数据隐私方法伦理和实践

Methodology, Ethics and Practice of Data Privacy

5. 攻击

Attacks

张兰
中国科学技术大学 计算机学院

2020春季

Attacking Sanitized Data

- In some cases, however, an attacker can still glean sensitive information from the sanitized data using varying amounts of skill, creativity, and effort.

- Means available to an attacker — these include the attacker’s statistical skills, external knowledge, and availability of computational resources.

Attacking Sanitized Data

- Attack type
  - Verifying information about specific individuals (including verifying their presence in the data)
  - Discovering information about individuals that are outliers in the sanitized data
  - Making a claim that a breach is possible
  - Partially reconstructing the original data

- An attack against a sanitized data set can be considered a success if an attacker, through reasonable means, believes that he or she has discovered sensitive information.

1 Direct Attacks

[1] D. Kifer, “Attacks on privacy and de Finetti’s theorem,” in SIGMOD, 2009.
[2] K. Liu, C. Giannella, and H. Kargupta, *A Survey of Attack Techniques on Privacy-preserving Data Perturbation Methods*. chapter 15, pp. 357–380,Springer, 2008.

Combinatorial Methods

- It aims at the vulnerability of certain query answering systems.
  - If an attacker knows some unique characteristics of a target individual, then by posing a few carefully chosen queries and using linear algebra, additional characteristics of the target individual can also be determined.
  - E.g., “what is the total salary of employees under 30” and “What is the total salary of male employees under 30?” to determine Shirley’s salary.

Combinatorial Methods

- It aims at the vulnerability of certain query answering systems.
  - How to compute combinatorial upper and lower bounds on the number of tuples in the table with particular attributes.

[1] A. Dobra, “statistical tools for disclosure limitation in multiway contingency tables,” PhD thesis, Carnegie Mellon University, 2002.

[2] S. E. Fienberg and A. B. Slavkovic, “Preserving the confidentiality of categorical statistical data bases when releasing information for association rules,” *Data Mining Knowledge Discovery*, vol. 11, no. 2, pp. 155–180, 2005.

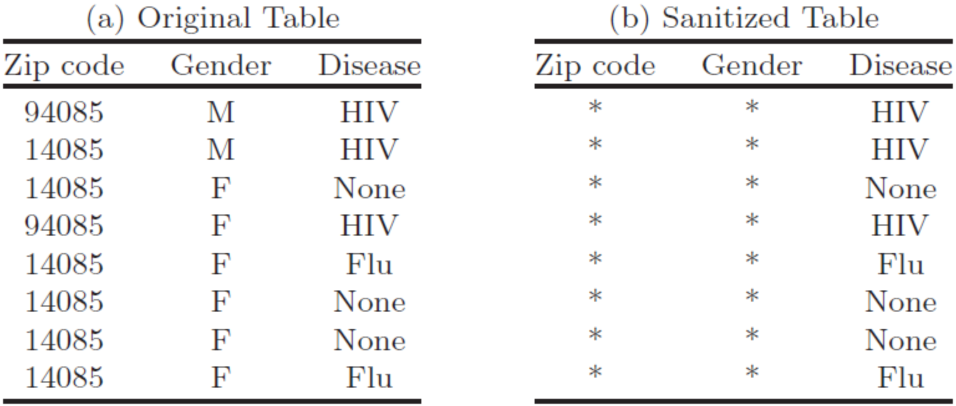
Optimality Attacks

- Observation: usually data anonymization is framed as a constrained optimization problem: produce the table with the smallest distortion that also satisfies a given set of privacy requirements.
- An attacker will usually need to know the non-sensitive information of many individuals in the table, the privacy policy, and the algorithm used for anonymization.

- How you sanitize the data → features of the original data

Optimality Attacks

- Suppose the following privacy policy is desired: for each published combination of gender and zip code, at most half the corresponding patients have HIV.
- Anonymization algorithm: suppress a set of attributes



(a) Original Table

Zip code	Gender	Disease
94085	M	HIV
14085	M	HIV
14085	F	None
94085	F	HIV
14085	F	Flu
14085	F	None
14085	F	None
14085	F	Flu

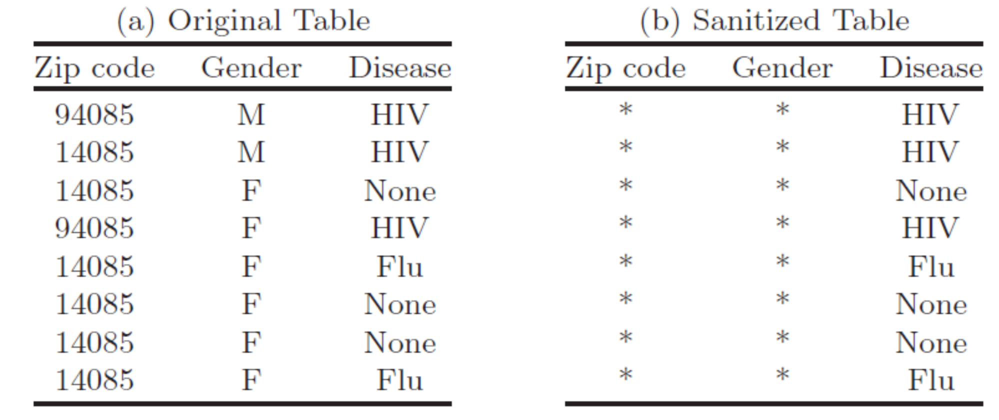
(b) Sanitized Table

Zip code	Gender	Disease
*	*	HIV
*	*	HIV
*	*	None
*	*	HIV
*	*	Flu
*	*	None
*	*	None
*	*	Flu

Optimality Attacks

» An attacker who knows the non-sensitive attributes of every individual in the table sees (b):

- if all the HIV patients were female, or if exactly one of the HIV patients were male, then the privacy requirement could have been achieved by suppressing only the zip code. Therefore both male patients must have HIV.
- if at most one of the patients from zip code 94085 had HIV, then the privacy requirement could have been satisfied by suppressing only gender. Thus both patients from zip code 94085 must have HIV.



(a) Original Table

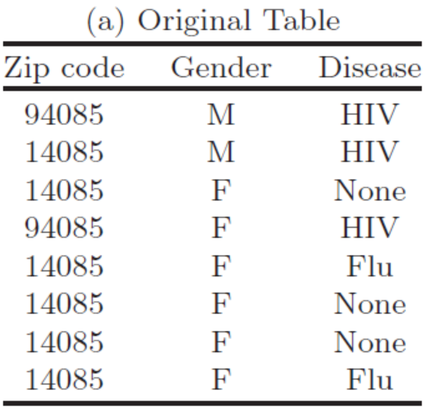
Zip code	Gender	Disease
94085	M	HIV
14085	M	HIV
14085	F	None
94085	F	HIV
14085	F	Flu
14085	F	None
14085	F	None
14085	F	Flu

(b) Sanitized Table

Zip code	Gender	Disease
*	*	HIV
*	*	HIV
*	*	None
*	*	HIV
*	*	Flu
*	*	None
*	*	None
*	*	Flu

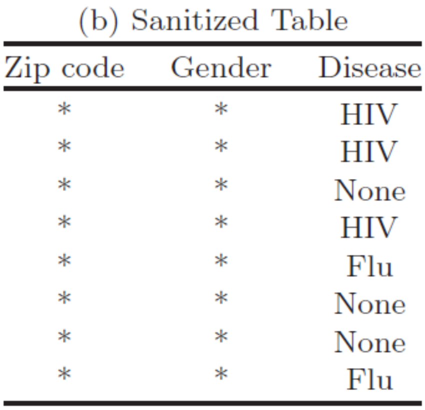
Optimality Attacks

- An attacker who knows the non-sensitive attributes of every individual in the table sees (b):
  - Sampling the original data makes the attack more difficult.



(a) Original Table

Zip code	Gender	Disease
94085	M	HIV
14085	M	HIV
14085	F	None
94085	F	HIV
14085	F	Flu
14085	F	None
14085	F	None
14085	F	Flu



(b) Sanitized Table

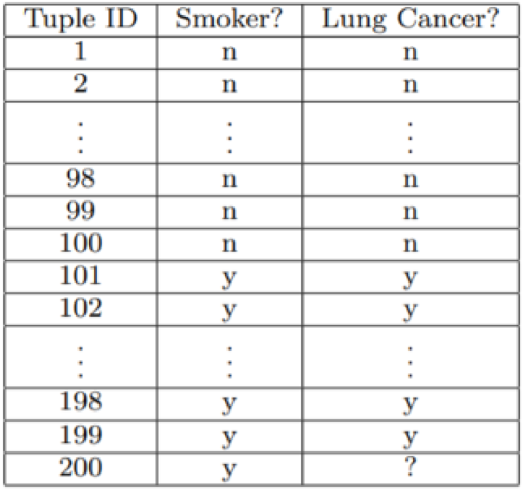
Zip code	Gender	Disease
*	*	HIV
*	*	HIV
*	*	None
*	*	HIV
*	*	Flu
*	*	None
*	*	None
*	*	Flu

Alternative Reasoning

- The sanitized data themselves can leak more information than the data publisher anticipated, even when an attacker does not have detailed knowledge of the sanitization algorithm.

Alternative Reasoning

- Modeling attackers

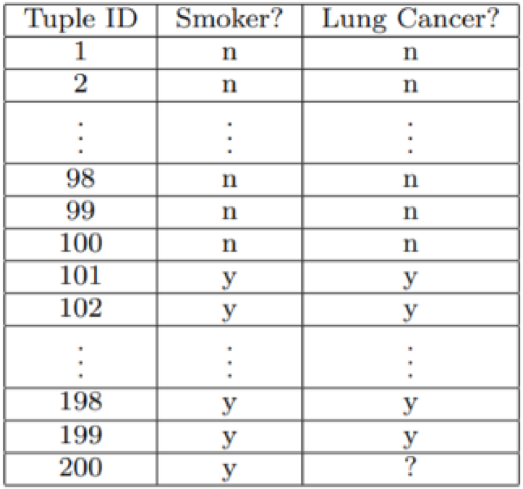


Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

$P^{200}(cancer) = ?$

Alternative Reasoning

- Modeling attackers
  - Random worlds



Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

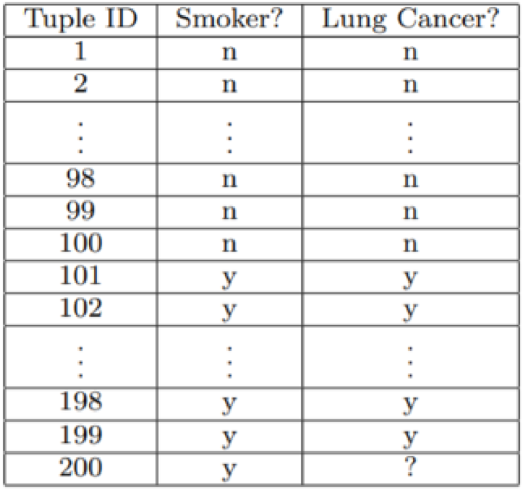
Start with the belief that every table of 200 tuples is equally likely.

$$P^{200}(cancer) = 0.5$$

The same probability we would have given before we had seen the Table.

Alternative Reasoning

- Modeling attackers
  - i.i.d. model

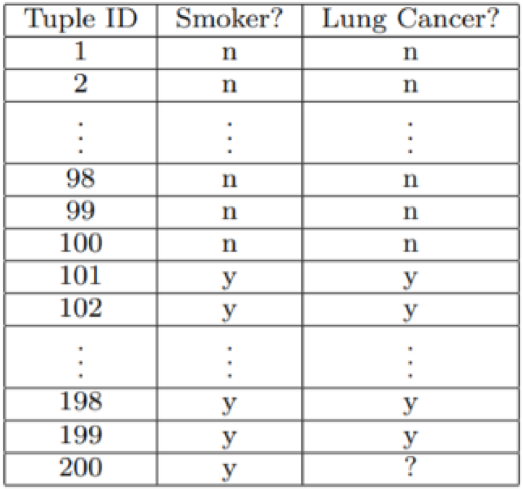


Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

Start with the belief t tuples are generated i.i.d. by the probability distribution P (known to the attacker).
Prior knowledge:
$P_1 = P(smoker \wedge lung\ cancer)$,
$P_2 = P(nonsmoker \wedge lung\ cancer)$
$P_3 = P(smoker \wedge no\ lung\ cancer)$
$P_4 = P(nonsmoker \wedge no\ lung\ cancer)$
$P^{200}(cancer) = P_1/(P_1+P_3)$
The same probability we would have given before we had seen the Table, if we know 200 was a smoker.

Alternative Reasoning

- Modeling attackers
  - Tuple-independent model



Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

Start with the belief each tuple t_i has independent probability p_i of appearing in a database instance.

Prior knowledge:

$P_1^x = P^x(smoker \wedge lung\ cancer),$

$P_2^x = P^x(nonsmoker \wedge lung\ cancer)$

$P_3^x = P^x(smoker \wedge no\ lung\ cancer)$

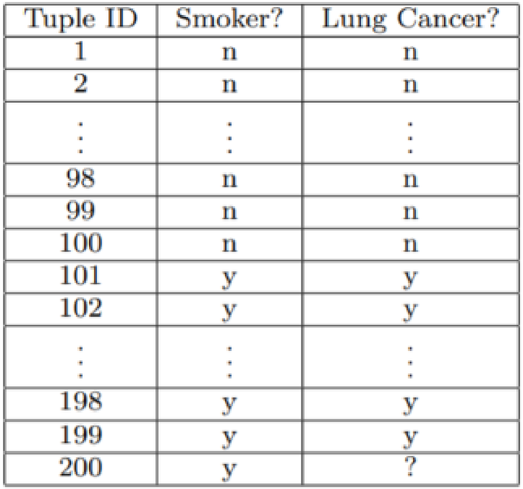
$P_4^x = P^x(nonsmoker \wedge no\ lung\ cancer)$

$P^{200}(cancer)=P_1^{200}/(P_1^{200}+P_3^{200})$

The same probability we would have given before we had seen the Table, if we know 200 was a smoker.

Alternative Reasoning

- Modeling attackers
    - Random worlds, i.i.d. model, Tuple-independent model



Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

Even though there appears to be a strong correlation between smoking and lung cancer. None of these three model accounted for it.

The table did nothing to change our beliefs.

Alternative Reasoning

- Modeling attackers
  - The error in reasoning (for random worlds, the i.i.d. model, and the tuple-independent model) is all three models assume that the tuples are independent of each other and that we believe they are generated by a particular distribution P.
  - Exchangeability: A sequence X1, X2, . . . of random variables is exchangeable if every finite permutation of these random variables has the same distribution.

The flips of a coin are i.i.d.
The probability of seeing HHHT T is the same as the probability of seeing THTHH.

Every i.i.d. sequence of random variables is Exchangeable.

Alternative Reasoning

Modeling attackers

- A game: there are two coins, a host randomly selects one and flip it to generate a sequence.

P(head)=1

P(tail)=1

- H?????
- T?????
- From the first coin flip, we learn more about the coin and thus we are able to better predict the second coin flip.
- Exchangeable but non-iid.

Alternative Reasoning

Modeling attackers

- A game: there are infinite number of possible coins, a host randomly selects one and flip it to generate the data.

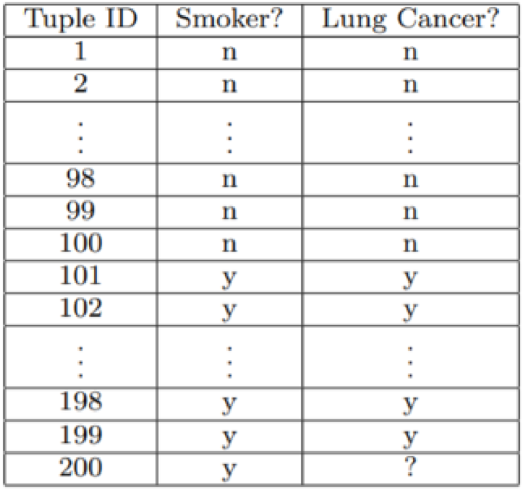
P(head)=p1 ……

P(head)=pn

- The generation of data from an exchangeable sequence of random variables as a two-step process:
  - select the parameters of a probability distribution (such as the bias of a coin) from a prior probability over parameters.
  - Use these parameters, we generate the data (such as coin flips). Based on the data, we can use Bayesian reasoning to compute the posterior distribution of the parameters.

Alternative Reasoning

- Modeling attackers



Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

$p_1 = P(smoker \ \wedge \ lung\ cancer),$
$p_2 = P(nonsmoker \ \wedge \ lung\ cancer)$
$p_3 = P(smoker \ \wedge \ no\ lung\ cancer)$
$p_4 = P(nonsmoker \ \wedge \ no\ lung\ cancer)$

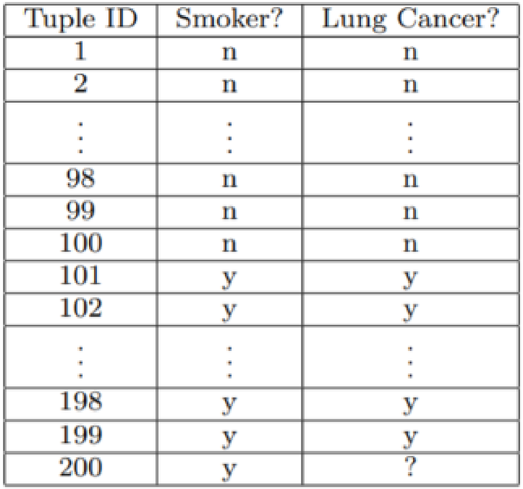
Treat them as unknown parameters with a uniform prior. Any choice of $(p1, p2, p3, p4)$ for $p1 + p1 + p2 + p4 = 1$ is equally likely.

The probability of generating Table 1 :

$$\frac{1}{3!}\int_{\substack{0\le p_1\le 1,0\le p_2\le 1\\ 0\le p_3\le 1,0\le p_4\le 1\\ p_1+p_2+p_3+p_4=1}} p_1^{100}p_2^{0}p_3^{0}p_4^{100}\,dp_1\ dp_2\ dp_3\ dp_4$$
$$+\frac{1}{3!}\int_{\substack{0\le p_1\le 1,0\le p_2\le 1\\ 0\le p_3\le 1,0\le p_4\le 1\\ p_1+p_2+p_3+p_4=1}} p_1^{99}p_2^{0}p_3^{1}p_4^{100}\,dp_1\ dp_2\ dp_3\ dp_4$$

Alternative Reasoning

- Modeling attackers



Tuple ID	Smoker?	Lung Cancer?
1	n	n
2	n	n
⋮	⋮	⋮
98	n	n
99	n	n
100	n	n
101	y	y
102	y	y
⋮	⋮	⋮
198	y	y
199	y	y
200	y	?

The probability of generating Table 1 :

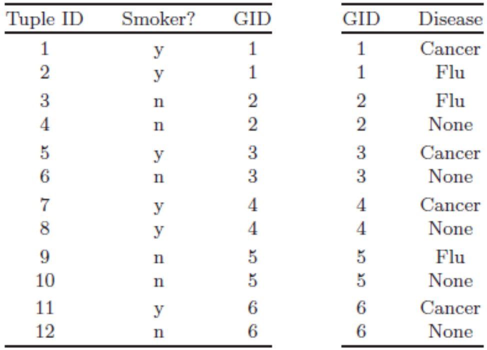
$$\frac{1}{3!}\frac{100!100!}{203!} + \frac{1}{3!}\frac{99!100!}{203!}$$

$P^{200}(cancer)$=

$$\frac{\frac{1}{3!}\frac{100!100!}{203!}}{\frac{1}{3!}\frac{100!100!}{203!} + \frac{1}{3!}\frac{99!100!}{203!}} = \frac{100}{101}$$

Alternative Reasoning

- Exchangeability can be used to attack bucketization schemes
  - no group consisting entirely of nonsmokers has cancer as one of the diseases. Tuple 12, which is a nonsmoker, belongs to a group that has exactly one cancer patient. Since the other tuple belongs to a smoker, we could conclude that tuple 12 is less likely to have cancer.



Tuple ID	Smoker?	GID
1	y	1
2	y	1
3	n	2
4	n	2
5	y	3
6	n	3
7	y	4
8	y	4
9	n	5
10	n	5
11	y	6
12	n	6

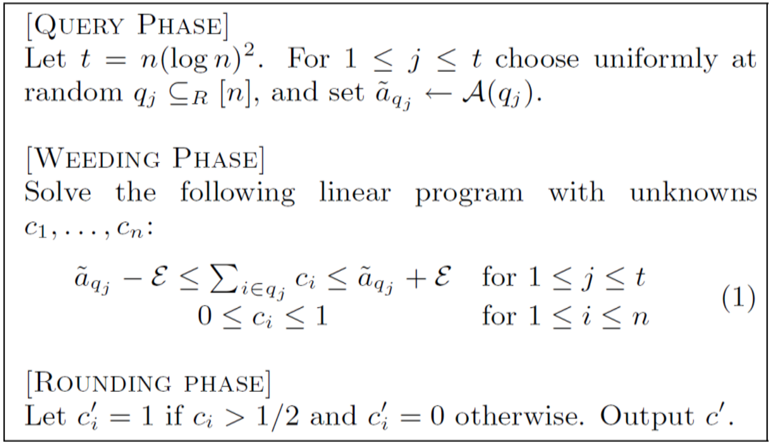
GID	Disease
1	Cancer
1	Flu
2	Flu
2	None
3	Cancer
3	None
4	Cancer
4	None
5	Flu
5	None
6	Cancer
6	None

Alternative Reasoning

- Modeling attackers
  - The random worlds model, the i.i.d. model, and the tuple-independent have been considered to be reasonable. However, they do not sufficiently protect privacy.
  - Reasoning using exchangeability provided better inference of sensitive attributes.
  - An attacker does not have to correctly guess the true value x.S in order to cause harm to x. For instance, suppose the attacker decides that x.S =AIDS with probability 0.9.
  - Sensitive information may also be an aggregate property of a subset of the data. E.g., the average salary of a company.

Denoising

- Attacks have been used against schemes that add noise
  - Many methods to remove noise or reidentification [2]
  - **Linear Programming**



[QUERY PHASE]
Let $t = n(\log n)^2$. For $1 \leq j \leq t$ choose uniformly at random $q_j \subseteq_R [n]$, and set $\tilde{a}_{q_j} \leftarrow \mathcal{A}(q_j)$.

[WEEDING PHASE]
Solve the following linear program with unknowns $c_1, \ldots, c_n$:

$$\begin{array}{rl} \tilde{a}_{q_j} - \mathcal{E} \leq \sum_{i \in q_j} c_i \leq \tilde{a}_{q_j} + \mathcal{E} & \text{for } 1 \leq j \leq t \\ 0 \leq c_i \leq 1 & \text{for } 1 \leq i \leq n \end{array} \qquad (1)$$

[ROUNDING PHASE]
Let $c'_i = 1$ if $c_i > 1/2$ and $c'_i = 0$ otherwise. Output c'.

  - Query: q ⊆ $[n]$, $\sum_{i\in q} d_i$. $n(logn)^2$ $queries$, the output candidate c' satisfies $dist(c, c') < \varepsilon n = o(n)$. A is within $o(\sqrt{n})$ perturbation, D is $\exp(n)$-non-private.

Denoising

- Attacks have been used against schemes that add noise
  - Many methods to remove noise or reidentification [2]
  - **Linear Programming**

If weighted query: $q \subseteq [n],\ \sum_{i \in q} a_i d_i$, where the weights are generated according to the standard Gaussian distribution. With O(n) queries, even if pn (for p < 0.239) queries are arbitrarily wrong and noise bounded by α is added to the rest of the answers, the output candidate c' satisfies $dist(c, c') < O(\alpha)$.

  - These results say that the amount of statistically meaningful information in a data set is sub-linear in the size of the data.
  - For time series data, **linear filters and linear regression** (on points known to an attacker) to remove some of the variance caused by the addition of noise.

Undesired Uses of Data

- Building a particular model over the data may be considered a violation of privacy
    - E.g., an employee may build a statistical model on the data to predict wages
    - Palley and Simonoff demonstrated how to build a linear regression model from a database that only allowed count, average, and sum-of-squares queries over subsets of the data.
    - 1) build a 1D histogram on each attribute
    - 2) use these histograms to identify regions of the domain that should be queried, to construct artificial data sets that would give the same answers to such queries
    - 3) create a linear regression model from such data sets.

2. Attacks using External Information

Linking Attack

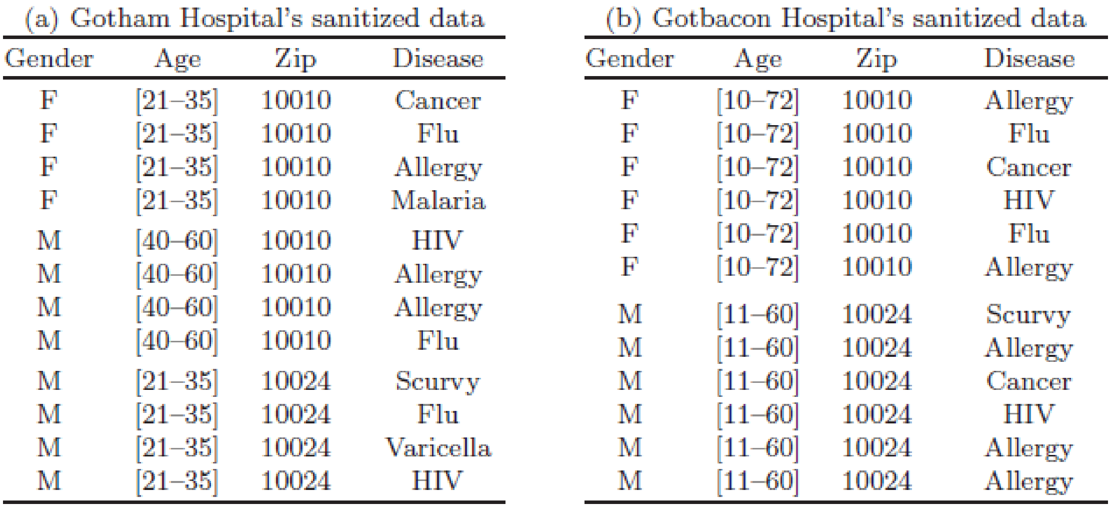
- The most common form of attack, the re-identification attack, uses record linkage techniques to link tuples in external data to sanitized data.
  - Given two files A and B containing lists of tuples, classify pairs (a, b) (where a ∈ A and b ∈ B) as match or non-match with various levels of uncertainty. B is sanitized.
  - Bayesian methods, discriminant analysis, bipartite matching, and nearest neighbor methods.
  - It may also be possible to link free text, images, and videos.

Composition

- A privacy breach can occur when the external data themselves are also sanitized.
- This attack is possible when several data publishers own data sets that are not disjoint (e.g., two hospitals in the same city).

Composition

- Bob just finished his Master’s degree, is living in zip code 10024



(a) Gotham Hospital’s sanitized data

Gender	Age	Zip	Disease
F	[21–35]	10010	Cancer
F	[21–35]	10010	Flu
F	[21–35]	10010	Allergy
F	[21–35]	10010	Malaria
M	[40–60]	10010	HIV
M	[40–60]	10010	Allergy
M	[40–60]	10010	Allergy
M	[40–60]	10010	Flu
M	[21–35]	10024	Scurvy
M	[21–35]	10024	Flu
M	[21–35]	10024	Varicella
M	[21–35]	10024	HIV

(b) Gotbacon Hospital’s sanitized data

Gender	Age	Zip	Disease
F	[10–72]	10010	Allergy
F	[10–72]	10010	Flu
F	[10–72]	10010	Cancer
F	[10–72]	10010	HIV
F	[10–72]	10010	Flu
F	[10–72]	10010	Allergy
M	[11–60]	10024	Scurvy
M	[11–60]	10024	Allergy
M	[11–60]	10024	Cancer
M	[11–60]	10024	HIV
M	[11–60]	10024	Allergy
M	[11–60]	10024	Allergy

Attacks using similar data

- it is possible to attack one sanitized data set with the help of a second data set even if they are disjoint
  - A company sells items $I_1, \ldots, I_m$and maintains a database which is a list $\{T_1, \ldots, T_n\}$ of transactions.
  - The company applies a perfect hash function (i.e., there are no collisions) to each item in each transaction in the data set and publicly releases the result.
  - For each item $I_1, \ldots, I_m$, the attacker has a belief function which gives upper and lower bounds on the frequency of an item in the company’s data set.
  - The goal of the attacker is to match each item to its hashed value.

Attacks using similar data

- it is possible to attack one sanitized data set with the help of a second data set even if they are disjoint
  - It is possible to use co-occurrence information, e.g., user identity and IP address of websites' logs.
  - Search query log: frequency and co-occurrence information from the un-encoded search logs can be used to recover many of the hashed tokens from the hashed data set.
  - Social network: if an attacker participates in the social network, then in many cases it is possible for the attacker to identify nodes corresponding to accounts under his control.

Instance Level Background Knowledge

- it is often important to consider the role of instance-level background knowledge.
  - There are many ideas based on logic
  - It is also possible to incorporate statistical knowledge. The statistical knowledge comes in the form of linear constraints and linear inequalities on probabilities. The inference of an attacker is modeled using maximum entropy.
  - To simulate and quantify the amount of knowledge an attacker may have, some works also use association rules mined from the original data.

THANKS!

Any questions?

You can find me at:

- » zhanglan@ustc.edu.cn