CS5562: Trustworthy Machine Learning

Part II Lecture 2: Inference Attacks

Reza Shokri[a]

Aug 2023

[a]Acknowledgment. The wonderful teaching assistants: Hongyan Chang, Martin Strobel, Jiashu Tao, Yao Tong, Jiayuan Ye

Contents

Setting: Inference Attacks

- Let D be a private dataset consisting of records from multiple people
- An algorithm runs on D, and releases useful information to the adversary
  - sanitized data
  - summary statistics
  - machine learning models
- Information leakage: how much could an adversary infer about the secret information (of sensitive data), given the released outputs?

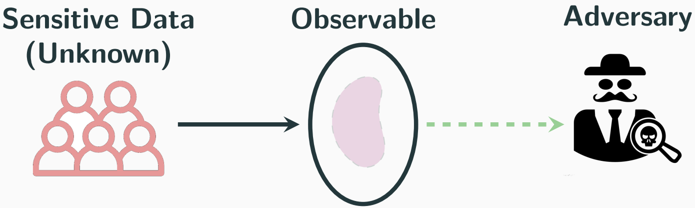
Setting: Inference Attacks

- Let D be a private dataset consisting of records from multiple people
- An algorithm runs on D, and releases useful information to the adversary
    - sanitized data
    - summary statistics
    - machine learning models
- Information leakage: how much could an adversary infer about the secret information (of sensitive data), given the released outputs?

Setting: Inference Attacks

- Let D be a private dataset consisting of records from multiple people
- An algorithm runs on D, and releases useful information to the adversary
  - sanitized data
  - summary statistics
  - machine learning models
- Information leakage: how much could an adversary infer about the secret information (of sensitive data), given the released outputs?

Inference Attack



Could the adversary infer certain secret of the private dataset D?

- (Membership Inference) Does D contain a given person's record?
- (Reconstruction) What are the values of possible data records in D?

Membership Inference Attacks

Why do we care about membership inference attack?

Membership may reveal highly sensitive private information

Example: Genome-Wide Association Studies

- Given two groups of participants: people with a given (sensitive) disease (e.g. HIV) and similar people without the disease.
- Certain information about these two groups used to be released by organizations, e.g. the US national Institute of Health (NIH).
- Membership inference: does a given individual belong to the first group $\Rightarrow$ does this individual have the disease $\Rightarrow$ privacy violation

Source: [Homer et al., 2008]

Why do we care about membership inference attack?

Membership may reveal highly sensitive private information

Example: Genome-Wide Association Studies

- Given two groups of participants: people with a given (sensitive) disease (e.g. HIV) and similar people without the disease.
- Certain information about these two groups used to be released by organizations, e.g. the US national Institute of Health (NIH).
- Membership inference: does a given individual belong to the first group $\Rightarrow$ does this individual have the disease $\Rightarrow$ privacy violation

Source: [Homer et al., 2008]

Why do we care about membership inference attack?

Membership may reveal highly sensitive private information

Example: Genome-Wide Association Studies

- Given two groups of participants: people with a given (sensitive) disease (e.g. HIV) and similar people without the disease.
- Certain information about these two groups used to be released by organizations, e.g. the US national Institute of Health (NIH).
- Membership inference: does a given individual belong to the first group $\Rightarrow$ does this individual have the disease $\Rightarrow$ privacy violation

Source: [Homer et al., 2008]

Why do we care about membership inference attack?

Membership information is useful for recovering more complicated information (reconstruction) of records

Example: next-word prediction

- A next-word prediction model is trained on text that may contain sensitive information (such as address and phone number)
- Using membership inference for reconstruction: given a sentence "Lebowski's social security number is ___", among all possible values for the blank (SSN), find the token that has the highest probability of being a member of the training dataset $\Rightarrow$ successful reconstruction!

Source: [Carlini et al., 2019, Carlini et al., 2021]

Why do we care about membership inference attack?

Membership information is useful for recovering more complicated information (reconstruction) of records

Example: next-word prediction

- A next-word prediction model is trained on text that may contain sensitive information (such as address and phone number)
- Using membership inference for reconstruction: given a sentence "Lebowski's social security number is ___", among all possible values for the blank (SSN), find the token that has the highest probability of being a member of the training dataset $\Rightarrow$ successful reconstruction!

Source: [Carlini et al., 2019, Carlini et al., 2021]

Why do we care about membership inference attack?

Membership information is useful for recovering more complicated information (reconstruction) of records

Example: next-word prediction

- A next-word prediction model is trained on text that may contain sensitive information (such as address and phone number)
- Using membership inference for reconstruction: given a sentence "Lebowski's social security number is ___", among all possible values for the blank (SSN), find the token that has the highest probability of being a member of the training dataset $\Rightarrow$ successful reconstruction!

Source: [Carlini et al., 2019, Carlini et al., 2021]

Why do we care about membership inference attack?

Membership inference attack is useful for quantifying the privacy risk in machine learning systems

- many AI regulations and Guidelines (e.g. those required by NIST, ICO and GDPR) consider membership inference as crucial risk
- "...membership inferences show that AI models can inadvertently contain personal data"
- "Attacks that reveal confidential information about the data include membership inference..."

Why do we care about membership inference attack?

Membership inference attack is useful for quantifying the privacy risk in machine learning systems

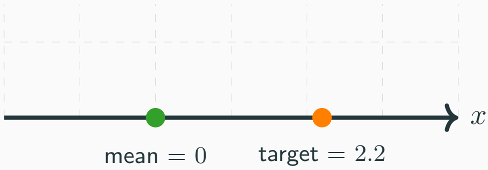
- many AI regulations and Guidelines (e.g. those required by NIST, ICO and GDPR) consider membership inference as crucial risk
- "...membership inferences show that AI models can inadvertently contain personal data"
- "Attacks that reveal confidential information about the data include membership inference..."

Setting: Membership Inference

- Membership Inference: after releasing P, if an adversary is given a target record z, could it infer whether $z \in D$ or not?

Example: membership inference given released statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- We are given the released mean of the records in D
- Membership Inference: is a target record $z = 2.2$ in the dataset?

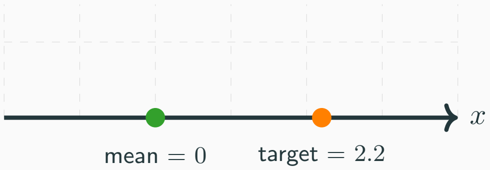


Intuitive reasoning: probably not, because 2.2 is far from the mean 0

However, from how far exactly should the adversary guess non-member?
$\Rightarrow$ we need some other data as comparison baseline (we need to know more about the underlying distribution)

Example: membership inference given released statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- We are given the released mean of the records in D
- Membership Inference: is a target record $z = 2.2$ in the dataset?

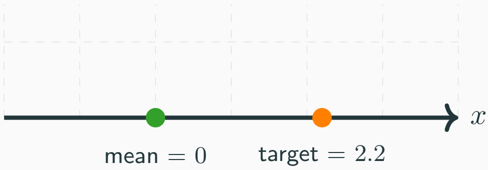


Intuitive reasoning: probably not, because 2.2 is far from the mean 0

However, from how far exactly should the adversary guess non-member?
$\Rightarrow$ we need some other data as comparison baseline (we need to know more about the underlying distribution)

Example: membership inference given released statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- We are given the released mean of the records in D
- Membership Inference: is a target record $z = 2.2$ in the dataset?



Intuitive reasoning: probably not, because 2.2 is far from the mean 0

However, from how far exactly should the adversary guess non-member?
$\Rightarrow$ we need some other data as comparison baseline (we need to know more about the underlying distribution)

Membership Inference Attack using Population Data

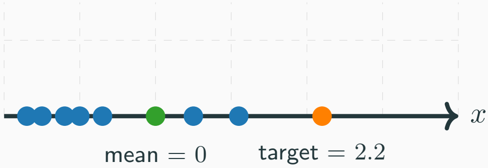
Use a large *pool* of (population) data points to launch the attack.

- Is the specific record z closer to the released statistics P than a randomly selected record x from the population is to P?

Source: [Homer et al., 2008]

Example: membership inference via population records

- A dataset D consists of unknown records drawn from 1D Gaussian.
- We are given the released mean of the records in D.
- Membership Inference: is the target record $z = 2.2$ in the dataset?

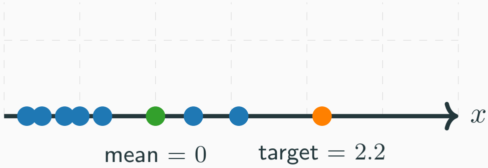


Population records: drawn from the population pool (the pool is large s.t. with high probability, the drawn records do not overlap with D)

Adversary: predicts non-member because the population distance $1.1 < 2.2$ (the distance between target and mean)

Example: membership inference via population records

- A dataset D consists of unknown records drawn from 1D Gaussian.
- We are given the released mean of the records in D.
- Membership Inference: is the target record $z = 2.2$ in the dataset?

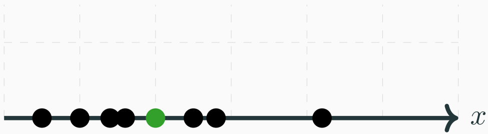


Population records: drawn from the population pool (the pool is large s.t. with high probability, the drawn records do not overlap with D)

Adversary: predicts non-member because the population distance $1.1 < 2.2$ (the distance between target and mean)

Is membership inference via population records always correct?

- Adversary might make errors in his inference attack.



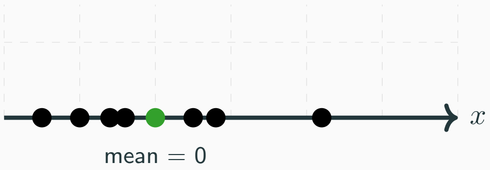
Average distance between population records and released mean: 1.1

Assume D **includes** $z = 2.2$ (black circles)

Adversary wrongly predicts the record $x = 2.2$ as non-member Because the distance between 2.2 and mean 0 is much larger than population distance $1.1 \Rightarrow$ fails on outlier members (e.g. at the tail)

Is membership inference via population records always correct?

- Adversary might make errors in his inference attack.



Average distance between population records and released mean: 1.1

Assume D **includes** $z = 2.2$ (black circles)

Adversary wrongly predicts the record $x = 2.2$ as non-member Because the distance between 2.2 and mean 0 is much larger than population distance 1.1 $\Rightarrow$ fails on outlier members (e.g. at the tail)

MIA via population records may not work on outlier members

- MIA via population records is essentially checking how well a target record z mixes with population records
- An outlier target record z (e.g. at the tail) is always far from the released statistic even when it is actually a member of the target dataset
    - MIA via population record may not work accurately, if the dataset size is large (as comparison to the dimension of the data).

$\Rightarrow$ We need a better comparison reference for "distance" that adapts to outliers

MIA via population records may not work on outlier members

- MIA via population records is essentially checking how well a target record z mixes with population records
- An outlier target record z (e.g. at the tail) is always far from the released statistic even when it is actually a member of the target dataset
  - MIA via population record may not work accurately, if the dataset size is large (as comparison to the dimension of the data).

$\Rightarrow$ We need a better comparison reference for "distance" that adapts to outliers

MIA via population records may not work on outlier members

- MIA via population records is essentially checking how well a target record z mixes with population records
- An outlier target record z (e.g. at the tail) is always far from the released statistic even when it is actually a member of the target dataset
  - MIA via population record may not work accurately, if the dataset size is large (as comparison to the dimension of the data).

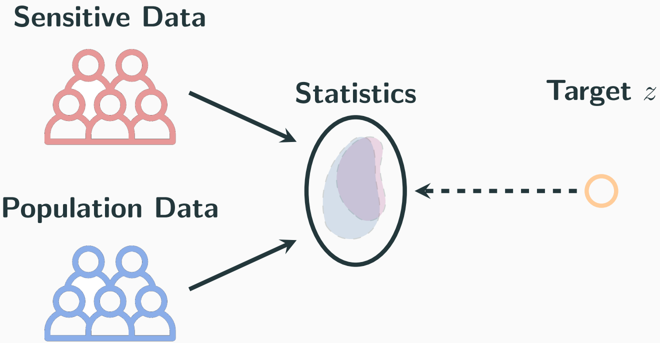
$\Rightarrow$ We need a better comparison reference for "distance" that adapts to outliers

MIA via population records may not work on outlier members

- MIA via population records is essentially checking how well a target record z mixes with population records
- An outlier target record z (e.g. at the tail) is always far from the released statistic even when it is actually a member of the target dataset
    - MIA via population record may not work accurately, if the dataset size is large (as comparison to the dimension of the data).

$\Rightarrow$ We need a better comparison reference for "distance" that adapts to outliers

Membership Inference Attack using Reference Statistics



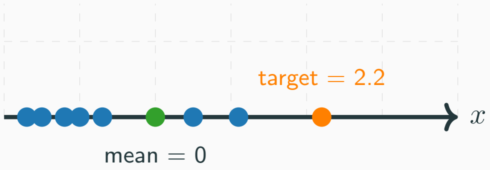
Compute reference statistics P' over population data that does not contain z.

- Is z more similar to the released statistics P or the reference statistics P'?

Source: [Sankararaman et al., 2009]

Example: membership inference via reference statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- Given released mean of the records.
- Membership Inference: is the target record $z = 2.2$ in the dataset?



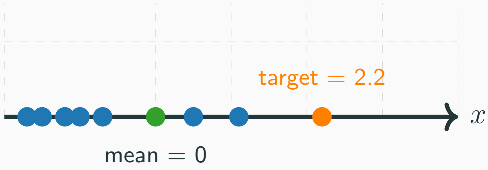
Population records: drawn from the population pool (the pool is large s.t. with high probability, the drawn records do not overlap with D)

Reference statistics: mean of population records is approximately -0.64

Adversary predicts member because the target 2.2 is closer to the released mean 0 than it is to the population mean -0.64

Example: membership inference via reference statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- Given released mean of the records.
- Membership Inference: is the target record $z = 2.2$ in the dataset?



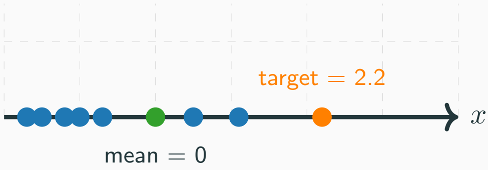
Population records: drawn from the population pool (the pool is large s.t. with high probability, the drawn records do not overlap with D)

Reference statistics: mean of population records is approximately -0.64

Adversary predicts member because the target 2.2 is closer to the released mean 0 than it is to the population mean -0.64

Example: membership inference via reference statistics

- A dataset D consists of unknown records drawn from 1D Gaussian.
- Given released mean of the records.
- Membership Inference: is the target record $z = 2.2$ in the dataset?



Population records: drawn from the population pool (the pool is large s.t. with high probability, the drawn records do not overlap with D)

Reference statistics: mean of population records is approximately -0.64

Adversary predicts member because the target 2.2 is closer to the released mean 0 than it is to the population mean -0.64

Extending reference statistics to reference ML models

- Given loss $\ell(x;\theta)$ of a model θ on a data record x.
- The training objective of ML model is $\arg\min_\theta \sum_{x \in D} loss(x;\theta)$
- **Similarity between record and released model:** we quantify the similarity with $Pr[x|\theta] \propto e^{-loss(x;\theta)}$, that is, the higher the loss value, the lower the similarity

Membership Inference:

After releasing θ, if the adversary knows z, how to infer whether $z \in D$ via reference models?

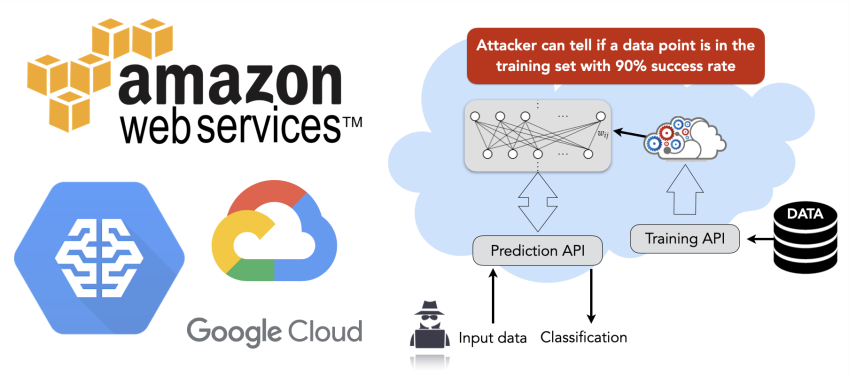
Extending reference statistics to reference ML models

After releasing the target model θ, if the adversary knows z, could it infer whether $z \in D$?

- Obtain "reference models" $\hat{\theta}$ that are trained on other records (that do not contain z), e.g.,
  - data drawn from population pool (excluding z)
- Attack via reference models
  If $Pr[x; \theta] \gg Pr[x; \hat{\theta}]$, output "member"

Source: [Shokri et al., 2017, Ye et al., 2022]

Real World Attacks against Machine Learning Models



Source: [Shokri et al., 2017]

Reconstruction attacks

Reconstruction attacks on released summary statistics

Given a private table with only 2 attributes for each record: race and age.

	Race	Age
Person 1	$\cdots$	$\cdots$

The following summary statistics about the table are released.

- There are 4 people who are Asian.
- The average age of people who are Asian is 20.
- The number of Asian people with age 18 is 3.

We can reconstruct some data records

(Asian, 26), (Asian, 18), (Asian, 18), (Asian, 18)

Reconstruction attacks on released summary statistics

Given a private table with only 2 attributes for each record: race and age.

	Race	Age
Person 1	$\cdots$	$\cdots$

The following summary statistics about the table are released.

- There are 4 people who are Asian.
- The average age of people who are Asian is 20.
- The number of Asian people with age 18 is 3.

We can reconstruct some data records

(Asian, 26), (Asian, 18), (Asian, 18), (Asian, 18)

Reconstruction attacks on released summary statistics

Given a private table with only 2 attributes for each record: race and age.

	Race	Age
Person 1	$\cdots$	$\cdots$

The following summary statistics about the table are released.

- There are 4 people who are Asian.
- The average age of people who are Asian is 20.
- The number of Asian people with age 18 is 3.

We can reconstruct some data records

(Asian, 26), (Asian, 18), (Asian, 18), (Asian, 18)

Formulating reconstruction under released summary statistics

- Given a dataset D with n records.
- Each record has one sensitive binary attribute.
- Given m queries asking the sum of records in different subsets $S_1, \cdots, S_m \subseteq D$ (specified by different predicates or conditions).
- Let the m released summary statistics be $a_1, \cdots, a_m$. That is,

$$\begin{cases} \sum_{x \in S_1} x = a_1 \\ \dots \\ \sum_{x \in S_m} x = a_m \end{cases}$$

Goal: obtain the solution $\{x : x \in D\} \Rightarrow$ Dataset reconstruction!

How to solve: With large m, the above linear equation system may become overdetermined, thus admitting a (unique) solution

Formulating reconstruction under released summary statistics

- Given a dataset D with n records.
- Each record has one sensitive binary attribute.
- Given m queries asking the sum of records in different subsets $S_1, \cdots, S_m \subseteq D$ (specified by different predicates or conditions).
- Let the m released summary statistics be $a_1, \cdots, a_m$. That is,

$$\begin{cases} \sum_{x \in S_1} x = a_1 \\ \dots \\ \sum_{x \in S_m} x = a_m \end{cases}$$

Goal: obtain the solution $\{x : x \in D\}$ $\Rightarrow$ Dataset reconstruction!

How to solve: With large m, the above linear equation system may become overdetermined, thus admitting a (unique) solution

Reconstruction attacks on released (noisy) statistics

If the statistics are only approximate (e.g. with noise), would reconstruction attack be difficult?

Example
The adversary asks each query with repeatedly a number of times.

- How many people are Asian? 3,5,4,6,3,...
- The average age of people who are Asian is 19.4, 20.2, 20.6, 19.8, 19.5...
- The number of Asian people with age 18 is 2, 2, 4, 3, 4...

Still able to reconstruct, but need to ask more queries!

Explanation: the averages of responses to repeated queries converge to their mean $(4, 20, 3)$, which enables reconstructing (Asian, 26).

Reconstruction attacks on released (noisy) statistics

If the statistics are only approximate (e.g. with noise), would reconstruction attack be difficult?

Example
The adversary asks each query with repeatedly a number of times.

- How many people are Asian? 3,5,4,6,3,...
- The average age of people who are Asian is 19.4, 20.2, 20.6, 19.8, 19.5...
- The number of Asian people with age 18 is 2, 2, 4, 3, 4...

Still able to reconstruct, but need to ask more queries!

Explanation: the averages of responses to repeated queries converge to their mean $(4, 20, 3)$, which enables reconstructing (Asian, 26).

Reconstruction attacks on released (noisy) statistics

If the statistics are only approximate (e.g. with noise), would reconstruction attack be difficult?

Example
The adversary asks each query with repeatedly a number of times.

- How many people are Asian? 3,5,4,6,3,...
- The average age of people who are Asian is 19.4, 20.2, 20.6, 19.8, 19.5...
- The number of Asian people with age 18 is 2, 2, 4, 3, 4...

Still able to reconstruct, but need to ask more queries!

Explanation: the averages of responses to repeated queries converge to their mean $(4, 20, 3)$, which enables reconstructing (Asian, 26).

Formulating reconstruction under (noisy) summary statistics

- Given private dataset where each record has one binary attribute
- Given m queries asking the sum of records in different subsets $S_1, \cdots, S_m \subseteq D$.
- Given m released noisy summary statistics $a_1, \cdots, a_m$ where $a_i = \sum_{x \in S_i} x + \mathcal{N}(0, \Delta_i)$ with Gaussian noise
- With high probability

$$
\begin{cases}
a_1 - 3\Delta_1 \leq \sum_{x \in S_1} x \leq a_1 + 3\Delta_1 \\
\dots \\
a_m - 3\Delta_m \leq \sum_{x \in S_m} x \leq a_m + 3\Delta_m
\end{cases}
$$

For large m, small Δ_i, this linear program is (partially) solvable with small error with high probability $\Rightarrow$ accurate dataset reconstruction!

Formulating reconstruction under (noisy) summary statistics

- Given private dataset where each record has one binary attribute
- Given m queries asking the sum of records in different subsets $S_1, \cdots, S_m \subseteq D$.
- Given m released noisy summary statistics $a_1, \cdots, a_m$ where $a_i = \sum_{x \in S_i} x + \mathcal{N}(0, \Delta_i)$ with Gaussian noise
- With high probability

$$\begin{cases} a_1 - 3\Delta_1 \leq \sum_{x \in S_1} x \leq a_1 + 3\Delta_1 \\ \dots \\ a_m - 3\Delta_m \leq \sum_{x \in S_m} x \leq a_m + 3\Delta_m \end{cases}$$

For large m, small Δ_i, this linear program is (partially) solvable with small error with high probability $\Rightarrow$ accurate dataset reconstruction!

Formulating reconstruction under (noisy) summary statistics

See [Dinur and Nissim, 2003] for more discussions regarding:

- How many noisy queries m could we answer while preserving privacy (in terms of preventing accurate reconstruction up to error e under noise Δ)?
- What is the relationship between noise scale Δ and the maximal number of queries m that could be answered?
- ...

Source: [Dinur and Nissim, 2003]

Reconstruction Attack in Practice

- The attack is on a production system Diffix that answers statistical queries
- Diffix adds noise to the query response before releasing
- Example query: how many clients have loan status ‘C’
  SELECT count(*) FROM loans
  WHERE status = ‘C’ AND client-id BETWEEN 2000 and 3000

Source: [Cohen and Nissim, 2018]

Reconstruction Attack in Practice

- Attack goal: find out the loan status associated with each client ID
- Adversary's query:
  - SELECT count(clientId) FROM loans
    WHERE floor(100 * ((clientId * 2)∧0.7) + 0.5)
    = floor(100 * ((clientId * 2)∧0.7))
    AND clientId BETWEEN 2000 and 3000
    AND loanStatus = 'C'
- Adversary could vary the highlighted terms to construct subsets of client IDs that have high granularity (e.g., each subset only contains a few IDs)

Source: [Cohen and Nissim, 2018]

Summary: Reconstruction Attacks

- Released statistics impose hard (or soft) constraints on the underlying data.
- Revealing "too many" statistics "too accurately" allows an adversary to solve the constraints and reconstruct data.
- For machine learning problems: the constraints may be more complicated and could be non-linear $\Rightarrow$ solve with other optimization algorithms besides linear programming ([Zhu et al., 2019])

What is privacy under inference attacks?

Setting

- Consider $x_1, x_2, \cdots, x_n$ sensitive data points, where x_i belongs to individual i
- An analyst is interested in running some computation on x_is
- **Absolute Information Leakage:** whether the computation leaks secret information about sensitive data through inference attacks

Setting

- Consider $x_1, x_2, \cdots, x_n$ sensitive data points, where x_i belongs to individual i
- An analyst is interested in running some computation on x_is
- **Absolute Information Leakage:** whether the computation leaks secret information about sensitive data through inference attacks

Setting

- Consider $x_1, x_2, \cdots, x_n$ sensitive data points, where x_i belongs to individual i
- An analyst is interested in running some computation on x_is
- **Absolute Information Leakage:** whether the computation leaks secret information about sensitive data through inference attacks

Classical Intuition for Privacy

- Dalenius (1977): "If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S, a disclosure has taken place"
  - Privacy means that anything that can be learned about a respondent from the statistical database can be learned without access to this particular database

Classical Intuition for Privacy

- Dalenius (1977): "If the release of statistics S makes it possible to determine the value [of private information] more accurately than is possible without access to S, a disclosure has taken place"
  - Privacy means that anything that can be learned about a respondent from the statistical database can be learned without access to this <u>particular</u> database

Problems with Interpreting Classical Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
- What if my (incorrect) prior is that every person has three arms?
- How much is "too much?"
  - Is it possible to achieve cryptographically small levels of disclosure and keep the data useful?
  - By useful, we mean analyst should learn useful population-level trend about the data

Absolute disclosure guarantees (with regard to the prior) are unachievable! (Prior may be wrong.)

Problems with Interpreting Classical Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
- What if my (incorrect) prior is that every person has three arms?
- How much is "too much?"
    - Is it possible to achieve cryptographically small levels of disclosure and keep the data useful?
    - By useful, we mean analyst should learn useful population-level trend about the data

Absolute disclosure guarantees (with regard to the prior) are unachievable! (Prior may be wrong.)

Problems with Interpreting Classical Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
- What if my (incorrect) prior is that every person has three arms?
- How much is "too much?"
  - Is it possible to achieve cryptographically small levels of disclosure and keep the data useful?
  - By useful, we mean analyst should learn useful population-level trend about the data

Absolute disclosure guarantees (with regard to the prior) are unachievable! (Prior may be wrong.)

Problems with Interpreting Classical Intuition

- Popular interpretation: prior and posterior views about an individual shouldn't change "too much"
- What if my (incorrect) prior is that every person has three arms?
- How much is "too much?"
    - Is it possible to achieve cryptographically small levels of disclosure and keep the data useful?
    - By useful, we mean analyst should learn useful population-level trend about the data

Absolute disclosure guarantees (with regard to the prior) are unachievable! (Prior may be wrong.)

Absolute Guarantees are problematic

- Absolute guarantees are problematic
  - Your privacy can be "breached" (per absolute definition of privacy) even if your data is not in the database

Example

- I know that you are 2 inches taller than the average Russian
- Database allows computing average height of a Russian
- This database breaks your privacy according to this definition... even if your record is not in the database!

Absolute Guarantees are problematic

- Absolute guarantees are problematic
  - Your privacy can be “breached” (per absolute definition of privacy) even if your data is not in the database

Example

- I know that you are 2 inches taller than the average Russian
- Database allows computing average height of a Russian
- This database breaks your privacy according to this definition... even if your record is not in the database!

What is privacy under inference attacks

- Absolute guarantees are problematic
    - Your privacy can be "breached" (per absolute definition of privacy) even if your data is not in the database

Relative privacy guarantee
Whatever is learned would be learned regardless of your participation

- Dual: Whatever is already known, situation won't get worse
- Analyst is supposed to learn useful population-level information.

Example: Smoking causes cancer

- A hospital maintains an access mechanism for its patient database
- The hospital wants to answer useful statistical queries, e.g., correlations between symptoms.
- However, the hospital should also preserve patients' privacy.

Now, the hospital reveals a correlation that "smoking causes cancer," Bob (who never went to this hospital but has cancer), complains that now people could infer that he smokes. Is this a violation for Bob's privacy?

- No. Because this correlation applies to the whole population rather than a specific dataset.

Example: Smoking causes cancer

- A hospital maintains an access mechanism for its patient database
- The hospital wants to answer useful statistical queries, e.g., correlations between symptoms.
- However, the hospital should also preserve patients' privacy.

Now, the hospital reveals a correlation that "smoking causes cancer," Bob (who never went to this hospital but has cancer), complains that now people could infer that he smokes. Is this a violation for Bob's privacy?

- No. Because this correlation applies to the whole population rather than a specific dataset.

Example: Smoking causes cancer

- A hospital maintains an access mechanism for its patient database
- The hospital wants to answer useful statistical queries, e.g., correlations between symptoms.
- However, the hospital should also preserve patients' privacy.

Now, the hospital reveals a correlation that "smoking causes cancer," Bob (who never went to this hospital but has cancer), complains that now people could infer that he smokes. Is this a violation for Bob's privacy?

- No. Because this correlation applies to the whole population rather than a specific dataset.

Relative privacy risk under inference attacks

- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
  - observes obs
  - knows the mechanism and has some background knowledge about D,
  - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} \cdot \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Relative privacy risk under inference attacks

- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
  - observes obs
  - knows the mechanism and has some background knowledge about D,
  - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} \cdot \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Relative privacy risk under inference attacks

- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
    - observes obs
    - knows the mechanism and has some background knowledge about D,
    - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} \cdot \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Relative privacy risk under inference attacks

- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
  - observes obs
  - knows the mechanism and has some background knowledge about D,
  - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} \cdot \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Relative privacy risk under inference attacks

- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
  - observes obs
  - knows the mechanism and has some background knowledge about D,
  - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} \cdot \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Relative privacy risk under inference attacks

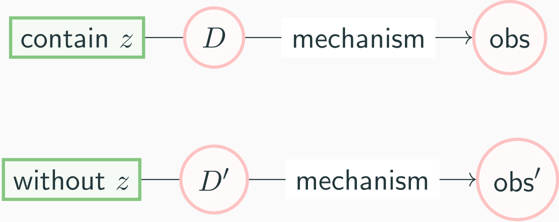
- D (sensitive data) $\longrightarrow$ mechanism $\longrightarrow$ obs (observables)
- Attacker
  - observes obs
  - knows the mechanism and has some background knowledge about D,
  - wants to infer (learn) secret information about D

$$\Pr[D|\text{obs}] = \frac{\overbrace{\Pr[\text{obs}|D]}^{\text{mechanism}} . \overbrace{\Pr[D]}^{\text{knowledge}}}{\Pr[\text{obs}|D]}$$

- **Violation of Privacy for $z \in D$:** how much the observer learns about z's secret after seeing obs on D, that could not be learned had the dataset been another dataset D' (without z)?

Quantifying relative privacy risk with indistinguishability

Two datasets that differ in the secret (information of a record) :
Neighboring datasets D and D' that differ in one sensitive record z

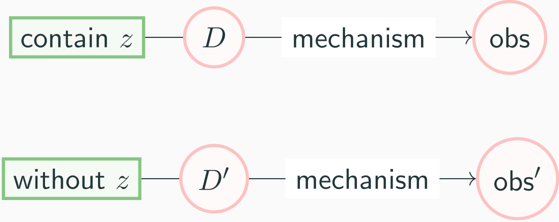


Goal of privacy
The distribution of inferred secret $\Pr(\text{secret}|\text{obs})$ and $\Pr(\text{secret}|\text{obs}')$ should be indistinguishable!

$\rightarrow$ we need to reason about this "indistinguishability" to quantify privacy

Two datasets that differ in the secret (information of a record) :

Neighboring datasets D and D' that differ in one sensitive record z



Goal of privacy

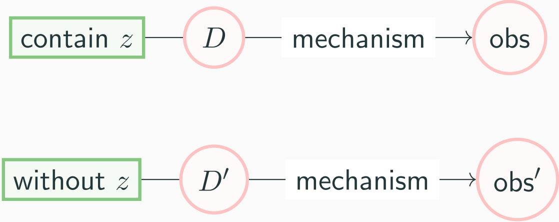
The distribution of inferred secret $\Pr(\text{secret}|\text{obs})$ and $\Pr(\text{secret}|\text{obs}')$ should be indistinguishable!

→ we need to reason about this “indistinguishability” to quantify privacy

Quantifying relative privacy risk with indistinguishability

Two datasets that differ in the secret (information of a record) :

Neighboring datasets D and D' that differ in one sensitive record z



Goal of privacy

The distribution of inferred secret $\Pr(\text{secret}|\text{obs})$ and $\Pr(\text{secret}|\text{obs}')$ should be indistinguishable!

$\rightarrow$ we need to reason about this "indistinguishability" to quantify privacy

Next Lecture: Quantitative reasoning about data privacy

- Meaningful evaluation metrics for “indistinguishability”
- Understand what risk is possible (stronger inference attacks)
- Understanding how to mitigate certain risks (differential privacy upper bound)

Reading Assignments

- Read the survey of attacks [Dwork et al., 2017]
- Read Chapters 1 and 2 of [Dwork and Roth, 2014]

References i

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. (2019). **The secret sharer: Evaluating and testing unintended memorization in neural networks.** In 28th USENIX Security Symposium (USENIX Security 19), pages 267–284.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). **Extracting training data from large language models.** In 30th USENIX Security Symposium (USENIX Security 21), pages 2633–2650.

- Cohen, A. and Nissim, K. (2018).
  Linear program reconstruction in practice.
  arXiv preprint arXiv:1810.05692.
- Dinur, I. and Nissim, K. (2003).
  Revealing information while preserving privacy.
  In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 202–210.
- Dwork, C. and Roth, A. (2014).
  The algorithmic foundations of differential privacy.
  Foundations and Trends in Theoretical Computer Science, 9(3–4):211–407.

References iii

Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017).
Exposed! a survey of attacks on private data.
Annual Review of Statistics and Its Application, 4:61–84.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008).
Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays.
PLoS genetics, 4(8):e1000167.

- Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. (2009).
Genomic privacy and limits of individual detection in a pool.
Nature genetics, 41(9):965–967.

- Shokri, R., Stronati, M., Song, C., and Shmatikov, V. (2017).
Membership inference attacks against machine learning models.
In 2017 IEEE symposium on security and privacy (SP), pages 3–18. IEEE.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. (2022).
Enhanced membership inference attacks against machine learning models.
In ACM Conference on Computer and Communications Security (CCS).

Zhu, L., Liu, Z., and Han, S. (2019).
Deep leakage from gradients.
Advances in Neural Information Processing Systems, 32.