WHY ARE ORGANISATIONS SLOW TO ADOPT PETS?

Differential Privacy as a case study

Theresa Stadler, SuRI at EPFL 2018
“What are you talking about? Everybody wants data privacy as fast as possible!”

More and more organisations show their commitment to protecting user privacy by adopting privacy-enhancing technologies. Google, ...
“What are you talking about? Everybody wants data privacy as fast as possible!”

Why the Census Bureau Adopted Differential Privacy for the 2020 Census of Population

WATCH Series - John Abowd - US Census - June 6 - Noon - Room 2030

June 6, 2018 12:00 PM to
June 6, 2018 1:00 PM
NSF Room 2030

More and more organisations show their commitment to protecting user privacy by adopting privacy-enhancing technologies. Google, US Census Bureau, ...
“What are you talking about? Everybody wants data privacy as fast as possible!”

More and more organisations show their commitment to protecting user privacy by adopting privacy-enhancing technologies. Google, US Census Bureau, Apple, ...
"What are you talking about? Everybody wants data privacy as fast as possible!"

But some are struggling to get it right
"What are you talking about? Everybody wants data privacy as fast as possible!"

But some are struggling to get it right... And this is just the ones who tried to use their data. Many other organisations would like to use their data (for good) but do not know what they can do, should do or which technologies are the right ones to use. Instead they are either locking down their data or rely on laborious manual access controls and human monitoring which is slowing down innovation.
“What are you talking about? Everybody wants data privacy as fast as possible!”

Despite the current push for stronger privacy regulations and an increased awareness amongst customers for data privacy, many organisations are slower to adopt PETs than one would expect given the current push for stronger privacy regulations. Why?
“What are the hard questions that need solving for PETs to become easier to adopt?”
“But academia already offers solutions such as Differential Privacy.”

**Industry need**
Safely release aggregate statistics

**Privacy-enhancing technology**
Differential Privacy

For all $D$ and for all $S$, $A$ is $\epsilon$-differentially private if:

$$\frac{\Pr[A(D) \in S]}{\Pr[A(D') \in S]} \leq e^\epsilon$$

$\epsilon$: the privacy budget.
$A$: a probabilistic (aka noisy) query answering system.
$S$: a subset of the possible results of $A$.
$D$: the dataset.
$D'$: the dataset minus any one record.

---

Source: Founders4Schools, LinkedIn Salary, SFR28/2017
Source: Dwork and Roth, 2014
DIFFERENTIAL PRIVACY

How to protect against privacy risks in aggregate statistics

• Enable to bound the information leakage about individuals
• Allows inference about groups

<table>
<thead>
<tr>
<th>Vote</th>
<th>COUNT C</th>
<th>NOISY COUNT C + L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50% 73%</td>
<td>50% 52.5%</td>
</tr>
<tr>
<td>Remain</td>
<td>961</td>
<td>962 896</td>
</tr>
<tr>
<td>Leave</td>
<td>446</td>
<td>440 447</td>
</tr>
<tr>
<td></td>
<td>50% 50.2%</td>
<td>50% 50.2%</td>
</tr>
</tbody>
</table>

| Remain | Leave  | 2107 | 348 |

Graph showing relative probability distribution over counts added.
Industry need (for real)
Safely release aggregate statistics multiple times about several related entities where data is aggregated from a relational database with high accuracy

Privacy-enhancing technology
Differential Privacy

Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis

Source: SFR28/2017, Johnson et al. 2017

“But academia already offers solutions such as Differential Privacy.”

### Theoretical risk

**Definition 4.** A fractional linear query is specified by a vector $b \in [0, 1]^n$; the exact answer is $q_b(s) = \frac{b^T s}{n}$ (which is in $[0, 1]$ as long as $s$ is binary). An answer $\hat{q}_b$ is $\alpha$-accurate if $|q_b - \hat{q}_b| \leq \alpha$.

If a collection of fractional linear query statistics, given by the rows of a matrix $B$, is answered to within some error $\alpha$, we get the following problem:

**Definition 5 (B-reconstruction problem).** Given a matrix $B$ and a vector $\hat{q} = \frac{1}{n}Bs + e$, where $\|e\|_\infty \leq \alpha$ and $s \in [0, 1]^n$, find $\hat{s}$ with $\text{Ham}(s, \hat{s}) \leq \frac{\alpha}{n}$. The reconstruction error is the fraction $\frac{\text{Ham}(s, \hat{s})}{n}$.

**Theorem 6 (Dinur & Nissim (2003)).** When $B \in \{0, 1\}^{2^m \times n}$ has all possible rows in $\{0, 1\}^n$, there is an attack $A$ that solves the B-reconstruction problem with reconstruction error at most $4\alpha$ (given $\alpha$-accurate query answers), for every $\alpha > 0$. In particular, every mechanism that releases such statistics is blatantly nonprivate when $\alpha < 1/40$.

**Theorem 8.** There exists an attack $A$ such that, if $B$ is chosen uniformly at random in $\{0, 1\}^{2^m \times n}$ and $1 \leq m \leq 2^n$ then, with high probability over the choice of $B$, $A(B, \hat{q})$, given any $\alpha$-accurate answers $\hat{q}$, solves B-reconstruction with error $\beta = o(1)$ as long as $\alpha = \alpha\left(\sqrt{\frac{\text{Ham}(s, \hat{s})}{n}}\right)$. In particular, there is a $c > 0$ such that every mechanism for answering the queries in $B$ with error $\alpha \leq c\sqrt{\frac{\text{Ham}(s, \hat{s})}{n}}$ is blatantly nonprivate.

Source: Dwork et al. 2016

---

**What organisations are worried about**

- **Government offers school pupil data to private companies**
- **When 2+2 Equals a Privacy Question**
- **New York City makes a hash of taxi driver data disclosure**
- **The New York Times**
  - **Strava Fitness App Can Reveal Military Sites, Analysts Say**
  - **AOL Proudly Releases Massive Amounts of Private Data**
“If these problems were all solved, will PETs become a plug-and-play technique?”

Privacy expert: We offer you strong privacy protection for your data product.

Privacy expert: $\epsilon = 0.5$

Client: Great. What’s the level of privacy?

Client: ...?

Client: But does it preserve my data utility?

Client: ...?
“What are we required to do?”

- Understanding regulations is hard for businesses
- Unclear what legal terms translate into
- Even privacy expert community can’t provide answers

Source: European Union Article 29 Data Protection Working Party Opinion on Anonymization
“What could we do?”

- There’s no good overview what technologies are out there
- There’s no clear overview which PETs are fit for which use case
“What should we do?”

<table>
<thead>
<tr>
<th>Questions</th>
<th>Singling out</th>
<th>Linkage</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential Privacy</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Aggregation</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Hashing</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Suppression</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Decentralisation</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

- There’s no good overview what technologies are out there
- No clear mapping from privacy harm to techniques to reduce the risk of a harm
- No best practice examples
- Few guidelines

Source: European Union Article 29 Data Protection Working Party Opinion on Anonymization
"What do we gain?"

Theory

\[ \epsilon: \text{the privacy budget.} \]

Industry

\[ k \text{ occurrences of each value under } k\text{-anonymity} \]

Recommended 192-bit Elliptic Curve Domain Parameters over \( \mathbb{F}_p \)

- What is the transactional value of privacy?
- Businesses want to measure the value of privacy in 
- This requires clearer measures of risk and risk reduction

Source: European Union Article 29 Data Protection Working Party Opinion on Anonymization
“What do we lose?”

Experiments

- Businesses want an easy answer to the question: “Will this impact my analytics results too much?”
- Need a clearer and use case specific way of measuring data utility

PART II

“Can you give an example of a business facing these challenges?”
“What is the data we have.”

- Data: mobile phone location traces and customers’ demographic data such as age, gender, home location
- Value of the data: movement patterns of different demographic groups

<table>
<thead>
<tr>
<th>IMSI</th>
<th>Location</th>
<th>Datetime</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>Regents Park</td>
<td>Mon, 09:37am</td>
</tr>
<tr>
<td>0001</td>
<td>UCL</td>
<td>Mon, 09:51am</td>
</tr>
<tr>
<td>0002</td>
<td>Hyde Park</td>
<td>Mon, 09:31am</td>
</tr>
<tr>
<td>0001</td>
<td>Waterloo</td>
<td>Mon, 11:06am</td>
</tr>
<tr>
<td>0002</td>
<td>UCL</td>
<td>Mon, 09:46am</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IMSI</th>
<th>Age</th>
<th>Gender</th>
<th>Home location</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>45-50</td>
<td>Female</td>
<td>Battersea</td>
</tr>
<tr>
<td>0002</td>
<td>30-35</td>
<td>Male</td>
<td>Peckham</td>
</tr>
<tr>
<td>0003</td>
<td>20-25</td>
<td>Male</td>
<td>Bethnal Green</td>
</tr>
<tr>
<td>0004</td>
<td>30-35</td>
<td>Female</td>
<td>Peckham</td>
</tr>
<tr>
<td>0005</td>
<td>40-45</td>
<td>Female</td>
<td>Islington</td>
</tr>
</tbody>
</table>
“What we are worried about.”

**UNICITY:** \( \mathcal{E}_p \)

Quantifies the average risk of re-identification of a dataset knowing \( p \) points

Not a privacy guarantee but a risk measure

- DeMontjoye et al. 2013 studied mobile phone traces of 1.5M users in a European country over 15 months
- Showed that knowing 4 spatiotemporal points is enough to uniquely identify the location trace of 95% of the individuals

Credit: Yves-Alexandre de Montjoye, Source: Montjoye et al. 2013
“What we plan to do.”

- Aggregate data by user defined spatial areas and time windows
- Publish aggregate statistics only such as counts of people in certain regions grouped by origin and destination
- Suppress small counts to protect individuals

getOrigin(8) at [UCL, Mon, 09:45am – 10am]

From Regents Park: 7    From Hyde Park: -
“What should we do?”

“Alice entered region R at 10:23” – Query 2

- Raw aggregates are still vulnerable to differencing attacks
“What could we do?”

- Differential Privacy to the rescue: Seems to be a good fit for noise addition
- Benefits of using Differential Privacy
  - Formal privacy guarantee
  - Quantifiable privacy loss
  - Quantifiable accuracy loss
  - Future proof

\[ C = C_{\text{raw}} + L \]

<table>
<thead>
<tr>
<th>ORIGIN</th>
<th>COUNT (C_{\text{raw}})</th>
<th>Small noise</th>
<th>Medium noise</th>
<th>Large noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hide Park</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>28</td>
</tr>
<tr>
<td>Regents Park</td>
<td>111</td>
<td>108</td>
<td>97</td>
<td>83</td>
</tr>
<tr>
<td>Battersea Park</td>
<td>608</td>
<td>605</td>
<td>594</td>
<td>580</td>
</tr>
</tbody>
</table>
“What do we gain?”

Definition 4. A fractional linear query is specified by a vector \( b \in \{0, 1\}^n \); the exact answer is \( q_b(x) = \frac{1}{b^T x} \) (which lies in \([0, 1]\) as long as \( x \) is binary). An answer \( \hat{q}_b \) is \( \alpha \)-accurate if \( |q_b - \hat{q}_b| \leq \alpha \).

If a collection of fractional linear query statistics, given by the rows of a matrix \( B \), is answered to within some error \( \epsilon \), we get the following problem:

Definition 5 (B-reconstruction problem). Given a matrix \( B \) and a vector \( \hat{q} = \frac{1}{b^T x} + \epsilon \), where \( |\epsilon| \leq \alpha \) and \( x \in \{0, 1\}^n \), find \( x \) with \( \text{Ham}(x, x') \leq \frac{\epsilon}{\alpha} \). The reconstruction error is the fraction \( \frac{\text{Ham}(x, x')}{\alpha} \).

Theorem 6 (Dinur & Nissim (2003)). When \( B \in \{0, 1\}^{2^n \times n} \) has all possible rows in \( \{0, 1\}^n \), there is an attack \( A \) that solves the B-reconstruction problem with reconstruction error at most \( 4\alpha \) (given \( \alpha \)-accurate query answers), for every \( \alpha > 0 \). In particular, every mechanism that releases such statistics is blatantly nonprivate when \( \alpha < 1/40 \).

Theorem 8. There exists an attack \( A \) such that, if \( B \) is chosen uniformly at random in \( \{0, 1\}^{m \times n} \) and \( 1.2n \leq m \leq 2^n \) then, with high probability over the choice of \( B \), \( A(B, \hat{q}) \), given any \( \alpha \)-accurate answers \( \hat{q} \), solves B-reconstruction with error \( \beta = \alpha(1) \) as long as \( \alpha = o(\sqrt{\frac{m}{2^n n}}) \). In particular, there is \( \alpha > 0 \) such that every mechanism for answering the queries in \( B \) with error \( \alpha \leq c \sqrt{n \log(2^n)} \) is blatantly nonprivate.

- Accuracy loss through noise addition needs to be justified by high risk
- The probability of these attacks happening in the real world hard to measure
- Hard to compare the protection from classical statistical disclosure control to the Differential Privacy guarantee and demonstrate the “gain in privacy”

Source: https://teachprivacy.com/the-funniest-hacker-stock-photos-4-0/
“What do we lose?”

What about insights about smaller groups?

Will a breakdown into smaller subregions be consistent with the roll-up of the table?

<table>
<thead>
<tr>
<th>PLACE</th>
<th>TIME</th>
<th>COUNT</th>
<th>NOISY COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCL</td>
<td>08:05</td>
<td>4422</td>
<td>4431</td>
</tr>
<tr>
<td>WATERLOO</td>
<td>08:05</td>
<td>2341</td>
<td>2341</td>
</tr>
</tbody>
</table>

- Accuracy ≠ utility: need for a use case specific utility measure
- Worried whether noise will wash out the signal and whether insights will be preserved
- Worried about consistency issues of noise addition that can lead to false conclusions and confusion of the data analyst
- Question about the “operating envelope” of Differential Privacy: Minimum sample size? Maximum number of statistics?

Will the temporal trend be preserved under noise addition?
”How do we do it?”

- **Noise addition**
  - How do we tune the noise to have optimal privacy-utility trade-off?
  - What should the query-rate limit be?
  - What should the minimum query set size be?
  - How do we communicate uncertainty?

- **Generalisation**
  - What should the minimum temporal aggregation window be?
  - What should the minimum spatial aggregation area be?

- **Monitoring**
  - 
  - 
  - 

- How to evaluate the privacy-utility trade-off?
- How to set all implementation parameters?
PART III

“So how do we accelerate the adoption of PETs?”
“What do we need to work on?”

Improved PUT Gaussian mechanism

Data changing over time

Local Differential Privacy for Evolving Data
Matthew Joseph* Aaron Roth† Jonathan Ullman† Bo Waggoner‡
February 21, 2018

Calculate query sensitivity from a relational database

- Tackle the specific technical challenges of PETs such as in Differential Privacy
- More work like Balle and Wang, Neel et al., Joseph et al., McSherry, Song et al.

“What do we need to work on?”

- Develop new techniques for new data use cases
- More work like McMahan et al. 2017, DP team at Apple 2018

Source: Google AI Blog, Apple Machine Learning Journal
“What do we need to work on?”

- Quantify disclosure risk
- Find the right definitions of privacy
- More work like DeMontoye et al., Papernot et al.
- More engagement with customers and business: What are businesses worried about? What do people consider as a privacy breach? What are their privacy expectations?

Source: cleverhans-blog by Goodfellow and Papernot, DeMontjoye et al. 2013, The New Yorker
“What do we need to work on?”

A Review of Statistical Disclosure Control Techniques Employed by Web-Based Data Query Systems

Gregory J. Matthews, PhD; Ofer Harel, PhD; Robert H. Aseltine Jr, PhD

- Demonstrate the practicality of attacks
- Show that theoretical attacks need to be considered as a real threat

Source: Matthews et al. 2017, teachprivacy.com
“What do we need to work on?”

- Easier to interpret utility measures
- Tailor utility measures to use case
- More collaboration with industry partners who have specific data use cases

Source: SFR37/2017, FFT Education Datalab
“What do we need to work on?”

- Principled ways of setting epsilon
- Relating privacy parameters to regulations
- More collaborations between academics, lawyers, practitioners, users

Source: Founders4Schools, SFR28/2017, SFR37/2017
“What do we need to work on?”

- Highlight the advantages of PETs rather than only talking about the drawbacks

Source: cleverhans-blog by Ian Goodfellow and Nicolas Papernot
“What do we need to work on?”

John is concerned that a potential health insurance provider will deny him coverage in the future, if it learns certain information about his health, such as his HIV-positive status, from a medical research database that health insurance providers can access via a differentially private mechanism. If the insurer bases its coverage decision with respect to John in part on information it learns via this mechanism, then its decision corresponds to an event defined over the outcome of a differentially private analysis.

- Find stories and analogies to explain the hard concepts in data privacy
- More work like Nissim et al. 2017

Source: Nissim et al. 2017
“So, what do I do now?!”

- Aim to make privacy-technologies manageable for non-experts
- Translate abstract parameters into more interpretable ones
- Find stories to explain hard concepts in data privacy
- Talk to individuals about their expectations of privacy
- Talk to lawyers and regulators to learn their language and share your expertise with them
- Watch out for the synergies between privacy and utility