Impossibility of inferential privacy and limits of anonymisation

Subhashis Banerjee, Subodh Sharma

Department of Computer Science and Engineering
IIT Delhi

November 20, 2020

Inferential and differential privacy

Differential Privacy, Dwork, ICALP, 2006

Dalenius (1971) desideratum:

Nothing about an individual should be learnable from the database that cannot be learned without access to the database.

General impossibility of inferential privacy:

A formalization of Dalenius' goal along the lines of *semantic security* cannot be achieved.

[Semantic security against an eavesdropper says that nothing can be learned about a plaintext from the ciphertext that could not be learned without seeing the ciphertext]

Inferential and differential privacy

The *Fundamental Law of Information Recovery* states that overly accurate answers to too many questions will eventually destroy privacy in a spectacular way.

The goal of algorithmic research on *differential privacy* is to postpone this inevitability as long as possible.

Anonymisation vs utility

Data cannot be fully anonymised and remain useful.

- Publish data after *anonymisation* or removal of *personally identifiable information*.
- However, the richness of the data enables "naming" an individual by a sometimes surprising collection of fields
  - combination of zip code, date of birth, and sex.
  - names of three movies and the approximate dates on which an individual watched these movies.
  - Geolocation, six times a day.
  - Even Census.
  - medical records of the Governor of Massachussetts were identified by matching anonymised medical encounter data with (publicly available) voter registration records.
  - viewing history of Netflix subscribers from anonymised training data released by Netflix.

Re-Identification of "anonymized" records is not the only risk

- A collection of medical encounter records from a specific urgent care center on a given date may list only a small number of distinct complaints or diagnoses.
- The additional information that a neighbour visited the facility on the date in question gives a fairly narrow range of possible diagnoses for the neighbour's condition.

Queries over large sets are not protective

Consider the *differencing attack*:

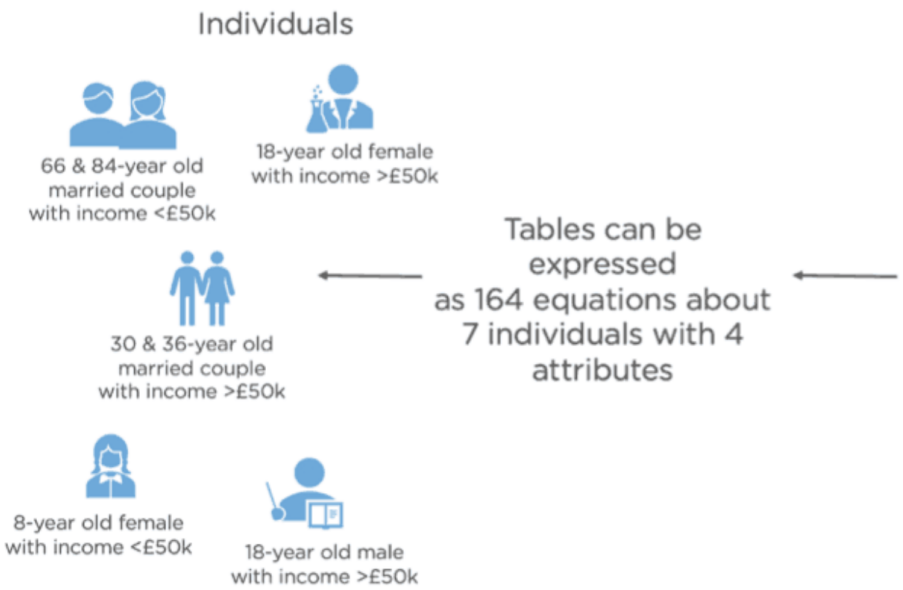
1. How many people in the database have Covid?
2. How many people, not named *suban*, in the database have Covid?

Query auditing is problematic

- Refusing to answer a query may itself sometimes be disclosive.
- For a sufficiently rich query language there may not even exist an algorithmic procedure for deciding if a pair of queries constitutes a differencing attack.

Summary statistics are not "Safe"

- Census.
- Linear equation attacks.
- Subsetting attacks.
- Differencing attacks.



Summary Statistics

	Count	Median Age	Mean Age
Total	7	30	38
Breakdown by gender			
Female	4	30	33.5
Male	3	30	44
Breakdown by annual income bracket			
<£50k	4	51	48.5
>£50k	3	24	24
Breakdown by marital status			
Married	4	51	54
Breakdown by gender & annual income bracket			
Female & <£50k	3	36	36.7

[Ref: Simson Garfinkel, Senior Scientist at the US Census Bureau's team for disclosure avoidance in PETS 2019]

“Ordinary” facts are not “OKay”

- Consider Mr. SugarLover, who regularly buys bread, year after year, until suddenly switching to rarely buying bread.
- An analyst (IT expert/Machine Learner) might conclude Mr. SugarLover most likely has been diagnosed with Type 2 diabetes.
- The analyst might be correct, or might be incorrect; either way Mr. SugarLover is harmed.

Just a few

There is often a claim that technique is adequate, as it compromises the privacy of "just a few" participants.

- The outliers may be precisely those people for whom privacy is most important.
- Not intrinsically without merit. There may be a social judgment - a weighing of costs and benefits - possible.
- Need to develop a well-articulated definition of privacy consistent with the "just a few" philosophy.

Impossibility of absolute disclosure prevention

Differential Privacy, Dwork, ICALP, 2006

The setting:

- A "sanitized" version of the collected data.
  - the literature uses terms such as "anonymisation" and "de-identification". Traditionally, sanitisation employs techniques such as data perturbation and sub-sampling, as well as removing well-known identifiers such as names, birthdates, phone numbers and social security numbers.
- Some notion of utility – after all, a mechanism that always outputs the empty string, or a purely random string, clearly preserves privacy.
- An adversary has access to an auxiliary information generator.

Impossibility of absolute disclosure prevention

Consider:

- An *utility vector w* (binary) of a fixed length.
- A *privacy breach* for a database is described by a Turing machine $\mathcal{C}$ that takes as input a description of a distribution $\mathcal{D}$ on databases, a database DB drawn according to this distribution, and a string – the purported privacy breach – and outputs $1/0$.
- The adversary *wins*, with respect to $\mathcal{C}$ and for a given $(\mathcal{D}, DB)$ pair, if it produces a string s such that $\mathcal{C}(\mathcal{D}, DB, s)$ accepts.
- An *auxiliary information generator* is a Turing machine that takes as input a description of the distribution $\mathcal{D}$ from which the database is drawn as well as the database DB itself, and outputs a string, z, of auxiliary information. This string is given both to the adversary and to a *simulator*.
- The simulator has no access of any kind to the database; the adversary has access to the database via a *privacy mechanism*.

Impossibility: the main result

Assumption

1. $\forall 0 < \gamma < 1$, $\exists n_\gamma$ $\Pr_{DB \in_R \mathcal{D}}[\mid DB \mid > n_\gamma] < \gamma$; moreover n_γ is computable by a machine given $\mathcal{D}$ as input.
2. There exists an ℓ such that both the following conditions hold:
   - 2.1 Conditioned on any privacy breach of length ℓ, the min-entropy of the utility vector is at least ℓ.
   - 2.2 Every $DB \in \mathcal{D}$ has a privacy breach of length ℓ.
   - 2.3 $\Pr[\mathcal{B}(\mathcal{D}, San(DB)) \text{ wins }] \leq \mu$ for all interactive Turing machines $\mathcal{B}$, where μ is a suitably small constant. The probability is over the coin flips of $\mathcal{B}$, the $San()$, and the choice of $DB \in_R \mathcal{D}$.

Theorem

Fix any privacy mechanism $San()$, privacy breach decider $\mathcal{C}$ and a suitable large constant Δ. There is an auxiliary information generator $\mathcal{X}$ and an adversary $\mathcal{A}$ such that for all distributions $\mathcal{D}$ satisfying the assumption and for all adversary simulators $\mathcal{A}^*$,

$$\Pr[\mathcal{A}(\mathcal{D}, San(\mathcal{D}, DB), \mathcal{X}(\mathcal{D}, DB)) \text{ wins }] - \Pr[\mathcal{A}^*(\mathcal{D}, \mathcal{X}(\mathcal{D}, DB)) \text{ wins }] \geq \Delta$$

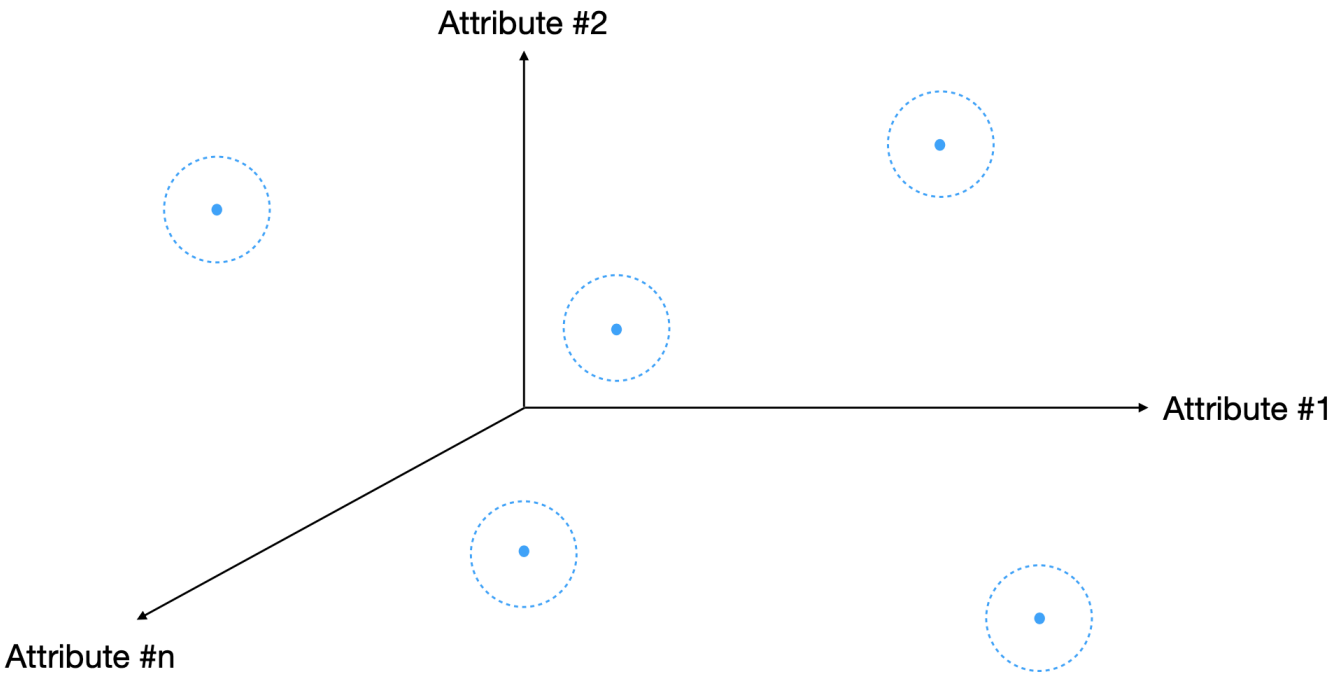
Proof sketch

- Assume (a special case) that the adversary learns the entire utility vector w. Otherwise the assumptions needs some modifications.
- For any fixed γ it is possible to find an ℓ_γ such that ℓ_γ satisfies both clauses of Assumption (2). Note that all but a γ fraction of the support of $\mathcal{D}$ is strings of length at most n_γ.
- On input $DB \in_R \mathcal{D}$, $\mathcal{X}$ randomly chooses a privacy breach y for DB of length $\ell = \ell_\gamma$, which exists with probability $1 - \gamma$. It also computes the utility vector, w. Finally, it chooses a seed s and uses a strong randomness extractor to obtain from w an ℓ-bit almost-uniformly (within ϵ of U_ℓ) distributed string r; i.e., $r = \mathrm{Ext}(s, w)$. The auxiliary information will be $z = (s, y \oplus r)$.
- Adversary learns all of w, from s it can obtain $r = \mathrm{Ext}(s, w)$ and hence y.
- $\mathcal{A}^*$ wins with probability (atmost) bounded by μ, yielding a gap of at least $1 - (\gamma + \mu + \epsilon)$.

Robust de-anonymisation of large sparse datasets

Narayanan and Shmatikov, IEEE S&P, 2008

Anonymisation is a myth



Individuals map to points in a sparse high-dimensional space where they are uniquely identifiable even after adding a lot of noise.

The setting

- A database $\mathcal{D}$ is an $N \times M$ matrix where each row is a record associated with some individual, and the columns are attributes. No attributes is a *quasi-identifiers*.
- The shopping history of even the most profligate Amazon shopper contains only a tiny fraction of all available items. These attributes *non-null* (denoted as $\perp$).
- The set of non-null attributes is the *support* of a record (denoted $\text{supp}(r)$). Similarly *support* of a column.
- The distribution of per-attribute support sizes is typically heavy- or long-tailed, roughly following the power law.
- Similarity over two records r_1, r_2 is defined using the *Cosine* similarity:

$$\text{Sim}(r_1, r_2) = \frac{\sum_i \text{Sim}(r_{1i}, r_{2i})}{|\ \text{supp}(r_1) \cup \text{supp}(r_2)\ |}$$

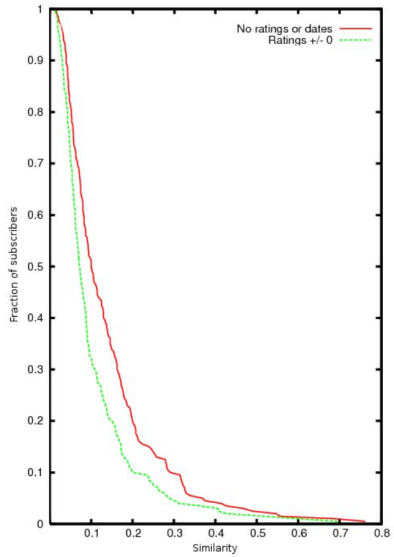
Sparse databases

Sparsity

A database $\mathcal{D}$ is (ϵ, δ)-sparse w.r.t. the similarity measure Sim if

$$\Pr_r \left[\text{Sim}(r, r') > \epsilon, \quad \forall r' \neq r\right] \leq \delta$$

Netflix Prize dataset (500,000) is overwhelmingly sparse



De-anonymisation model

Adversary

- For a record r randomly from the database, give auxiliary information or background knowledge related to r to the adversary. Aux : $X^M \rightarrow X^M$ of r's attributes.
- Given this auxiliary information and an anonymized sample D' of D, the adversary's goal is to reconstruct attribute values of the entire record r.

Privacy breach

An arbitrary subset $\hat{D}$ of a database D can be (θ, ω)-deanonymised w.r.t. auxiliary information Aux if there exists an algorithm A which, on inputs $\hat{D}$ and Aux(r) where $r \leftarrow D$

- If $r \in \hat{D}$ outputs r' such that $\Pr[\text{Sim}(r, r') \geq \theta] \geq \omega$
- If $r \notin \hat{D}$ outputs $\perp$ with probability at least ω

Attack

Algorithm

- $\text{Score}(\text{aux}, r') = \sum_{i \in \text{supp}(\text{aux})} w(i)\text{Sim}(\text{aux}, r'_i)$ where $w(i) = 1/\log |\ \text{supp}(i)\ |$.
- Apply the matching criterion to the resulting set of scores and compute the matching set; if the matching set is empty, output $\perp$ and exit.
- Output $r' \in \hat{D}$ with the highest score, or a *probability distribution* based on the matching score, as appropriate.

Main results

Theorems

- Let $0 < \epsilon, \delta < 1$, and let D be the database. Let aux $=$ Aux(r) consist of at least $m \geq \frac{\log N - \log \epsilon}{-\log 1-\delta}$ randomly selected attribute values of the target record r, where $\forall i \in \text{supp(aux)}, \text{Sim}(aux_i, r_i) \geq 1 - \epsilon$. Then D can be $(1 - \epsilon - \delta, 1 - \epsilon)$-deanonymized w.r.t. Aux.
- If r' is a false match then $\Pr_{i \in \text{supp}(r)} [\text{Sim}(r_i, r'_i) \geq 1 - \epsilon] < 1 - \delta$.
- Let ϵ, δ and aux be as in above. If the database D is $(1 - \epsilon - \delta, \epsilon)$-sparse, then D can be $(1, 1 - \epsilon)$-deanonymised.

Netflix prize data

The Data

Netflix publicly reeased a dataset containing 100, 480, 507 movie ratings, created by 480, 189 Netflix subscribers between December 1999 and December 2005.

FAQ item

"Is there any customer information in the dataset that should be kept private?"

FAQ answer

"No, all customer identifying information has been removed; all that remains are ratings and dates. This follows our privacy policy [. . .] Even if, for example, you knew all your own ratings and their dates you probably couldn't identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isn't a privacy problem is it?"

[Details of the attack in the paper.]

Sanitized?

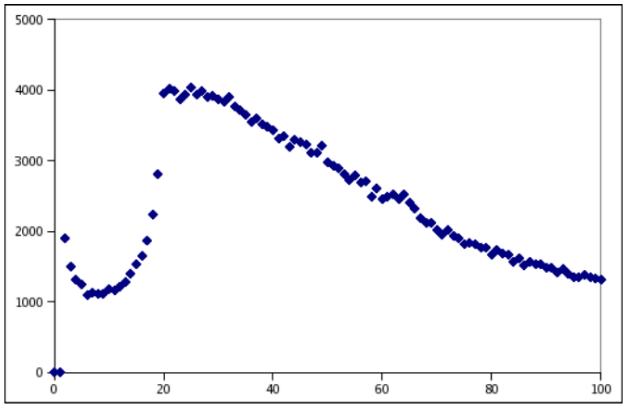


Figure 2. For each $X \leq 100$, the number of subscribers with X ratings in the released dataset.

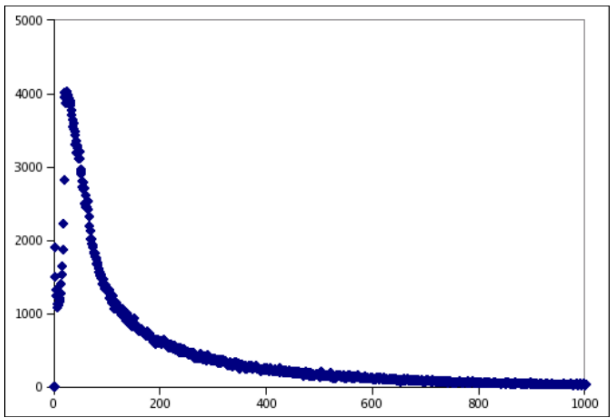


Figure 3. For each $X \leq 1000$, the number of subscribers with X ratings in the released dataset.

Deanonimization

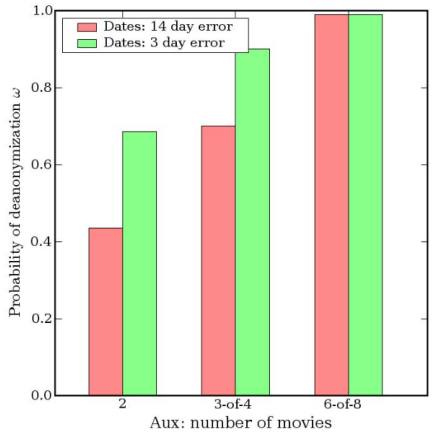


Figure 4. Adversary knows exact ratings and approximate dates.

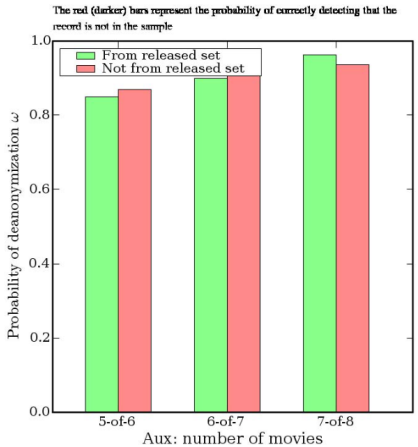


Figure 5. Same parameters as Fig. 4, but the adversary must also detect when the target record is not in the sample.

Entropic deanonimization

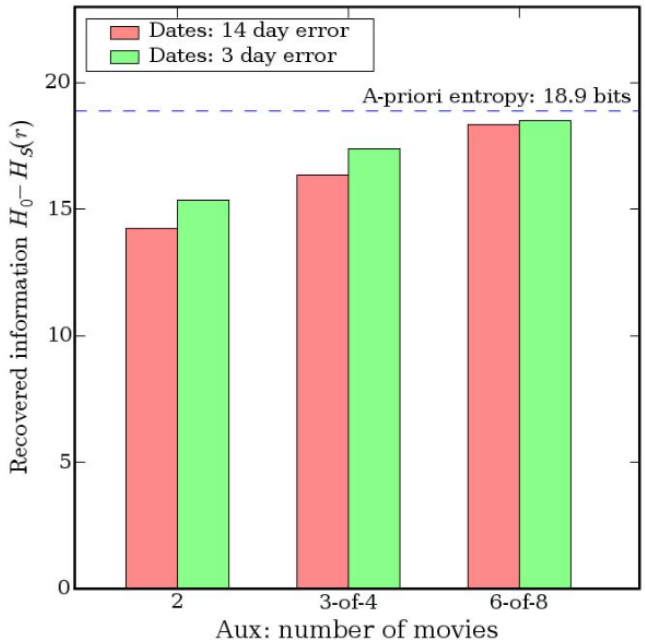


Figure 6. Entropic de-anonymization: same parameters as in Fig. 4.

Dependence on adversary knowledge

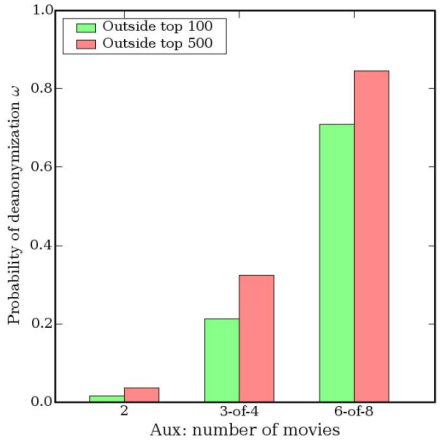


Figure 8. Adversary knows exact ratings but does not know dates at all.

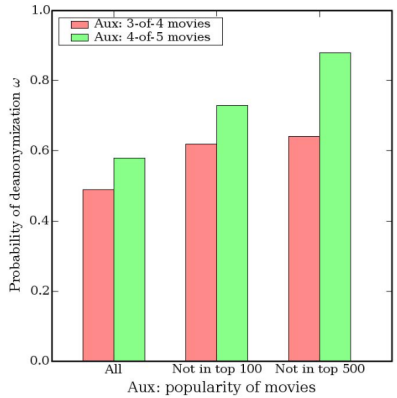


Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).

Differential privacy

A randomized function $\mathcal{K}$ gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq \text{Range}(\mathcal{K})$,

$$\Pr\left[\mathcal{K}(D_1) \in S\right] \leq \exp\left(\epsilon\right) \times \Pr\left[\mathcal{K}(D_2) \in S\right]$$

- For instance, Subodh's presence or absence in the database will not significantly change his chance of qualifying for insurance coverage.
- Defensive, because the concept only covers for *additional* harm.