PRIVACY ENHANCING TECHNOLOGIES

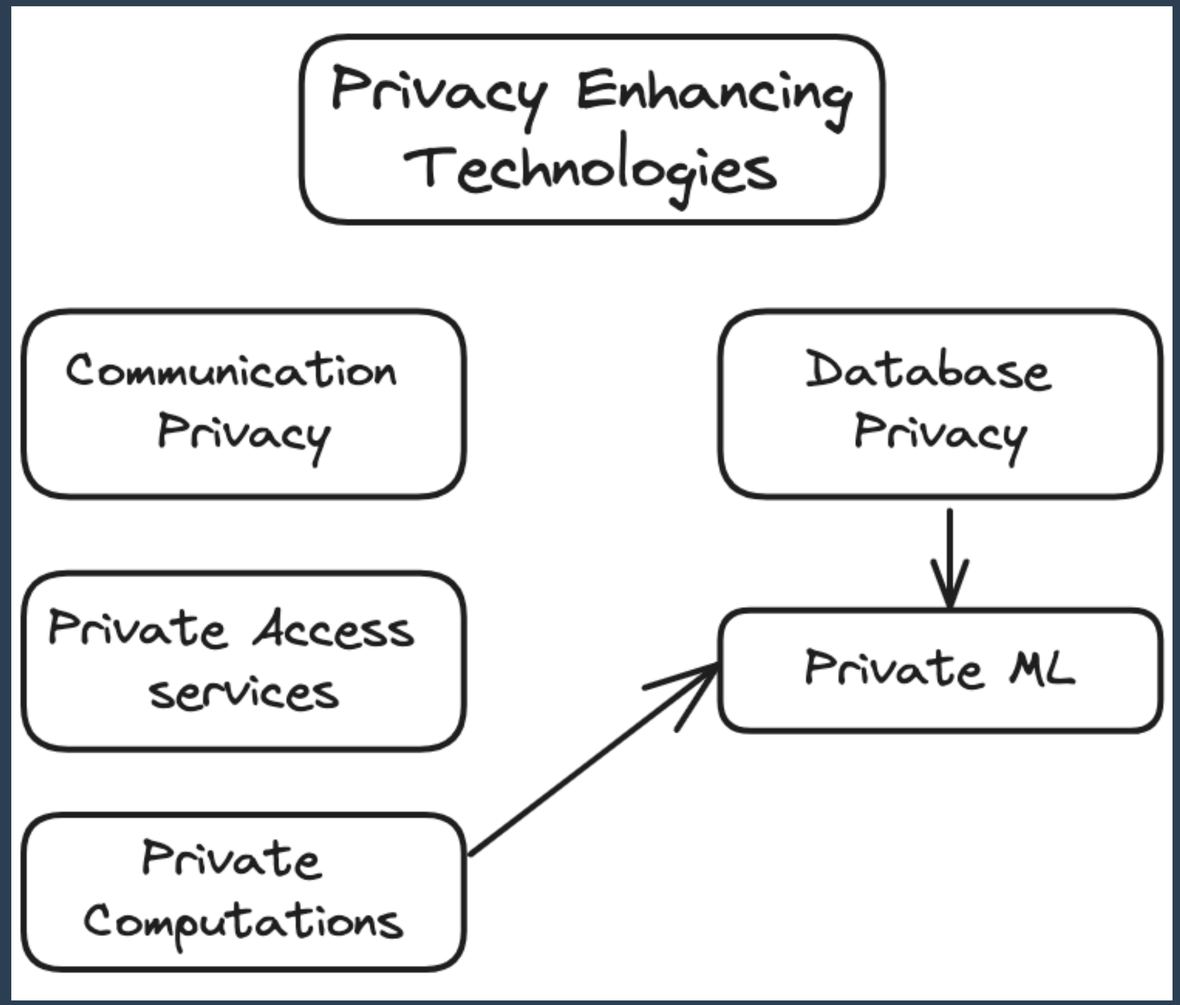
Database Privacy and Private ML Training Approaches

Acknowledgement: Some of the slides in this set are adaptations of lecture slides of Dr. Olaf Hartig (Linköping University).

2024-01-26 | JENNI REUBEN

INTRODUCTION

- Hard Privacy
  - avoid or reduce as much as possible in placing any trust in the parties involved in serving the service to the end-user



Sign Post of Day I and Day II topics

WHOSE PRIVACY

- **Respondent Privacy**

  Protecting the information of the individuals to which the records in a database corresponds to

- **Owner Privacy**

  Protecting the information of each entities that are coming together for computing a query

- **End-user Privacy**

  Protecting end-user's queries to an interactive databases such as search engines.

STATISTICAL DATABASES

- enable its users to retrieve statistical knowledge from a subset of the population that the database represents
- exploited for variety of reasons such as disease control, market research, medical research
- we should be interested in the public availability of such data:

  results from such data can contribute to expanding our knowledge about e.g., diseases
- However, those datasets contain confidential information about the respondents who have given their information to the database
- Can the users (researchers, analysts or the data consumers) of such databases be trusted?

WHAT ARE THE PRIVACY RISKS?

- Anonymity in terms of unlinkability:
  - The anonymity of a subject w.r.t an attribute may be defined as unlinkability of this subject and this attribute [Pfitzmann17]
- Two types of linkage from an adversary's perspective;
  - Record linkage: re-identify the individual that the records in the published database corresponds to, by linking the publicly available information to the information in the published data (that is presumably free of explicit identifiers)
  - Attribute linkage: accurately infer the confidential attribute values of an individual or a set of individuals represented in the underlying database, such as inference would have been possible without the access to the data.

RECORD LINKAGE EXAMPLE

- In Massachusetts, USA, the Group Insurance Commission (GIC) is responsible for purchasing health insurance for state employees
- Sweeney paid $20 to buy the voter registration list for Cambridge, MA
  - Former governor (William Weld) of MA lives in Cambridge, MA hence his record is in the Voters DB
  - 6 people in Voters DB shares his DOB
  - Of which only 3 of them were men
  - Of which only 1 record matches the Weld's ZIP code.
  - Mr. Weld's medical information, learned!

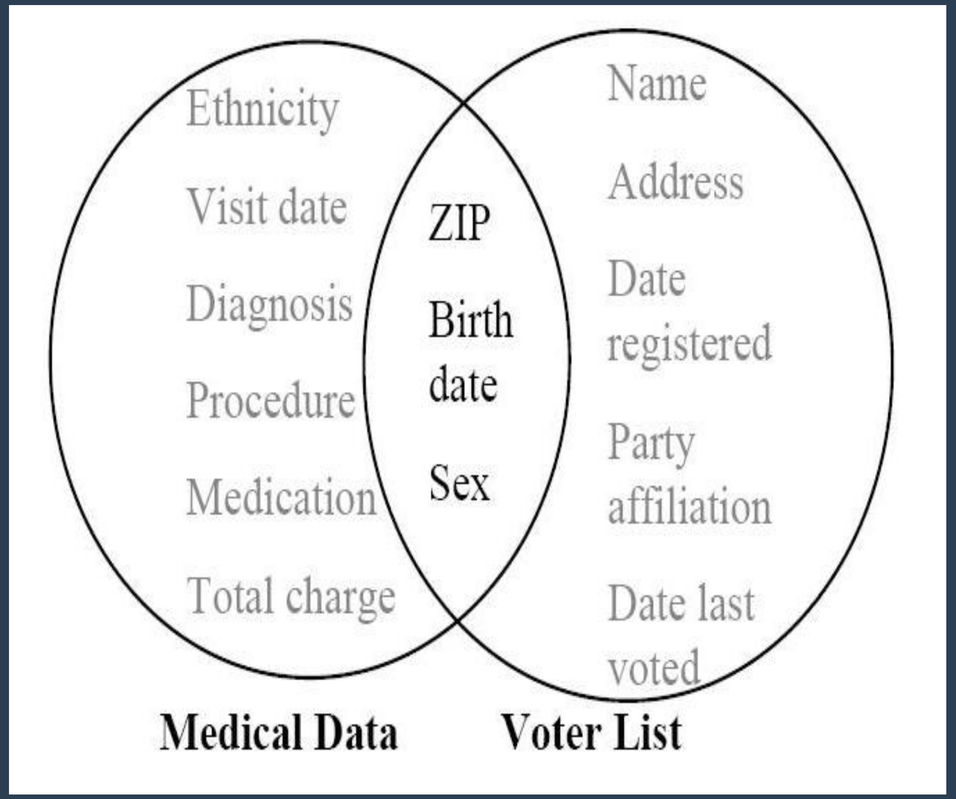


Figure taken from [Fung10]

CATEGORIES OF IDENTIFIERS

- Explicit Identifiers:
  - Attributes that unambiguously identify the respondent. E.g., name, social security number, IP address, etc.
- Quasi Identifiers:
  - A set of non-sensitive attributes that when combined may lead to unambiguously identify the respondent. E.g., gender, age, telephone number, zip code etc.
- Sensitive attributes:
  - Attributes that contain sensitive information of the respondents. E.g., disease, salary. etc.
- Non-sensitive attributes:
  - All other attributes that captures the respondents' non-sensitive information

THE CHALLENGE

- Statistical databases such as the databases of the U.S census Bureau contain confidential information such as age, sex, income, credit ratings, types of disease, etc.
- how to publish statistics about the underlying population, which is based on their confidential attributes while not revealing anything about those individual. The privacy, utility trade-off
- We need a non-trivial way to limit the disclosure of confidential information
- Fact: 87% of the US population can be identified by the combination of ZIP, DOB and sex.
- Statistical Disclosure Control (SDC) or Statistical Disclosure Limitation (SDL)
  - limits the disclosure of confidential information from the published statistics

K-ANONYMITY DEFINITION

- A dataset or datable T is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in T is shared by at least $k-1$ records.
- Let T be a table and X be a subset of the attributes of T. For every record t in T we write $t[X]$ to denote the sequence of values that t has for the attributes in X.
- Example:
  - If X = {ZIP, Age, Sex} and say t is the first tuple in T
  - then, $t[X]$ is (12211, 18, M)
  - If X = {ZIP, Sex}, then $t[X]$ is (12211, M)

12211	18	M	Arthritis
12244	19	M	Cold
12245	27	M	Heart problem
12377	27	M	Flu
12377	27	F	Arthritis
12391	34	F	Diabetes
12391	45	F	Flu

T

K-ANONYMITY DEFINITION

- Let T be a table and QI_T be the quasi-identifier of T. T satisfies k-anonymity if for every tuple t in T there exist (at least) $k-1$ other tuples $t_1, t_2, \cdots, t_{k-1}$ in T such that we have t[QI_T] = t1[QI_T] = t2[QI_T] = tk-1[QI_T].

12211	18	M	Arthritis
12244	19	M	Cold
12245	27	M	Heart problem
12377	27	M	Flu
12377	27	F	Arthritis
12391	34	F	Diabetes
12391	45	F	Flu

T

122**	18-19	M	Arthritis
122**	18-19	M	Cold
*	27	*	Heart problem
*	27	*	Flu
*	27	*	Arthritis
12391	≥ 30	F	Diabetes
12391	≥ 30	F	Flu

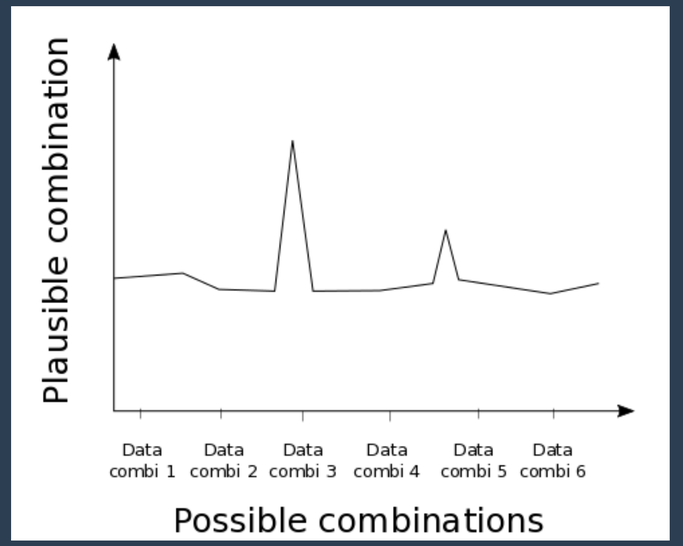
2-anonymous table *T**

DATABASE RECONSTRUCTION ATTACK (DRA)

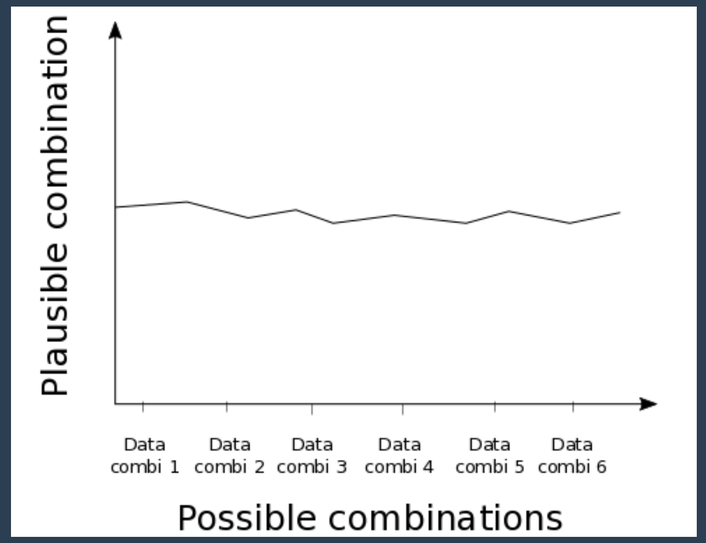
- It turns out k-anonymity is not sufficient against inference attacks, so what if only aggregate data is released
- But by simply observing the query answers/results of some random queries, one can recover the confidential data of the individuals in the underlying population.
- Take for example:
  - U.S census bureau database which contains answers given by the citizens of the United States
  - The census bureau publishes statistics such as how many people belonging to a race, live in a particular block
  - The attack then is to guess using brute force computation, all the possible combinations of answers that people could have given to questions concerning race and block, and find out the possible combinations that best fit the published statistics [Dinur03].

WAYS TO MEASURE OF PRIVACY

- measure of loss of respondent privacy is the level of certainty in an attacker's ability in determining the plausibility of some possible combinations of data.
- Publishing less statistics, then there are more plausible combinations of data that accurately fits the data
- Even lesser statistics are published means, increase in the amount of data combinations that plausibly fit the released statistics.
- Idea! to protect respondent privacy – make all possible combinations of data from the respondents to be equally plausible.
- There is an inevitable trade-off between accuracy of the published results and not revealing information of the record owners in the underlying database.



A few possible data combinations are plausible



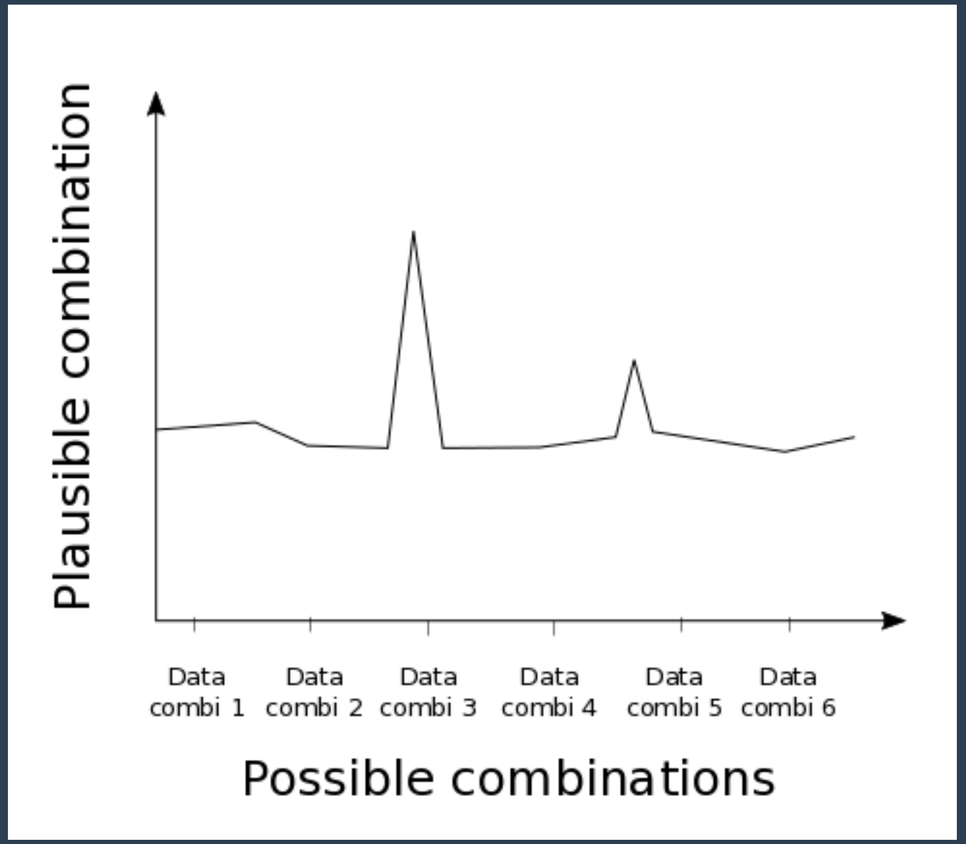
All possible data combinations are plausible

DIFFERENTIAL PRIVACY

- How then to publish data for data analyses?
- because increasing the uncertainty level of the adversaries, decreases the query results' accuracy
- Further, if random noise is added a bunch of times to a statistical query result, it is possible to get back the true results by taking the average of the noisy results, which cancels out the noise.
- Differential privacy model that provides a **strong** privacy guarantee, yet at the cost of small loss in the accuracy of the results.

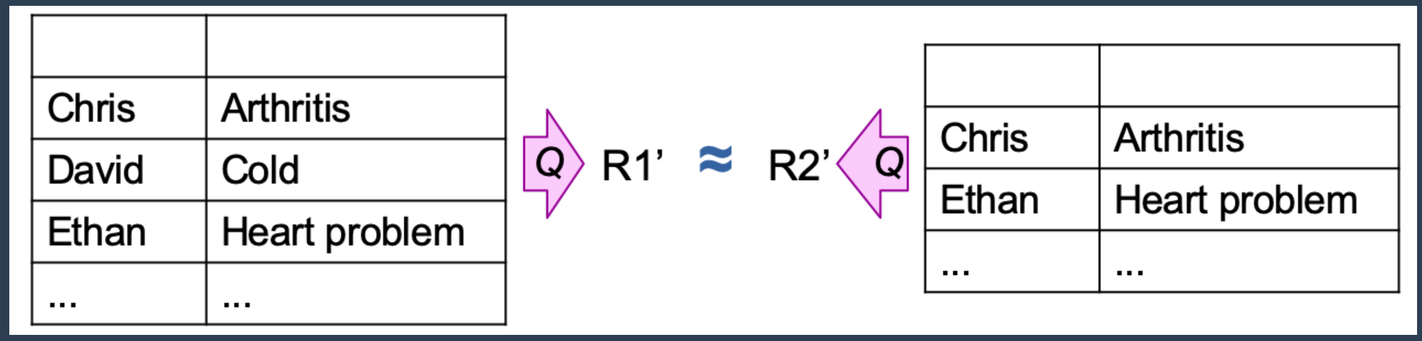
DIFFERENTIAL PRIVACY

- The differential privacy model provides a way to quantifies the plausibility peak (i.e the loss of privacy) and bounds (that is to say the maximum) the loss of privacy for the individuals in the underlying dataset, as a consequence of publishing results computed on their data.



The plausibility/possibility plot with a few peaks that stands out

DIFFERENTIAL PRIVACY EXAMPLE



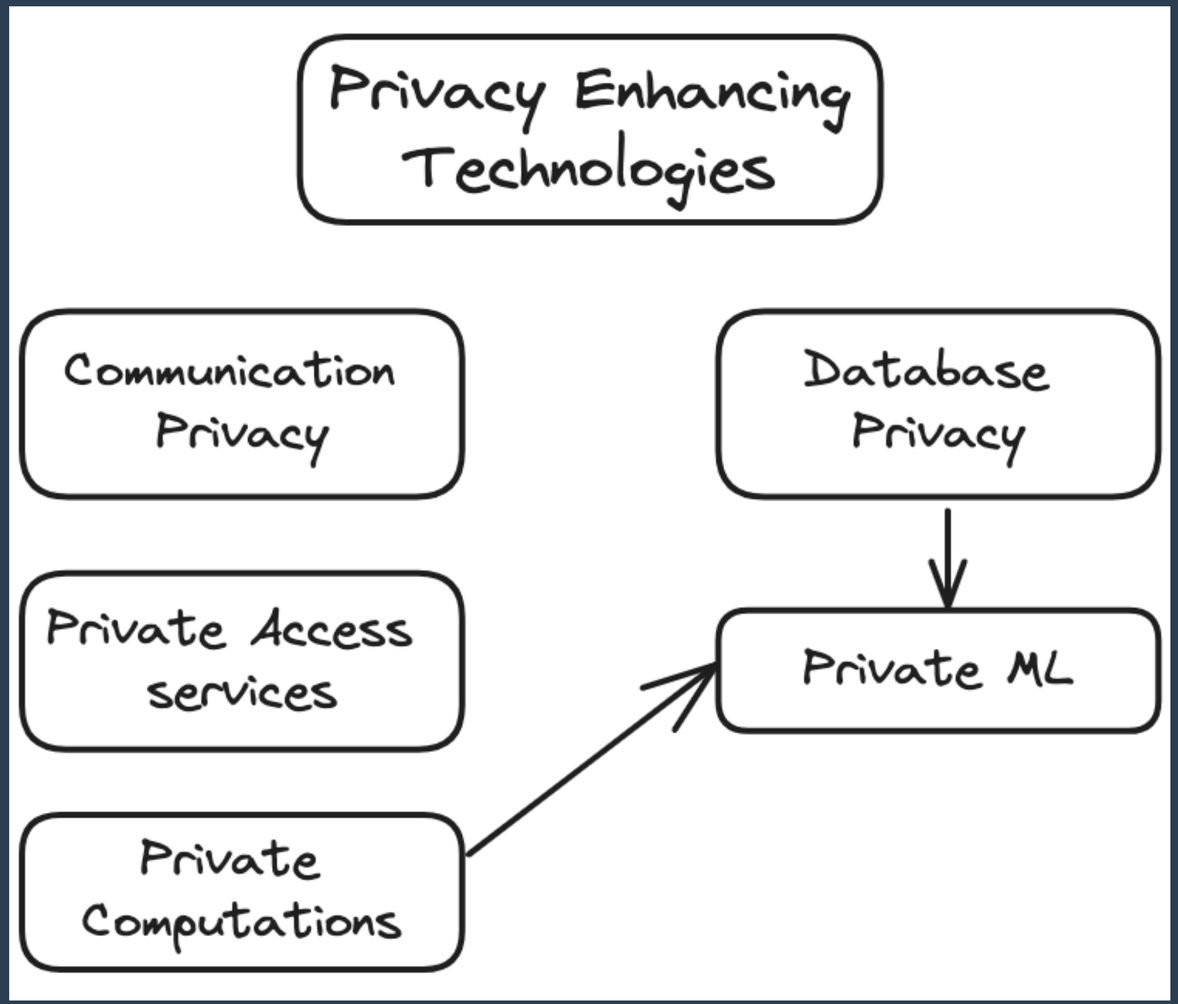
- Statistical Query: How many persons with a cold?, the answers from a differentially private computation will "nearly" be the same whether or not David is in the underlying database.
- Observation:
- The two databases where one contains David's data and the other do not contain his data – database neighbors. Generally speaking, any two databases D and D', which differ by at most one record but otherwise contain the same records are called database neighbors.
- The results of the query over D and D' doesn't look the same, what it means here is that the probability distributions of the query result are the same. So, the likelihood of getting answer 1 when database is D is the same likelihood for getting answer 1 from D'.

DIFFERENTIAL PRIVACY FORMAL DEFINITION

- Differential Privacy [Dwork06]:
  - A randomized query mechanism M_Q for query Q provides ε-differential privacy if
    - if for all databases D and D', where D and D' are database neighbors and
    - every subset O of the set of all possible outputs of M_Q,
- We have that: $Pr[M_Q(D) \ in \ O] \leq e^{\varepsilon} \cdot Pr[M_Q(D') \ in \ O]$

DIFFERENTIAL PRIVACY FORMAL DEFINITION CONT'D

- Observation:
- **Epsilon** is the measure of peak that stand out in the plausibility plot (is the measure of information gain in adversaries ability to confidently choose one combination of data over the other), and the above definition bounds the loss of privacy from releasing the query results.
- **Composition** The future releases also guarantee ε-differential privacy
  - if we publish the count of persons with cold with $\varepsilon = 3$ and publish the average age of persons with $\varepsilon = 3$, then the total privacy loss caused from the release of the two statistics is at most 6.



Sign Post of Day I and Day II topics