Data Privacy

Vitaly Shmatikov
Public Data Conundrum

◆ Health-care datasets
  - Clinical studies, hospital discharge databases ...

◆ Genetic datasets
  - $1000 genome, HapMap, DeCODE ...

◆ Demographic datasets
  - U.S. Census Bureau, sociology studies ...

◆ Search logs, recommender systems, social networks, blogs ...
  - AOL search data, online social networks, Netflix movie ratings, Amazon ...
Basic Setting

Users (government, researchers, marketers, ...)

query 1  
answer 1

query T
answer T

DB=

$x_1$
$x_2$
$x_3$
...
$x_{n-1}$
$x_n$

San

random coins
Examples of Sanitization Methods

- **Input perturbation**
  - Add random noise to database, release

- **Summary statistics**
  - Means, variances
  - Marginal totals
  - Regression coefficients

- **Output perturbation**
  - Summary statistics with noise

- **Interactive versions of the above methods**
  - Auditor decides which queries are OK, type of noise
Data “Anonymization”

◆ How?

◆ Remove “personally identifying information” (PII)
  - Name, Social Security number, phone number, email, address... what else?

◆ Problem: PII has no technical meaning
  - Defined in disclosure notification laws
    - If certain information is lost, consumer must be notified
  - In privacy breaches, any information can be personally identifying
    - Examples: AOL dataset, Netflix Prize dataset
Latanya Sweeney’s Attack (1997)

Massachusetts hospital discharge dataset

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>City</th>
<th>Date Of Birth</th>
<th>Sex</th>
<th>ZIP</th>
<th>Marital Status</th>
<th>Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Public voter dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>ZIP</th>
<th>DOB</th>
<th>Sex</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sue J. Carlson</td>
<td>1459 Main St.</td>
<td>Cambridge</td>
<td>02142</td>
<td>9/15/61</td>
<td>female</td>
<td>democrat</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Observation #1: Dataset Joins

◆ Attacker learns sensitive data by joining two datasets on common attributes
  
  • Anonymized dataset with sensitive attributes
    – Example: age, race, symptoms
  
  • “Harmless” dataset with individual identifiers
    – Example: name, address, age, race

◆ Demographic attributes (age, ZIP code, race, etc.) are very common in datasets with information about individuals
Observation #2: Quasi-Identifiers

- Sweeney’s observation:
  (birthdate, ZIP code, gender) uniquely identifies 87% of US population
  - Side note: actually, only 63% [Golle, WPES ‘06]
- Publishing a record with a quasi-identifier is as bad as publishing it with an explicit identity
- Eliminating quasi-identifiers is not desirable
  - For example, users of the dataset may want to study distribution of diseases by age and ZIP code
k-Anonymity

- Proposed by Samarati and/or Sweeney (1998)
- Hundreds of papers since then
  - Extremely popular in the database and data mining communities (SIGMOD, ICDE, KDD, VLDB)
- NP-hard in general, but there are many practically efficient k-anonymization algorithms
- Most based on generalization and suppression
Anonymization in a Nutshell

- Dataset is a relational table
- Attributes (columns) are divided into quasi-identifiers and sensitive attributes

<table>
<thead>
<tr>
<th>Race</th>
<th>Age</th>
<th>Symptoms</th>
<th>Blood type</th>
<th>Medical history</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Generalize/suppress quasi-identifiers, don’t touch sensitive attributes (keep them “truthful”)
k-Anonymity: Definition

◆ Any (transformed) quasi-identifier must appear in at least k records in the anonymized dataset
  • k is chosen by the data owner (how?)
  • Example: any age-race combination from original DB must appear at least 10 times in anonymized DB

◆ Guarantees that any join on quasi-identifiers with the anonymized dataset will contain at least k records for each quasi-identifier
Two (and a Half) Interpretations

- **Membership disclosure**: Attacker cannot tell that a given person in the dataset
- **Sensitive attribute disclosure**: Attacker cannot tell that a given person has a certain sensitive attribute
- **Identity disclosure**: Attacker cannot tell which record corresponds to a given person

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers.

But this does not imply any privacy!

Example: k clinical records, all HIV+
Achieving k-Anonymity

◆ Generalization
  • Replace specific quasi-identifiers with more general values until get k identical values
    – Example: area code instead of phone number
  • Partition ordered-value domains into intervals

◆ Suppression
  • When generalization causes too much information loss
    – This is common with “outliers” (come back to this later)

◆ Lots of algorithms in the literature
  • Aim to produce “useful” anonymizations
    ... usually without any clear notion of utility
Generalization in Action

\[ Z_2 = \{02100\} \]
\[ Z_1 = \{02130, 02140\} \]
\[ Z_0 = \{02138, 20239, 02141, 02142\} \]

\[ \text{DGH}_{z_0} \]

\[ \text{VGH}_{z_0} \]

\[ M_2 = \{\text{not\_released}\} \]
\[ M_1 = \{\text{once\_married, never\_married}\} \]
\[ M_0 = \{\text{married, divorced, widow, single}\} \]

\[ \text{DGH}_{n_0} \]

\[ \text{VGH}_{n_0} \]
Curse of Dimensionality

◆ Generalization fundamentally relies on spatial locality
  • Each record must have k close neighbors

◆ Real-world datasets are very sparse
  • Many attributes (dimensions)
    – Netflix Prize dataset: 17,000 dimensions
    – Amazon customer records: several million dimensions
  • “Nearest neighbor” is very far

◆ Projection to low dimensions loses all info ⇒ k-anonymized datasets are useless

[Aggarwal VLDB '05]
k-Anonymity: Definition

◆ Any (transformed) quasi-identifier must appear in at least $k$ records in the anonymized dataset
  - $k$ is chosen by the data owner (how?)
  - Example: any age-race combination from original DB must appear at least 10 times in anonymized DB
◆ Guarantees that any join on quasi-identifiers with the anonymized dataset will contain at least $k$ records for each quasi-identifier

This definition does not mention sensitive attributes at all!

Assumes that attacker will be able to join **only** on quasi-identifiers

Does not say anything about the computations that are to be done on the data
Membership Disclosure

- With large probability, quasi-identifier is unique in the population
- But generalizing/suppressing quasi-identifiers in the dataset does not affect their distribution in the population (obviously)!
  - Suppose anonymized dataset contains 10 records with a certain quasi-identifier …
  … and there are 10 people in the population who match this quasi-identifier
- k-anonymity may not hide whether a given person is in the dataset
Sensitive Attribute Disclosure

Intuitive reasoning:

- k-anonymity prevents attacker from telling which record corresponds to which person
- Therefore, attacker cannot tell that a certain person has a particular value of a sensitive attribute

This reasoning is fallacious!
## 3-Anonymization

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>ID</th>
<th>Diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucas</td>
<td>78712</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian</td>
<td>78705</td>
<td>Shingles</td>
</tr>
<tr>
<td>Caucas</td>
<td>78754</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian</td>
<td>78705</td>
<td>Acne</td>
</tr>
<tr>
<td>AfrAm</td>
<td>78705</td>
<td>Acne</td>
</tr>
<tr>
<td>Caucas</td>
<td>78705</td>
<td>Flu</td>
</tr>
</tbody>
</table>

This is 3-anonymous, right?
Joining With External Database

Problem: sensitive attributes are not “diverse” within each quasi-identifier group
Another Attempt: I-Diversity

[Machanavajjhala et al. ICDE ’06]

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Code</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Shingles</td>
</tr>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Acne</td>
</tr>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Acne</td>
</tr>
<tr>
<td>Caucasian</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Acne</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Shingles</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Acne</td>
</tr>
<tr>
<td>Asian/African</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
</tbody>
</table>

Entropy of sensitive attributes within each quasi-identifier group must be at least L
Still Does Not Work

Original database

| ... | Cancer |
| ... | Cancer |
| ... | Cancer |
| ... | Flu |
| ... | Cancer |
| ... | Cancer |
| ... | Cancer |
| ... | Cancer |
| ... | Flu |

99% have cancer

99% cancer ⇒ quasi-identifier group is not “diverse”
...yet anonymized database does not leak anything

Anonymization A

| Q1 | Flu |
| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |
| Q2 | Cancer |

This leaks a ton of information

Anonymization B

| Q1 | Flu |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q1 | Cancer |
| Q2 | Cancer |

50% cancer ⇒ quasi-identifier group is “diverse”
Try Again: t-Closeness

Distribution of sensitive attributes within each quasi-identifier group should be “close” to their distribution in the entire original database.

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Identification</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Shingles</td>
</tr>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Acne</td>
</tr>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Acne</td>
</tr>
<tr>
<td>Caucas</td>
<td>787XX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Acne</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Shingles</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Acne</td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>78XXX</td>
<td>Flu</td>
</tr>
</tbody>
</table>

[Li et al. ICDE ‘07]

Trick question: Why publish quasi-identifiers at all??
**Anonymized “t-Close” Database**

<table>
<thead>
<tr>
<th></th>
<th>787XX</th>
<th>HIV+</th>
<th>Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucas</td>
<td>HIV-</td>
<td>Flu</td>
<td></td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>HIV+</td>
<td>Shingles</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>HIV-</td>
<td>Acne</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>HIV-</td>
<td>Shingles</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>HIV-</td>
<td>Acne</td>
<td></td>
</tr>
</tbody>
</table>
What Does Attacker Know?

Bob is white and I heard he was admitted to hospital with flu...

<table>
<thead>
<tr>
<th></th>
<th>787XX</th>
<th>HIV+</th>
<th>Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucas</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian/AfrAm</td>
<td>787XX</td>
<td>HIV-</td>
<td>Flu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV+</td>
<td>Shingles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV-</td>
<td>Acne</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV-</td>
<td>Shingles</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>HIV-</td>
<td>Acne</td>
</tr>
</tbody>
</table>

This is against the rules! “flu” is not a quasi-identifier

Yes... and this is yet another problem with k-anonymity
Issues with Syntactic Definitions

◆ What adversary do they apply to?
  • Do not consider adversaries with side information
  • Do not consider probability
  • Do not consider adversarial algorithms for making decisions (inference)

◆ Any attribute is a potential quasi-identifier
  • External / auxiliary / background information about people is very easy to obtain
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 the database

Similar to semantic security of encryption

- Anything about the plaintext that can be learned from a ciphertext can be learned without the ciphertext
Problems with Classic Intuition

- Popular interpretation: prior and posterior views about an individual shouldn’t change “too much”
  - What if my (incorrect) prior is that every Cornell graduate student has three arms?

- How much is “too much?”
  - Can’t achieve cryptographically small levels of disclosure and keep the data useful
  - Adversarial user is supposed to learn unpredictable things about the database
Absolute Guarantee Unachievable

Privacy: for some definition of “privacy breach,”
∀ distribution on databases, ∀ adversaries A, ∃ A’
such that $\Pr(A(San)=\text{breach}) - \Pr(A’()=\text{breach}) \leq \varepsilon$

- For reasonable “breach”, if San(DB) contains information about DB, then some adversary breaks this definition

Example

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

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

Relative guarantee: “Whatever is learned would be learned regardless of whether or not you participate”
  • Dual: Whatever is already known, situation won’t get worse
Indistinguishability

Differ in 1 row

DB =

x₁
x₂
x₃
…
xₙ₋₁
xₙ

random coins

San

transcript
S

Distance between distributions is at most $\varepsilon$

DB' =

y₁
y₂
y₃
…
yₙ₋₁
yₙ

random coins

San

transcript
S'

Distance between distributions is at most $\varepsilon$

query 1
answer 1

query T
answer T

query 1
answer 1

query T
answer T
Which Distance to Use?

Problem: $\varepsilon$ must be large
- Any two databases induce transcripts at distance $\leq n\varepsilon$
- To get utility, need $\varepsilon > 1/n$

Statistical difference $1/n$ is not meaningful!
- Example: release a random point from the database
  - $San(x_1,...,x_n) = (j, x_j)$ for random $j$
- For every $i$, changing $x_i$ induces statistical difference $1/n$
- But some $x_i$ is revealed with probability 1
  - Definition is satisfied, but privacy is broken!
Formalizing Indistinguishability

Definition: San is $\varepsilon$-indistinguishable if

$$\forall A, \forall DB, DB' \text{ which differ in 1 row}, \forall \text{ sets of transcripts } S$$

$$p(\text{San}(DB) \in S) \in (1 \pm \varepsilon) p(\text{San}(DB') \in S)$$

Equivalently, $\forall S$:

$$\frac{p(\text{San}(DB) = S)}{p(\text{San}(DB') = S)} \in 1 \pm \varepsilon$$
Laplacian Mechanism

- Intuition: $f(x)$ can be released accurately when $f$ is insensitive to individual entries $x_1, \ldots, x_n$

- Global sensitivity $\text{GS}_f = \max_{\text{neighbors } x, x'} \| f(x) - f(x') \|_1$
  - Example: $\text{GS}_{\text{average}} = 1/n$ for sets of bits

- Theorem: $f(x) + \text{Lap}(\text{GS}_f/\varepsilon)$ is $\varepsilon$-indistinguishable
  - Noise generated from Laplace distribution
Sensitivity with Laplace Noise

**Theorem**

If $A(x) = f(x) + \text{Lap}\left(\frac{\text{GS}_f}{\varepsilon}\right)$ then $A$ is $\varepsilon$-indistinguishable.

Laplace distribution $\text{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$

![Laplace distribution graph]

Sliding property of $\text{Lap}\left(\frac{\text{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\text{GS}_f}}$ for all $y, \delta$

*Proof idea:*

- $A(x)$: blue curve
- $A(x')$: red curve
- $\delta = f(x) - f(x') \leq \text{GS}_f$
Differential Privacy: Summary

San gives $\varepsilon$-differential privacy if for all values of DB and Me and all transcripts $t$:

$$\frac{\Pr[ San (DB - Me) = t]}{\Pr[ San (DB + Me) = t]} \leq e^\varepsilon \approx 1 \pm \varepsilon$$