Data Mining and Analysis

Data Privacy

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Resources

Li Xiong Data Privacy and Security course @Emory

http://www.mathcs.emory.edu/~lxiong/cs573_f16/schedule.html

Privacy-preserving data publishing: A survey of recent developments; k-anonymity (Sweeney)

A firm foundation for private data analysis, CACM, 2011; Designing statistical privacy for your data, CACM, 2015

Differential Privacy in the Wild. SIGMOD 2017 Tutorial. Machanavajjhala, Hay, and He.

Differential Privacy

Person 1
r_1
Person 2
r_2
Person 3
r_3
...
Person N
r_N
Hospital
DB
Census
DB
Google
DB
Doctors
Medical Researchers
Economists
Recommen-dation Algorithms
Information Retrieval Researchers

Statistical Data Privacy

Non-interactive vs interactive

Privacy goal: individual is protected

Utility goal: statistical information useful for analysis

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Ideas?

Anonymization or de-identification (input perturbation)

- Linkage attacks

Query auditing/restriction

- Query denial is itself disclosive, computationally infeasible

Summary statistics

- Differencing attacks

Query Auditing / Restriction

- Offline auditing
  - Given a collection of queries and answers to them, check whether anything “forbidden” was revealed
  - Detects privacy breaches after the fact
- Online auditing
  - Queries are presented to auditor one at a time; auditor checks if answering the current query (in combination with past answers) reveals “forbidden” information
  - Prevents privacy breaches on-the-fly

Query Auditing Example: Sum/Max

Variables d_i are real

Privacy breached if adversary learns some d_i

$\text{sum}(d_1,d_2,d_3)?$ →

← Answer=15

$\text{max}(d_1,d_2,d_3)?$ →

← Denied!

Summary Statistics: Differencing Attack

How many people have flu?

→

←

Answer=100

How many people
not named John Doe have flu?

→

←

Answer = 99

Differential Privacy

Promise: an individual will not be affected, adversely or otherwise, by allowing his/her data to be used in any study or analysis, no matter what other studies, datasets, or information sources, are available

Paradox: learning nothing about an individual while learning useful statistical information about a population

Differential Privacy

Statistical outcome is indistinguishable regardless of whether a particular user (record) is included in the data

An Example

Name	Age	HIV+
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…

Name	Age	HIV+
Alice	43	Y
Frank	42	Y
Bob	31	Y
Mary	28	Y
Dave	43	N
…	…	…

Original records

Original histogram

Perturbed histogram with differential privacy

Differential Privacy

[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …

D_1 D_2

O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \quad (\varepsilon > 0)$$

… that differ in one row

For every pair of inputs that differ in one row

For every output …

D_1 D_2

O

Simulate the presence or absence of a single record

… all pairs of datasets

For every pair of inputs that differ in one row

For every output …

Guarantee holds no matter what the other records are.

… all outputs

Should not be able to distinguish whether input was D_1 or D_2 no matter what the output

Privacy Parameters

For every pair of inputs that differ in one row

For every output …

D_1 D_2

O

$$\Pr[A(D_1) = O] \leq e^{\varepsilon} \Pr[A(D_2) = O]$$

Controls the degree to which D_1 and D_2 can be distinguished.
Smaller the ε more the privacy (and better the utility)

$$\frac{\Pr[\mathcal{A}(D_1) \in Y]}{\Pr[\mathcal{A}(D_2) \in Y]} \leq \exp(\epsilon)$$

$$\rightarrow$$

$$\exp(-\epsilon) \leq \frac{\Pr[\mathcal{A}(D_1) \in Y]}{\Pr[\mathcal{A}(D_2) \in Y]} \leq \exp(\epsilon)$$

Since $\exp(\epsilon) \approx 1 + \epsilon$ *for small* ϵ,

$$1 - \epsilon \lesssim \frac{\Pr[\mathcal{A}(D_1) \in Y]}{\Pr[\mathcal{A}(D_2) \in Y]} \lesssim 1 + \epsilon$$

Differential privacy: Some qualitative properties

Protection against presence/ participation of a single record

Quantification of privacy loss

Composition

Post-processing

Output Randomization

Add noise to answers such that:

- Each answer does not leak too much information about the database
- Noisy answers are close to the original answers

Laplace Mechanism

Laplace distribution

PDF: $$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

Denoted as Lap(b) when u=0

Mean u

Variance $2b^2$

How much noise for privacy?

Sensitivity: Consider a query q: *I* → R. S(q) is the smallest number s.t. for any neighboring tables D, D',

$$| q(D) - q(D') | \leq S(q)$$

Theorem: If **sensitivity** of the query is **S**, then the algorithm A(D) = q(D) + Lap(S(q)/ε) guarantees ε-differential privacy

$$f(D) = \sum_{i=1}^{n} d_i, \text{ where each } d_i \in \{0, 1\}$$

Example: COUNT query

Number of people having disease

Sensitivity =

1

("differing in one row" —> could change count by max of 1)

If we use Lap(1/ε) then we have ε-differential privacy

$$\mathcal{A}(D) = \sum_{i=1}^{n} d_i + Lap(1/\epsilon)$$

Disease (Y/N)
Y
Y
N
Y
N
N

$$f(D) = \frac{1}{n}\sum_{i=1}^{n} d_i, \text{ where each } d_i \in [0, M]$$

Example: Average query

Sensitivity =

M/n

$$\mathcal{A}(D) = \frac{1}{n}\sum_{i=1}^{n} d_i + Lap(\frac{M}{n\epsilon})$$

Why Composition?

Reasoning about privacy of a complex algorithm is hard.

Helps software design

> If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.

A bound on the number of queries

In order to ensure utility, a statistical database must leak some information about each individual

We can only hope to bound the amount of disclosure

Hence, there is a limit on number of queries that can be answered

Composition Theorems

Sequential composition

$\sum_i \varepsilon_i$ –differential privacy

Parallel composition

$\max(\varepsilon_i)$–differential privacy

Sequential Composition

If M1, M2, ..., Mk are algorithms that access a private database D such that each Mi satisfies εi -differential privacy,

then the combination of their outputs satisfies ε-differential privacy with ε=ε1+...+εk

Parallel Composition

If M1, M2, ..., Mk are algorithms that access disjoint databases D1, D2, ..., Dk such that each Mi satisfies εi -differential privacy,

then the combination of their outputs satisfies ε-differential privacy with ε= max{ε1,...,εk}

Postprocessing

If M1 is an ε differentially private algorithm that accesses a private database D,

then outputting M2(M1(D)) also satisfies ε- differential privacy.

Case Study: kmeans

kmeans with Differential Privacy

- Suppose we fix the number of iterations to T
- In each iteration (given a set of centers):

  1. Assign the points to the new center to form clusters
  2. Noisily compute the size of each cluster
  3. Compute noisy sums of points in each cluster

kmeans with Differential Privacy

- Suppose we fix the number of iterations to T

Each iteration uses ε/T privacy budget, total privacy loss is ε

- In each iteration (given a set of centers):

1. Assign the points to the new center to form clusters

2. Noisily compute the size of each cluster

3. Compute noisy sums of points in each cluster

kmeans with Differential Privacy

Exercise: Which of these steps expends privacy budget?

- In each iteration (given a set of centers):

  1. Assign the points to the new center to form clusters

  2. Noisily compute the size of each cluster

  3. Compute noisy sums of points in each cluster