive evaluation for anonymity from CNIL
  • French national data protection authority
  • But this cannot be considered definitive
• Plan “bug bounty” type evaluation
  • Cash prizes for breaking anonymity
  • First anonymity bug bounty
  • Release in September
Thanks!

francis@mpi-sws.org

solutions@aircloak.com