What is Meant by the Privacy of a Sampled Individual?

Palash Sarkar

Applied Statistics Unit
Indian Statistical Institute
palash@isical.ac.in

As part of the farewell ceremony
of
Professor Shibdas Bandyopadhyay
(Member, National Statistics Commission)

31st October 2012.

Statistics and Privacy

Examples of **sensitive** data:

- Population or agriculture censuses.
- Income tax returns.
- Medical records.

Risk: private, personal or corporate information can be disclosed.

Adverse effect: Loss of trust from citizens and businesses.

Problem: How to infer sufficiently detailed and useful statistics but, still protect the privacy of the sampled individuals?

Fundamental Principles of Official Statistics

Ten principles formulated by the United Nations:

`http://unstats.un.org/unsd/methods/statorg/FP-English.htm`

Fundamental Principles of Official Statistics

Ten principles formulated by the United Nations:

`http://unstats.un.org/unsd/methods/statorg/FP-English.htm`

The principle pertaining to privacy.

> *"Principle 6. Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes."*

What is Privacy?

Two separate strands of work:

- In the statistical literature as “statistical disclosure limitation” or “data confidentiality”.
  - Many articles in the Journal of Official Statistics.
- In the database community as “privacy-preserving data mining”.
  - Interaction between the theory of databases, complexity theory, cryptography and game theory.

What is Privacy?

Two separate strands of work:

- In the statistical literature as “statistical disclosure limitation” or “data confidentiality”.
  - Many articles in the Journal of Official Statistics.
- In the database community as “privacy-preserving data mining”.
  - Interaction between the theory of databases, complexity theory, cryptography and game theory.

Guarantee of privacy?

- Either, no formal guarantee of privacy;
- or, ensure privacy only against a specific suite of attacks.

What is Privacy?

Two separate strands of work:

- In the statistical literature as “statistical disclosure limitation” or “data confidentiality”.
  - Many articles in the Journal of Official Statistics.
- In the database community as “privacy-preserving data mining”.
  - Interaction between the theory of databases, complexity theory, cryptography and game theory.

Guarantee of privacy?

- Either, no formal guarantee of privacy;
- or, ensure privacy only against a specific suite of attacks.

The issue of *defining* privacy not adequately addressed.

How to Define Privacy?

Dalenius' criterion (1977):

> *Anything that can be learnt about a respondent from a sample should be learnable without access to the sample.*

How to Define Privacy?

Dalenius' criterion (1977):

> *Anything that can be learnt about a respondent from a sample should be learnable without access to the sample.*

Problems:

- Statistics does change knowledge about individuals.
  - Example: Suppose 'Smoking Causes Cancer' is inferred; this *does* change the view about a respondent who is a smoker.

How to Define Privacy?

Dalenius' criterion (1977):

> *Anything that can be learnt about a respondent from a sample should be learnable without access to the sample.*

Problems:

- Statistics does change knowledge about individuals.
  - Example: Suppose 'Smoking Causes Cancer' is inferred; this *does* change the view about a respondent who is a smoker.
- Combination with auxilliary information.
  - Example: Suppose auxilliary information 'Madhu is 3 inches taller than an average Bengali' is available; combined with average height of Bengalis inferred from a sample gives a good idea of the height of Madhu.
  - The crux is that Madhu need *not* even be in the sample.

Differential Privacy

Introduced by Dwork et al (TCC, 2006) and Dwork (ICALP, 2006).
The rough idea is the following:

> *Whatever can be inferred from a sample with an individual present can also be essentially inferred from the sample without the individual.*

- Differential in the sense of two samples differing in the presence of exactly one respondent.
- The presence or absence of an individual does not make much difference to the inferred value.
- Requires the inference algorithm to be probabilistic.

Differential Privacy

- A query $\mathcal{Q}$ is a randomised algorithm which is applied to a sample to infer some quantity about the underlying population.
- Consider two samples $\mathcal{D}$ and $\mathcal{D}'$ which differ in exactly one respondent.
- For any subset of $S \subset \text{Range}(\mathcal{Q})$, let $\Pr[\mathcal{Q}(D) \in S]$ be the probability that the output of the query is in the set S.
- The idea of differential privacy is to quantify that $\Pr[\mathcal{Q}(D) \in S]$ and $\Pr[\mathcal{Q}(D') \in S]$ are not much different.

Differential Privacy

- A query $\mathcal{Q}$ is a randomised algorithm which is applied to a sample to infer some quantity about the underlying population.
- Consider two samples $\mathcal{D}$ and $\mathcal{D}'$ which differ in exactly one respondent.
- For any subset of $S \subset \text{Range}(\mathcal{Q})$, let $\Pr[\mathcal{Q}(D) \in S]$ be the probability that the output of the query is in the set S.
- The idea of differential privacy is to quantify that $\Pr[\mathcal{Q}(D) \in S]$ and $\Pr[\mathcal{Q}(D') \in S]$ are not much different.

ϵ-differentially private:

$$\Pr[\mathcal{Q}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{Q}(D') \in S].$$

The probability is over the internal random bits of $\mathcal{Q}$.

Achieving Differential Privacy

Suppose f be a deterministic inference function.

- Input perturbation: add random noise to the sampled values and apply f.
  - Becomes difficult to apply for complex queries.
- Output perturbation: apply f and then add random noise to the output.
  - Does not work directly, since, the same query can be repeated and the results averaged to cancel out the noise.
  - Storing previous queries is impractical; further, deciding semantic equivalence of queries is undecidable.

Achieving DP: A Simple Idea

Suppose the range of f is a finite subset $\mathcal{S}$ of $\mathbb{R}$. Define the sensitivity of f as follows: $\Delta f = \max_{D,D'} \|f(D) - f(D')\|_1$ for all samples D and D' differing in only one individual.

Consider the distribution $\mathrm{Lap}(\lambda)$ with pdf $p_\lambda(y) = \frac{1}{2\lambda} \exp\left(\frac{-|y|}{\lambda}\right)$. For $r \in \mathcal{S}$, the probability mass assigned to r is $q_\lambda(r) = p_\lambda(r)/\theta$, where $\theta = \sum_{y \in \mathcal{S}} p_\lambda(y)$ and $\lambda = \Delta f/\varepsilon$.

Query mechanism $\mathcal{Q}$: Compute $f(D)$ and return $f(D) + r$ where r is a noise following $q_\lambda(r)$.

This achieves ε-DP.

Can be extended to a sequence of queries $f_1, f_2, \ldots, f_\ell$.

DP and Statistics

DP inherently introduces noise. So, the following question is important.

Is it possible to form reasonably accurate statistical estimates while maintaining differential privacy?

DP and Statistics

DP inherently introduces noise. So, the following question is important.

Is it possible to form reasonably accurate statistical estimates while maintaining differential privacy?

- Differentially private estimate: For a large class of parametric probability models, Smith (2008) has shown a DP estimator whose distribution converges to that of the MLE.

DP and Statistics

DP inherently introduces noise. So, the following question is important.

Is it possible to form reasonably accurate statistical estimates while maintaining differential privacy?

- Differentially private estimate: For a large class of parametric probability models, Smith (2008) has shown a DP estimator whose distribution converges to that of the MLE.
- Robust statistics: The goal is to ensure that a single data item does not have a significant effect on the outcome.
  - Similar in spirit to the idea of DP.
  - This similarity has led to robust statistical estimators being used to construct DP estimators for interquartile distance, the median and regression (Dwork-Lei, 2008).

Some References and Links


C. Dwork, A firm foundation for private data analysis, Communications of the ACM.


C. Dwork, A. Smith, Differential privacy for statistics: What we know and what we want to learn, Journal of Privacy and Confidentiality, 1.


A. Smith, Efficient, differentially private point estimators, CoRR abs/0809.4794.


C. Dwork, J. Lei, Differential privacy and robust statistics, in: Proceedings of the 41th Annual ACM Symposium on Theory of Computing (STOC), 2008.

Links:

`http://research.microsoft.com/en-us/projects/databaseprivacy/`

`http://en.wikipedia.org/wiki/Differential_privacy`

Thank you for your attention!