Privacy in databases

Gijs Kopmeiners en Thomas van Harskamp

Outline

- **Statistical Disclosure Control**
    - What is Statistical Disclosure Control?
    - Why/when is Statistical Disclosure Control used?
    - Risk assessment
    - SDC techniques

- **Differential privacy**
    - Introduction
    - Definition of differential privacy
    - ε-differential privacy
    - Composability of ε-differential privacy
    - Group privacy
    - Laplace mechanism
    - Randomised response
    - Usages

Definitions

- **Data**
  - Microdata/Macrodata
  - Aggregate/tabular data
  - Surveys/Frequency tables/etc.
- **Data intruder**
  - The “Attacker” / Adversary
  - Something or someone seeking to identify population units within a dataset
  - Legal/Illegal access
- **Key variable**
  - Information known to a data intruder about a population unit which is also present on an anonymised dataset
  - Indirect identifier that could be used to re-identify individuals

Statistical Disclosure Control

What is Statistical Disclosure Control (SDC)?

Definition: Statistical Disclosure Control is the practice of reducing the risk of finding people or entities in data (re-identification) and/or associating data with a person or entity (association).

Different types of disclosure

- Re-identification: the association of a particular record within a set of data with particular population unit
- Association: the association or disassociation of a particular attribute with a particular population unit
- Often together, but not always the case
- Important to balance between data utility and confidentiality/privacy
- SDC aims to disclose data in a way such that no information is leaked
- Examples?

Re-identification

- **The association of a particular record within a set of data with particular population unit**
- **Direct identifiers:**
  - Names, BSN, etc.
- **Indirect identifiers:**
  - What do you think?
- **Combining key values with information you know**

Re-identification example

- **What can the data intruder learn?**
- **What is/are the identifying/key variable(s)?**
- **How can you handle this problem without impacting usefulness of data?**

ID	Net worth in $	Hair color	American shoe size
45	186.900.000.000	Brown	13
46	243.000	Blonde	12
47	110.000	Brown	13

Attribute disclosure

- **The association or disassociation of a particular attribute with a particular population unit**
- **Can happen together with or without re-identification**
  - With: what is the shoe size of elon musk?
  - Without: how do you think?

Attribute disclosure example

- **What can be learned from this dataset?**
- **Do you need to identify units in the dataset?**

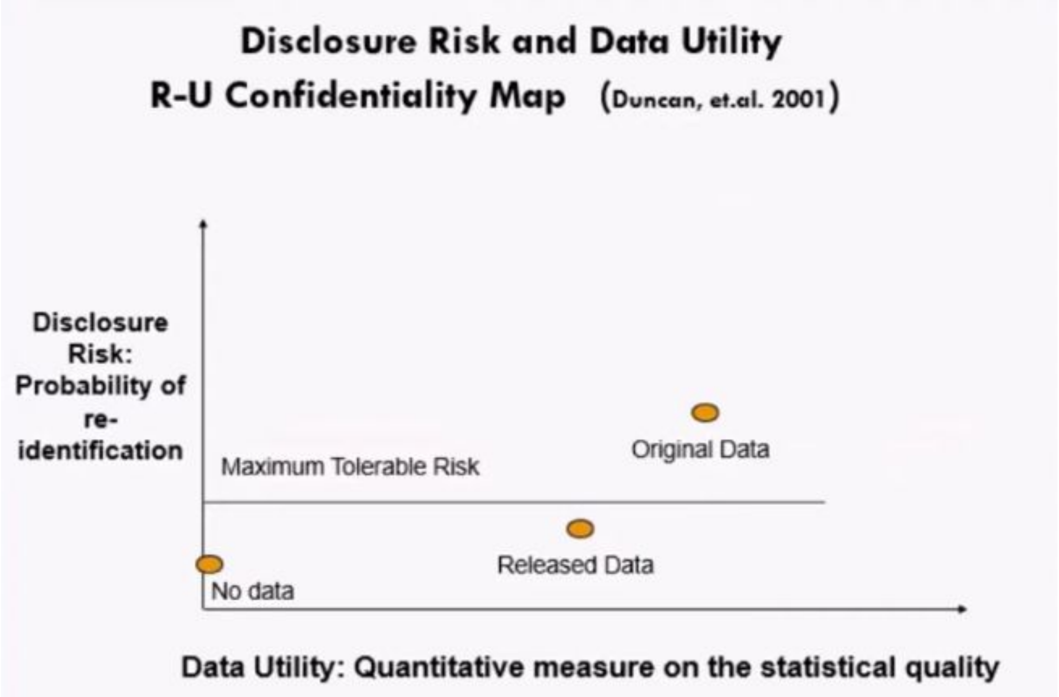
Age	Sex	% of people with cancer	% of people eating meat
50	F	27	30
50	M	29	40
51	F	100	41
51	M	40	23
52	F	56	26

Who uses SDC and why?

- **Researchers**
  - Research uses potentially sensitive data
  - Not interested in information about one single person
  - Balance between protecting confidentiality and data utility
- **Organizations have their own policies**
- **CBS**
  - Regulation
  - GDPR

Actual and perceived risk

- **Perceived risk is a complex psychosocial process**
- **Assessing and controlling disclosure risk are highly complex**
  - What is the actual risk?
  - What knowledge does data intruder have?
- **Key selection**
  - Key combinations
  - Exponantial growth
- **Data utility**



Risk assessment

- **Population uniqueness**
  - Proportion of population units that are unique on a given key
  - Assume the intruder information is 100% compatible
  - Unique =/= risky
- **Risky records**
  - Rare
  - Target for SDC
- **Matching**
  - Match records of two files
  - Very specific

Risk factors

- **Data divergence**
  - Inconsistencies
  - Reduces the probability of correct attribution or identification
- **Size of key**
  - Larger keys more options

SDC-in SDC-out

- **SDC-in aka pre-tabulation disclosure control: Doing something to the data before it is put in the tables**
  - Anonymization
- **SDC-out aka aka post-tabulation disclosure control: Doing something to the tables that contain the data**
  - Random noise
  - (Differential privacy)
  - Etc
- **Active research areas**

Data anonymization

- Process of protecting sensitive information by removing identifiers
- Enforced by GDPR
- Data masking: hiding data with altered values
- Pseudonymization: replacing identifiers with pseudonyms
- Limits ability to derive value and insight from data

Recoding

- **Collapse categories of a variable**
- **Low frequencies conjoined**
- **Topping**
  - Elon Musk example
- **Grouping**
  - Age, occupation
- **Visible**
- **Small benefit**
- **High cost**

ID	Net worth in $	Hair color	American shoe size
45	250.000+	Brown	13
46	200.000-250.000	Blonde	12
47	100.000-150.000	Brown	13

Cell suppression

- Leaving cell blank
- Sensitive
- Low count
- Suppress more cells
- Complementary suppressions
- Can quickly reduce analytical value

Rounding

- Disguise exact frequency
- Random rounding
- Cell counts may not add up

Masking/Blurring

- Add noise
- Add/subtract numbers to cells
- Changing values to other values
- Overwrite sensitive values

Data swapping

- Swap variables
- Data intruder doesn't know which ones
- Random swapping
- Record swapping

Impact on data

- Mathematical models
- Information loss
- Dependant on what user wants

Alternatives/additions

?

Alternatives/additions

- **Safe settings**
  - Data holders good view on what is happening
  - Hard to set up
  - No ease of access
- **Only allow queries**
- **Simulated data**
  - Mutation algorithms
  - Analytically equivalent?

Differential Privacy

Collecting analytical data

- Suppose we want to collect analytical data on a database
- We also want to preserve the privacy of individuals
- Can we just disclose statistics of the database?

Example

Sally:

- **Has a rare illness**
- **Is in a database of the local abortion clinic**

Problems?

Example

How do we solve this problem?

What is differential privacy?

?

In general

- **Collecting analytical data while preserving the privacy of individuals**

In general

- **Collecting analytical data while preserving the privacy of individuals**
- **Adds noise in a controlled way**

In general

- **Collecting analytical data while preserving the privacy of individuals**
- **Adds noise in a controlled way**
- **Introduced by Cynthia Dwork and Frank McSherry, et al. in 2006**

In general

- **Collecting analytical data while preserving the privacy of individuals**
- **Adds noise in a controlled way**
- **Introduced by Cynthia Dwork and Frank McSherry, et al. in 2006**
- **Individual record should not significantly outcome of analytical function**

In general

- **Collecting analytical data while preserving the privacy of individuals**
- **Adds noise in a controlled way**
- **Introduced by Cynthia Dwork and Frank McSherry, et al. in 2006**
- **Individual record should not significantly outcome of analytical function**
- **Standard for measuring privacy in data analysis**

Usefulness

Usefulness

- **Does not matter what the adversary knows or does**

Usefulness

- **Does not matter what the adversary knows or does**
- **Even when attacker has unlimited computing power**

Usefulness

- **Does not matter what the adversary knows or does**
- **Even when attacker has unlimited computing power**
- **Future-proof**

Mechanism

Mechanism

- **Computation on a database**

Mechanism

- **Computation on a database**
- **Randomised mechanism**

Mechanism

- **Computation on a database**
- **Randomised mechanism**
- **Considered differentially private if the probability of an outcome of the mechanism occurring is almost the same for all pairs of datasets that only differ in one record**

Mechanism

- Computation on a database
- Randomised mechanism
- Considered differentially private if the probability of an outcome of the mechanism occurring is almost the same for all pairs of datasets that only differ in one record
- A single person participating in the database should not make a difference

Mechanism

- **Computation on a database**
- **Randomised mechanism**
- **Considered differentially private if the probability of an outcome of the mechanism occurring is almost the same for all pairs of datasets that only differ in one record**
- **A single person participating in the database should not make a difference**
- **Next up: a mathematical definition**

Mechanism

Complete differential privacy is really hard to achieve, why is that?

In general

- The way of denoting privacy for a mechanism
- Mathematical notation
- Composability
- Group privacy

Mathematical notation

Defined as follows[1] for datasets D1 and D2 that differ in only one entry, mechanism K, S ⊆ Image(K), and $\varepsilon \geq 0$:

$$Pr[K(D_1) \in S] \leq e^{\varepsilon} \cdot Pr[K(D_2) \in S]$$

[1] Cynthia Dwork. Differential privacy. In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, *Automata, Languages and Programming*, pages 1–12, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Mathematical notation

Alternatively in literature[2]:

$$\left| ln\left(\frac{Pr[K(D_1) \in S]}{Pr[K(D_2) \in S]}\right)\right| \leq \varepsilon$$

Rewrite steps show this is the same.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Mathematical notation

- $\varepsilon = 0$ means we would have complete differential privacy
- $\varepsilon = 10$ is hardly worth anything, since $e^{10} = {\sim}22000$

Mathematical notation

- $\varepsilon = 0$ **means we would have complete differential privacy**
- $\varepsilon = 10$ **is hardly worth anything, since** $e^{10} = \sim 22000$
- **Next up: composition of mechanisms**

Composition

- **A way of measuring privacy of mechanisms together**

Composition

- **A way of measuring privacy of mechanisms together**

- **What do you think that the total privacy of the use of n privacy mechanisms used on the same database is, with respective differential privacy $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$?**

Composition

- **A way of measuring privacy of mechanisms together**

- **What do you think that the total privacy of the use of n privacy mechanisms used on the same database is, with respective differential privacy ε_1, ε_2, ..., ε_n?**

a) max(ε_i) over all 1 <= i <= n

b) $\prod_{i=1}^{n} \varepsilon_i$

c) $\sum_{i=1}^{n} \varepsilon_i$

Sequential Composition

- **Different mechanisms used on same database**

Sequential Composition

- **Different mechanisms used on same database**
- **No restrictions on mechanisms or data used for mechanisms**

Sequential Composition

- **Different mechanisms used on same database**
- **No restrictions on mechanisms or data used for mechanisms**
- **Mechanisms can be computed after each other**

Sequential Composition

- **Different mechanisms used on same database**
- **No restrictions on mechanisms or data used for mechanisms**
- **Mechanisms can be computed after each other**
- **Example: first requesting the average age in a database, and then based on this requesting the percentage of people with a certain illness**

Sequential Composition

Mathematical proof of privacy in sequential composition:

- **write exp(x) for e^x**
- **databases A and B that differ in at most one entry**
- **write $A \oplus B$ for difference between A and B,**
  $|A \oplus B|$ for number of entries different between A and B
- **sequence of mechanisms $M = \{M_i\}_{i=1}$**
- **sequence r of outcomes $r_i \in \text{Range}(M_i)$**
- **write M_i^r for mechanism M_i supplied with $r_1, \ldots, r_{i-1}$**
- **probability of output r from the sequence of $M_i^r(A)$:**

$$\Pr[M(A) = r] \quad = \quad \prod_i \Pr[M_i^r(A) = r_i]$$

Sequential Composition

Mathematical proof continued[3]:

$$\Pr[M(A) = r] = \prod_i \Pr[M_i^r(A) = r_i] \leq \prod_i \Pr[M_i^r(B) = r_i] \times \prod_i \exp(\epsilon_i \times |A \oplus B|) =$$

$$\Pr[M(B) = r] \times \exp\left(\sum_i \epsilon_i\right)$$

[3] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, page 19–30, New York, NY, USA, 2009. Association for Computing Machinery.

Sequential Composition

Mathematical proof continued[3]:

$$\Pr[M(A) = r] = \prod_i \Pr[M_i^r(A) = r_i] \leq \prod_i \Pr[M_i^r(B) = r_i] \times \prod_i \exp(\epsilon_i \times |A \oplus B|) =$$

$$\Pr[M(B) = r] \times \exp\left(\sum_i \epsilon_i\right)$$

This means that the total privacy of mechanisms used sequentially is equal to the sum of their ε-values.

[3] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, page 19–30, New York, NY, USA, 2009. Association for Computing Machinery.

Parallel Composition

?

Parallel Composition

- **Different mechanisms used on same database**

Parallel Composition

- **Different mechanisms used on same database**
- **Restriction: datasets on which mechanisms are used must be disjoint**

Parallel Composition

- **Different mechanisms used on same database**
- **Restriction: datasets on which mechanisms are used must be disjoint**
- **Mechanisms can be computed in parallel, since no mechanism gives information about the other one anyways**

Parallel Composition

- **Different mechanisms used on same database**
- **Restriction: datasets on which mechanisms are used must be disjoint**
- **Mechanisms can be computed in parallel, since no mechanism gives information about the other one anyways**
- **Example: requesting the average age of customers in a database, while also requesting the average cost of products in the database**

Parallel Composition

- Different mechanisms used on same database
- Restriction: datasets on which mechanisms are used must be disjoint
- Mechanisms can be computed in parallel, since no mechanism gives information about the other one anyways
- Example: requesting the average age of customers in a database, while also requesting the average cost of products in the database
- For the mathematical proof, we consider the mechanisms computed sequentially (should not matter given restriction)

Parallel Composition

Mathematical proof of privacy in sequential composition:

- **write exp(x) for e^x**
- **databases A and B that differ in at most one entry**
- **write $A \oplus B$ for difference between A and B, $|A \oplus B|$ for number of entries different between A and B**
- **split input domain into $D = D_1 \cup D_2 \cup \ldots$ where all D_i are disjoint**
- **write $A_i = A \cap D_i$ and $B = B \cap D_i$ to split A and B into disjoint subsets**
- **sequence of mechanisms $M = \{M_i\}_{i=1}$**
- **sequence r of outcomes $r_i \in \text{Range}(M_i)$**
- **write M_i^r for mechanism M_i supplied with $r_1, \ldots, r_{i-1}$**
- **probability of output r from the sequence of $M_i^r(A)$:**

$$\mathbf{Pr}[M(A) = r] \quad = \quad \prod_i \mathbf{Pr}[M_i^r(A_i) = r_i]$$

Parallel Composition

Mathematical proof continued[3]:

$$\mathbf{Pr}[M(A)=r] \;=\; \prod \mathbf{Pr}[M_i^r(A_i)=r_i]$$

$$\leq \prod_i \mathbf{Pr}[M_i^r(B_i)=r_i] \times \prod_i \exp(\epsilon_i \times |A_i \oplus B_i|)$$

$$\leq \prod_i \mathbf{Pr}[M_i^r(B_i)=r_i] \times \exp(\epsilon_j \times |A \oplus B|) \;\leq\; \mathbf{Pr}[M(B)=r] \times max_i\{\epsilon_i\}$$

[3] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, page 19–30, New York, NY, USA, 2009. Association for Computing Machinery.

Parallel Composition

Mathematical proof continued[3]:

$$\mathbf{Pr}[M(A)=r] \;=\; \prod \mathbf{Pr}[M_i^r(A_i)=r_i]$$

$$\leq \prod_i \mathbf{Pr}[M_i^r(B_i)=r_i] \times \prod_i \exp(\epsilon_i \times |A_i \oplus B_i|)$$

$$\leq \prod_i \mathbf{Pr}[M_i^r(B_i)=r_i] \times \exp(\epsilon_j \times |A \oplus B|) \quad \leq \quad \mathbf{Pr}[M(B)=r] \times max_i\{\epsilon_i\}$$

This means that the total privacy of mechanisms used in parallel is equal to the biggest ε-value of all mechanisms.

[3] Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, page 19–30, New York, NY, USA, 2009. Association for Computing Machinery.

Definition

Definition

- **Privacy of a group of individuals**

Definition

- Privacy of a group of individuals
- Does information about a group of c people leak?

Definition

- Privacy of a group of individuals
- Does information about a group of c people leak?
- For example: can you guess the age of a group of 3 people with high probability, when requesting the average age of everyone in the database?

Definition

- **Privacy of a group of individuals**
- **Does information about a group of c people leak?**
- **For example: can you guess the age of a group of 3 people with high probability, when requesting the average age of everyone in the database?**
- **Easily extended from definition of ε-differential privacy**

Mathematical notation

Mathematical notation

?

Mathematical notation

Just apply the definition of ε-differential privacy repeatedly.

Mathematical notation

Just apply the definition of ε-differential privacy repeatedly.

For c individuals:

$$Pr[K(D_1) \in S] \leq e^{c\varepsilon} \cdot Pr[K(D_2) \in S]$$

Mathematical notation

Just apply the definition of ε-differential privacy repeatedly.

For c individuals:

$$Pr[K(D_1) \in S] \leq e^{c\varepsilon} \cdot Pr[K(D_2) \in S]$$

So this gives c*ε-differential privacy.

Definition

Definition

- **Adds noise to analytical data in the form of the Laplace distribution**

Definition

- **Adds noise to analytical data in the form of the Laplace distribution**
- **Desired ε-differential privacy can be exactly chosen**

Definition

- **Adds noise to analytical data in the form of the Laplace distribution**
- **Desired ε-differential privacy can be exactly chosen**
- **Variations like Gaussian Mechanism exist**

Mathematical notation

We must first define the sensitivity of a function:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

Mathematical notation

We must first define the sensitivity of a function:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

This is basically the largest difference between two outcomes of a function on all possible pairs of datasets that differ in one entry.

Mathematical notation

We must first define the sensitivity of a function:

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1$$

This is basically the largest difference between two outcomes of a function on all possible pairs of datasets that differ in one entry.

An example:

When requesting the total number of women in a database, the sensitivity of the function is 1, since two databases that differ in one entry can have 1 woman more or less.

Mathematical notation

We must first define the sensitivity of a function:

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1$$

This is basically the largest difference between two outcomes of a function on all possible pairs of datasets that differ in one entry.

Using this, we can define the Laplace Mechanism for a desired differential privacy:

$$\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$$

Mathematical notation

We can now prove that this indeed gives the desired differential privacy using the following definition for ε-differential privacy:

$$\frac{Pr[K(D_1)\in S]}{Pr[K(D_2)\in S]}) \leq exp(\varepsilon)$$

Complete proof is on the next slide.

Mathematical notation

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1$$

$$\frac{Pr[K(D_1) \in S]}{Pr[K(D_2) \in S]}) \leq exp(\varepsilon)$$

$$\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$$

Proof:

$$\frac{\Pr\left(\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = r\right)}{\Pr\left(\mathcal{M}_{\text{Lap}}(y, f, \epsilon) = r\right)} = \frac{\Pr\left(f(x) + \text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r\right)}{\Pr\left(f(y) + \text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r\right)} = \frac{\Pr\left(\text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r - f(x)\right)}{\Pr\left(\text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r - f(y)\right)}$$

$$= \frac{\frac{1}{2b}\exp\left(-\frac{|r - f(x)|}{b}\right)}{\frac{1}{2b}\exp\left(-\frac{|r - f(y)|}{b}\right)} = \exp\left(\frac{|r - f(y)| - |r - f(x)|}{b}\right) \leq \exp\left(\frac{|f(y) - f(x)|}{b}\right) \leq \exp\left(\frac{\Delta f}{b}\right) = \exp(\epsilon)$$

Mathematical notation

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \|_1$$

$$\frac{Pr[K(D_1) \in S]}{Pr[K(D_2) \in S]}) \leq exp(\varepsilon)$$

$$\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = f(x) + \text{Lap}\left(\mu = 0, b = \frac{\Delta f}{\epsilon}\right)$$

Proof:

$$\frac{\Pr\left(\mathcal{M}_{\text{Lap}}(x, f, \epsilon) = r\right)}{\Pr\left(\mathcal{M}_{\text{Lap}}(y, f, \epsilon) = r\right)} = \frac{\Pr\left(f(x) + \text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r\right)}{\Pr\left(f(y) + \text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r\right)} = \frac{\Pr\left(\text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r - f(x)\right)}{\Pr\left(\text{Lap}\left(0, \frac{\Delta f}{\epsilon}\right) = r - f(y)\right)}$$

$$= \frac{\frac{1}{2b}\exp\left(-\frac{|r - f(x)|}{b}\right)}{\frac{1}{2b}\exp\left(-\frac{|r - f(y)|}{b}\right)} = \exp\left(\frac{|r - f(y)| - |r - f(x)|}{b}\right) \leq \exp\left(\frac{|f(y) - f(x)|}{b}\right) \leq \exp\left(\frac{\Delta f}{b}\right) = \exp(\epsilon)$$

So we can exactly choose what privacy we want to provide!

In general

In general

- **For asking questions about individuals in the database**

In general

- **For asking questions about individuals in the database**
- **Probability p of giving a truthful response, probability 1-p of giving a random response**

In general

- For asking questions about individuals in the database
- Probability p of giving a truthful response, probability 1-p of giving a random response
- Allows users the possibility to deny that data given about them was truthful

Limitations

Limitations

- Probability of giving a truthful answer must be decently high to ensure accuracy of analytical data

Limitations

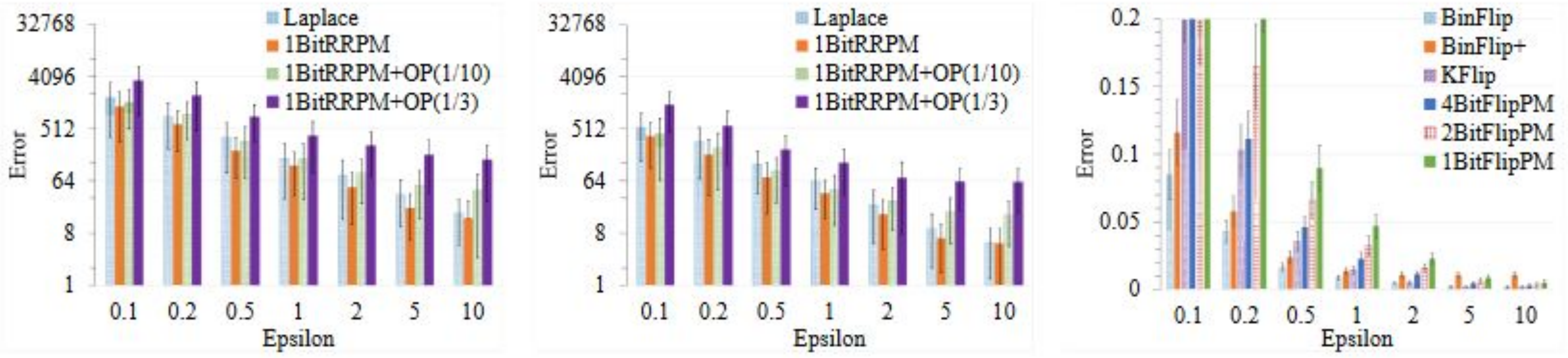
- Probability of giving a truthful answer must be decently high to ensure accuracy of analytical data

- This can impact the whole privacy goal of this mechanism

Usage in practice

Usage in practice

- **Collecting telemetry data in Windows:**



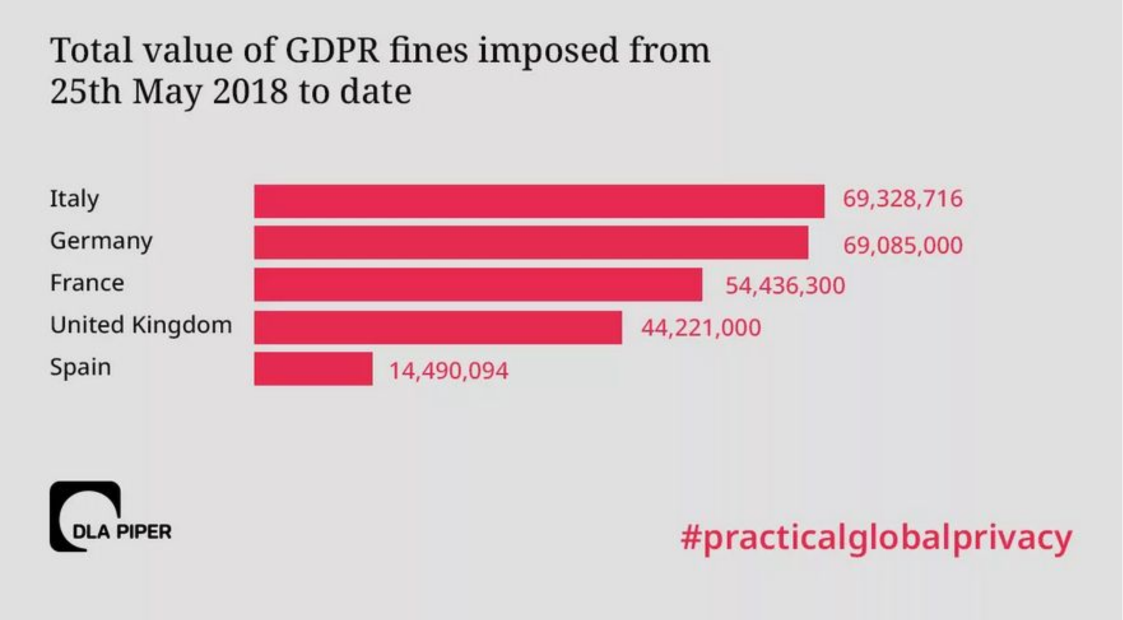
(a) Mean ($n = 0.3 \times 10^6$) (b) Mean ($n = 3 \times 10^6$) (c) Histogram ($n = 0.3 \times 10^6$)

Figure 2: Comparison of mechanisms for mean and histogram estimations on real-world datasets

https://www.microsoft.com/en-us/research/publication/collecting-telemetry-data-privately/

Usage in practice

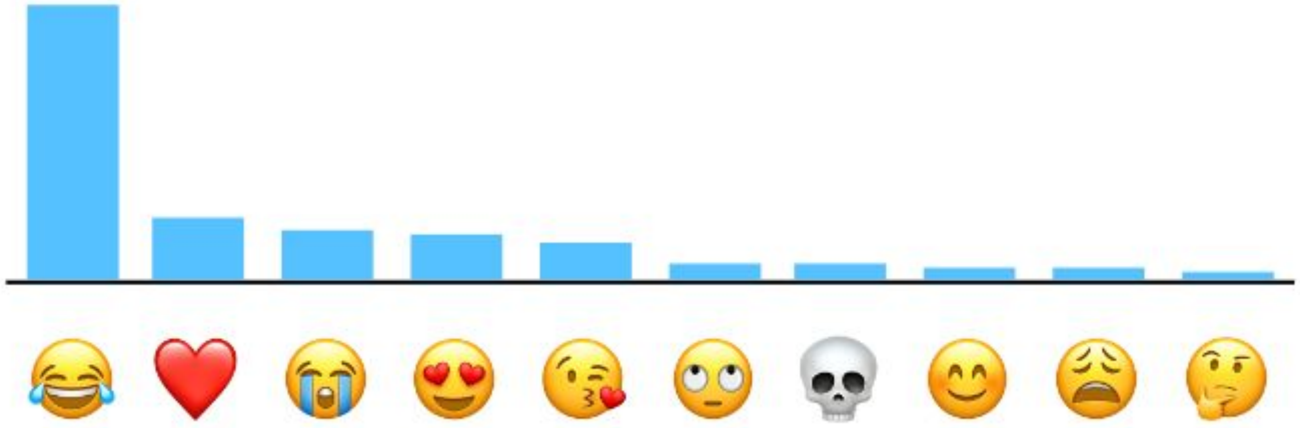
- **Differential privacy is important:**



https://www.dlapiper.com/en-us/insights/publications/2021/01/dla-piper-gdpr-fines-and-data-breach-survey-2021

Usage in practice

- **Most popular emojis by Count Mean Sketch used by Apple:**

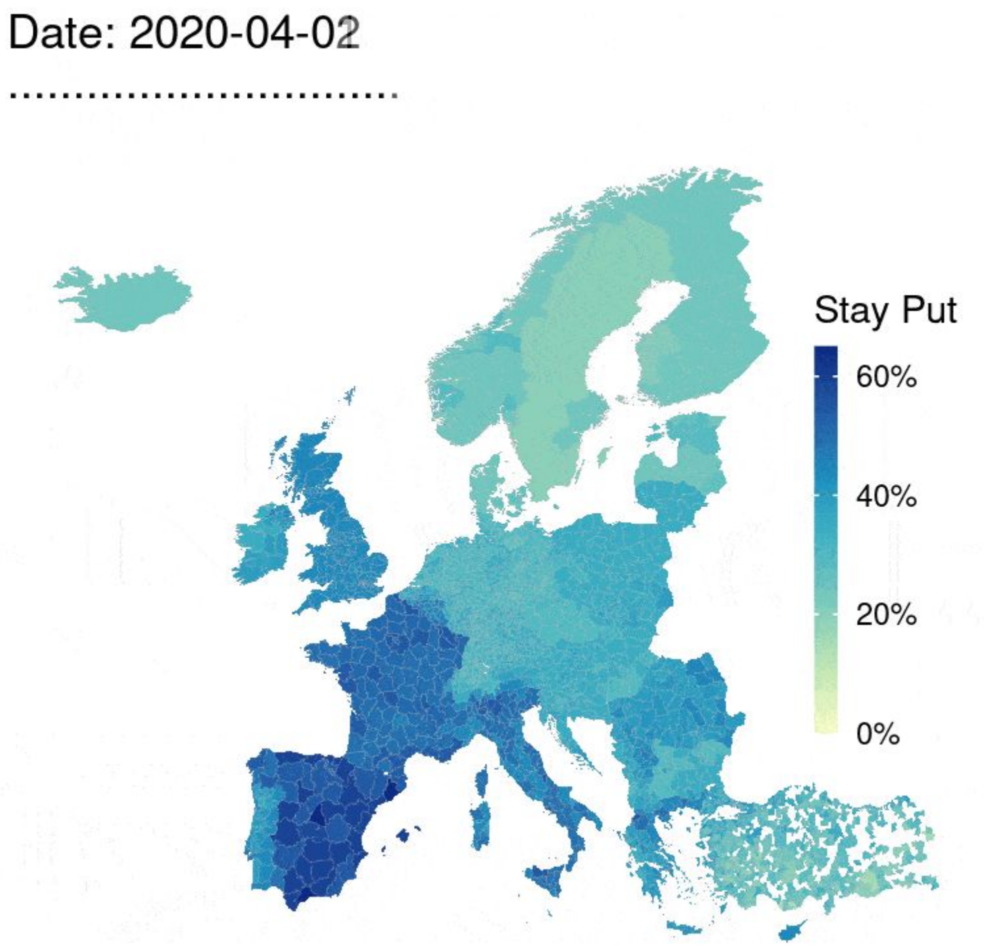


The Count Mean Sketch technique allows Apple to determine the most popular emoji to help design better ways to find and use our favorite emoji. The top emoji for US English speakers contained some surprising favorites.

https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Usage in practice

- **Even Facebook uses it:**
Location tracking during COVID



https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/

Do/would you use one of these methods?

?

Questions?