

Ethics in Data Science (Part 2, Disclosure)

Daniel Lawson — University of Bristol

Lecture 11.2 (v2.0.0)

Signposting

- ▶ Part 1 covers ethics and the law,
- ▶ This is part 2 covering Privacy and disclosure,
- ▶ Part 3 covers Fairness and interpretability.

Protecting Privacy: Anonymity

- ▶ It is not enough to **anonymise** data by removing identifiers. It may be de-anonymised.
- ▶ For example: the Netflix competition was **partially de-anonymised** by comparing to public datasets, IMDB ratings, etc., resulting in a lawsuit. **Narayanan and Shmatikov 2008**
- ▶ Formally: **quasi-identifiers** are statistically valuable information that can be combined with additional data to produce identifiers.

Statistical disclosure attacks

- ▶ Statistical disclosure describes how **legitimate access** to a database can be used to extract confidential information regarding identity, attributes, or membership.
- ▶ It works by:
 - ▶ Assuming that users can query **statistical**, anonymised properties,
 - ▶ Making repeated queries, specific information can be extracted using the **intersection of answers**,
 - ▶ Disclosure attacks are therefore a form of **elevation of access rights**: obtaining access that was not intended to be given.

Example of statistical disclosure attacks

- ▶ By knowing specific details, or observing large-scale results, additional information can be extracted about identifiers.
 - ▶ **attribute disclosure**: e.g. If we know when Mr R moved out of an area, we can obtain his salary by querying the average salary in the region before and after he moved.
 - ▶ **identity disclosure**: e.g. By making a large number of queries containing different attribute ranges, we can associate each identifier with a particular attribute value.
 - ▶ **membership disclosure**: e.g. Similarly, we might obtain information about whether an identifier is in a particular group such as “HIV patient”.

Protecting against statistical disclosure

- ▶ The main lines of defense are:
 - ▶ Limiting the **volume** of queries, i.e. not permitting more than M queries per actor.
 - ▶ Limiting the **detail** of queries, i.e. ensuring that all values are shared by at least k entries (Formally: k -anonymity).
 - ▶ Limiting the **accuracy** of queries, i.e. adding noise to reported answers.

Quantifying vulnerability to statistical disclosure attacks

- ▶ There is a robust theory called **differential privacy**¹ which formalises the vulnerability of a dataset/release mechanism to disclosure:
 - ▶ A dataset allows querying a summary statistic, A .
 - ▶ Adversary proposes two datasets S and S' that differ by only one row or example, and a test set Q .
- ▶ A is called ϵ -differentially private iff:

$$\left| \log \frac{\Pr(A(S) \in Q)}{\Pr(A(S') \in Q)} \right| \leq \epsilon,$$

- ▶ i.e. the change in log-probability is bounded.
- ▶ A is called (ϵ, δ) -differentially private iff:

$$\Pr(A(S) \in Q) \leq \exp(\epsilon) \Pr(A(S') \in Q) + \delta,$$

- ▶ where δ is typically smaller than any polynomial.
- ▶ i.e. $\delta = 0$ leads to ϵ -differential privacy.

¹The Algorithmic Foundations of Differential Privacy by Dwork and Roth (2014).

Continuous outcomes: the Laplace mechanism

- ▶ Continuous outcomes are particularly tricky because the answer can reveal location via the scale of the noise.
- ▶ Adding noise from a light-tailed distribution does not ensure “plausible deniability”, which can be exploited to de-noise the data.
 - ▶ e.g. if my salary is truly large, then it will remain large after perturbation.
- ▶ To address this, **heavy tailed** distributions are favoured, to place finite weight on “all” values.
- ▶ i.e. Instead of reporting $f(x)$ we report:

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + \Delta$$

- ▶ Where $\Delta \sim \text{Lap}(x|b) = (1/2b) \exp(-|x|/b)$,
 - ▶ $b = \Delta f / \epsilon$ is chosen in terms of the desired sensitivity $\Delta f = \max_{x,y:|x-y|=1} |f(x) - f(y)|$.
 - ▶ This is analogous to a standard deviation for the L-1 norm.
- ▶ i.e. we add noise scaled to the output at the scale of variation, but which can induce any value with non-vanishing probability.

Beyond simple differential privacy

- ▶ Privacy means maintaining **plausible deniability** so that *any outcome could have happened*, for any specific case.
 - ▶ However, protecting from disclosure attacks using audit is provably not possible in general.
- ▶ **Correlated** (worst case, repeated) queries raise particular problems:
 - ▶ It is possible to “learn” the noise, average it out, and obtain the true value.
- ▶ **Audit** of queries is clearly essential to prevent these and other attacks.
- ▶ A partial solution is to consider creating **correlated noise** in the response,
 - ▶ So if you ask a similar question, you get similar noise.
- ▶ But what if the individual exists in k different databases?
 - ▶ It turns out that we can still describe ϵ -privacy in this case, though our controls are more limited.
 - ▶ We essentially have to allow for k independent noise observations.

Example: Randomized response

- ▶ Consider the question, “Have you taken drugs this week?”
- ▶ We want to know population-level answers without revealing whether anyone specifically answered “yes”.
- ▶ We apply the following algorithm:
 1. Flip a coin.
 2. If **tails**, respond truthfully.
 3. If **heads**, flip a second coin and respond “Yes” on heads, and “No” on tails.
- ▶ This protocol is $(\ln 3, 0)$ -differentially private (see Worksheet).
 - ▶ Intuition: We compute the odds ratio of the truth, given the answer.
 - ▶ If someone is reported to have said “Yes”, there is a 3:1 odds that they really did take drugs this week.

Practice for sharing anonymised data

- ▶ The **ONS** suggest that:
- ▶ *The question to ask is – could an intruder discover any **protected information** from (provided information)? This breaks down into the following three questions:*
 1. Can **any individual be identified** from the table, with any degree of certainty?
 2. If so, is any **new information revealed** about them (attribute disclosure)?
 3. Is any information revealed about any **other living person** connected with them?
- ▶ *Common SDC techniques include:*
 - ▶ **collapsing categories** to reduce the sparsity of the table (for example, aggregating single year ages to five-year groups, or five-year age groups to 10-year groups) (non-perturbative)
 - ▶ **aggregating the data** over a greater period of time, or a larger geographical area (non-perturbative)
 - ▶ **rounding** to a specific base to avoid very small numbers (usually three or five) (perturbative)
 - ▶ **suppressing very small numbers** (usually numbers less than three) (perturbative)

Discussion

- ▶ Each form of anonymisation implies a slightly different question, changing the baseline.
- ▶ This sort of privacy protection must be considered alongside the usual forms of data security.
- ▶ The **consequences** of a data reveal are complex and contingent:
 - ▶ It may stop at learning that a single actor has an uninteresting value of a feature.
 - ▶ It may instead set off a cascade of consequential knowledge leading to a complete reveal of the whole database.
 - ▶ Typically, outside information is involved in the worst such problems.
- ▶ High profile failures include:
 - ▶ Linking voter registration to “anonymised” medical data **Sweeney 1997**,
 - ▶ Linking “anonymised” Netflix data back to individuals **Narayanan and Shmatikov 2008**.

Reflection

- ▶ What is the difference between “anonymising” data and protecting privacy?
- ▶ What role does cyber security, statistics, and other measures have in protecting privacy?
- ▶ What are the main approaches to protecting privacy?
- ▶ What impact does protecting privacy have on the utility of databases?
- ▶ By the end of the course, you should:
 - ▶ Be able to define ϵ and (ϵ, δ) -privacy,
 - ▶ Be able to state the methods by which privacy is protected,
 - ▶ Understand the value and limitations of the approach at a high level.

Signposting

- ▶ Still to come:
 - ▶ 12.1.3 Ensuring algorithms are fair and interpretable.
- ▶ References:
 - ▶ **The Algorithmic Foundations of Differential Privacy** by Dwork and Roth (2014).
 - ▶ **ONS** policy on disclosure control.
 - ▶ **Sweeney 1997**. Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine Ethics*, 25:98–110.
 - ▶ **Narayanan and Shmatikov 2008**. Robust de-anonymization of largesparse datasets (how to break anonymity of the netflix prize dataset). *IEEE Sec. and Priv.*
 - ▶ **Statistical Disclosure Attacks** by George Danezis.