Privacy in Data Publishing†

CPS 116

Introduction to Database Systems

†*Most contents drawn from tutorial by Gehrke & Machanavajjhala at the 2009 IEEE Sym. on Security & Privacy*

Announcements (Tue. Sep. 22)

- Homework #2 due in one week
  - Start now, if you haven't already
- Homework #1 sample solution was handed out last Thursday—see me if you didn't pick one up
- Project milestone #1 due in 2½ weeks
  - Come to my office hours and chat
- Midterm in class next Thursday
  - Sample midterm (from last year) available today
    - Sample solution available Thursday

An abundance of data

Estimated user data generated per day on the Web

- 8-10 GB public content
- ~4TB private content [Ramakrishnan et al., IEEE Computer 2007]
  - Emails
  - Instant messages
  - Tags/page views/annotations
  - Browsing and shopping histories
  - Social networks…

Beyond Web?

- 5 Exabytes (EB) of new information in 2002 alone [How Much Information 2003, UC Berkeley]

Exploiting user-generated content

- Social advertising: i.e., generate ads based on shopping histories of friends in the social network
- User-targeted subscriptions: e.g., recommend books based on those read by other readers
- Data mining for investigative reporting: e.g., mine tax returns of non-profits in conjunction with social networks

☞ Valuable information can be learned by sharing personal data

What about privacy?

"… Last week AOL did another stupid thing …
… but, at least it was in the name of science…"

— Alternet, August 2006

AOL data release

- AOL released a list of 21 million "anonymized" Web search queries
  - Uer ids were replaced by random numbers

```
junyang: duke registrar
junyang: sigmod 2010
junyang: dom se 6
junyang: youtube thumper
junyang: cefalexin
chunhu:  toddler ear infection
```

```
142857: duke registrar
142857: sigmod 2010
142857: dom se 6
142857: youtube thumper
142857: cefalexin
314159: toddler ear infection
```

AOL searcher #4417749

7

[New York Times, August 9, 2006]

… No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

…

AOL searcher #4417749

8

Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. In response, she plans to drop her AOL subscription. "We all have a right to privacy," she said. "Nobody should have found this all out."

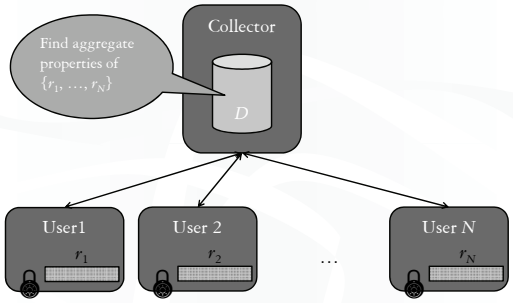
What is privacy?

9

- "*The claim of individuals, groups, or institutions to determine for themselves when, how and to what extent information about them is communicated to others*"
  - Westin, *Privacy and Freedom*, 1967
- "*… nothing about an individual should be learnable from the database that cannot be learned without access to the database…*"
  - T. Dalenius, 1977

Model I: untrusted data collector

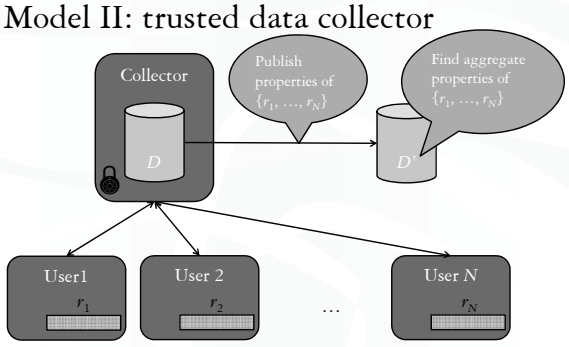
10



How do you collect data from users in a privacy-preserving way that still allows you to compute the aggregate properties?

Model II: trusted data collector

11



How do you publish data from users in a privacy-preserving way that supports the "largest" set of queries?

Types of disclosure

12

- Statistically private
  - Make private data too fuzzy to draw any meaningful conclusion
  - ☞Knowledge as distribution
    - Focus of this lecture
- Computationally private
  - Make data difficult to uncover/leak
  - ☞Cryptographic protocols, certificate revocation

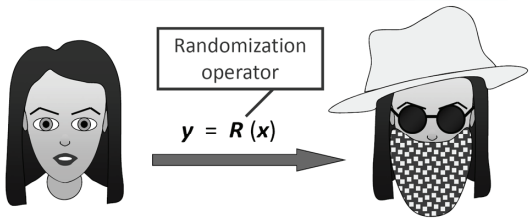
Untrusted collector: social survey

13

- Measures opinions, attitudes, behavior
- Problem: questions of sensitive nature
  - E.g.: sexuality, incriminating/embarrassing/threatening questions, controversial issues, etc.
  - The "non-cooperative" group leads to errors in surveys
  - Even though privacy is guaranteed, skepticism prevails

The randomized response model

14



- x: original data; private
- y: randomized data; responded
- Assumption: individual users are independent

[Stanley Warner, JASA 1965]

15

- Respondents are given:
  - A source of randomness—a biased coin
    - Bias p is known to the data collector
  - A statement S: e.g., I am a member of the XYZ party
- The procedure
  - Respondent flips the coin in private
    - Head = Yes, Tail = No
  - Respondent answers:
    - "Yes" if coin's answer coincides with the true answer
    - "No" otherwise

Randomized response

16

- The procedure
  - Respondent flips the coin in private
    - Head = Yes, Tail = No
  - Respondent answers:
    - "Yes" if coin's answer coincides with the true answer
    - "No" otherwise

	Truth = Yes	Truth = No
Head (Yes)	Answer = Yes	Answer = No
Tail (No)	Answer = No	Answer = Yes

- In other words, truth is flipped with probability $(1 - p)$
- As long as $0 < p < 1$, the collector doesn't know the truth for sure ⇒ protection for respondent

Analysis

17

- π = true probability of S in the population
- p = coin bias (probability of head)
- Y_i = 1 if the i-th respondent says "yes"
  0 if the i-th respondent says "no"

- $P(Y_i = 1) = \pi p + (1 - \pi)(1 - p) = p_{yes}$
- $P(Y_i = 0) = (1 - \pi) p + \pi (1 - p) = p_{no}$

Analysis (cont'd)

18

- Say we surveyed n respondents, of which m said "yes" and $(n - m)$ said "no"
- Likelihood of this outcome $L = (p_{yes})^m (p_{no})^{n-m}$
  - A function of π, p, n, m
- Maximum likelihood estimate of π (i.e., one that maximizes L) $\pi^{hat} = (p-1) / (2p-1) + m / [n(2p-1)]$
- Easy to show that
  - $E(\pi^{hat}) = \pi$; i.e., the estimator is unbiased
  - $Var(\pi^{hat}) = \{ 1/[16(p - 0.5)^2] - (\pi - 0.5)^2 \} / n$
    - When $p \to 0.5$, $Var(\pi^{hat})$ suffers

Summary of randomized response

19

- Collector can still estimate the answer (π) even though the responses are randomized
- But how about privacy?
  - $p = 1$?
  - $p = 0$?
  - $p = 0.5$?

☞ Clear trade-off between utility and privacy

Interval privacy

20

[Agrawal & Srikant, SIGMOD 2000]

- Same idea of randomization applied to numeric data
  - For a numeric attribute value x, share value $z = x + y$, where noise y is drawn from some known distribution
- Example: add to *age* a value drawn from a uniform distribution over [–30, 30]
  - If randomized *age* is 60
    - We know with 100% confidence *age* is between 30 and 90
    - With what confidence do we know *age* is between 33 and 87?
      - 90%
- Amount of privacy: width of interval to which adversary can localize original value with some confidence

An attack on interval privacy

21

[Agrawal & Aggarwal, PODS 2001]

- Noise Y is uniform over [–1, 1]
  - Claim: privacy = 2 at 100% confidence interval
- Attribute X with following distribution
  - With ½ probability value falls uniformly within [0, 1]
  - With ½ probability value falls uniformly within [4, 5]
- What if randomized value $Z=X+Y$ happens to be 0.5?
  - With 100% confidence X falls in [0, 1] ⇒ privacy = 1!
  - In general, $Z\in[-1,2] \Rightarrow X\in[0,1]$, while $Z\in[3,6] \Rightarrow X\in[4,5]$
  - Specifically, $Z=-0.5 \Rightarrow X\in[0,0.5]$, and $Z=-1 \Rightarrow X=0$

What went wrong?

22

- Original distribution of X was ignored
  - Some values of X may be highly unlikely
  - If we see "outlier" values of Z, they constrain the corresponding value of X
- Approach: quantify the information content of distribution of randomized records compared with the distribution of original records

An entropy-based measure

23

- (Continuous) entropy of a random variable X with probability density function $f(x)$: $h(X) = -\int_x f(x) \log f(x)\, dx$
  - A standard measure of average "surprisal" in data
  - X is uniform over [0, 1]: $h(X) = 0$
  - X is uniform over [0, a]: $h(X) = \log a$
- Define privacy of X as $\Pi(X) = 2^{h(X)}$
  - Intuitively, $\Pi(X)$ is the length of an interval over which a uniformly distributed random variable has as much uncertainty as X

An entropy-based measure (cont'd)

24

- Conditional entropy of X given disclosure Z: $h(X|Z) = -\int_{x,z} f_{X,Z}(x,z) \log f_{X|Z=z}(x)\, dx\, dz$
- Define (average) conditional privacy of X given Z as $\Pi(X|Z) = 2^{h(X|Z)}$
- Define conditional privacy loss of X given Z as $\text{Loss}(X|Z) = 1 - \Pi(X|Z) / \Pi(X)$
  - I.e., fraction of privacy of X which is lost by revealing Z

Example revisited

25

- Attribute X with following distribution
  - With ½ probability value falls uniformly within [0, 1]
  - With ½ probability value falls uniformly within [4, 5]
  - ☞ Can calculate privacy $\Pi(X) = 2$
- $Z = X + Y$, where noise Y is uniform over [–1, 1]
  - ☞ Can calculate Loss$(X|Z) = 1 - \Pi(X|Z)/\Pi(X) \approx 0.5796$
    - I.e., more than half of the privacy is lost

Are we done?

26

Turns out that this measure is still imperfect!

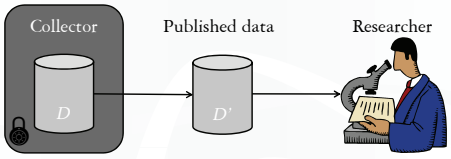
- It measures only the average loss of privacy
- But bad things still happen for rare disclosures
  - E.g., $Z \in [-1,-0.99]$ (rare, but possible) $\Rightarrow X \in [0,-0.01]$

☞ Better measures are needed, e.g., α-β privacy

- Check literature for more

Trusted collector: data publishing

27



- Tricks in publishing
  - Generalization/redaction, perturbation
  - Synthetic data generation
- Utility goals
  - "Difference" between published and original data is small
  - Difference in answers over a query workload is small

Example: detailed medical data

28

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer

- Medical records of a hospital near Ithaca serving patients from
  - Freeville (13068)
  - Dryden (13053)
  - Ithaca (14850, 14853)

Removing SSN…

29

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

(Or replacing them with random numbers)

- The remaining attributes are obviously useful to medical research
  - But at what cost to privacy?

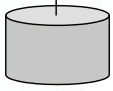
Linkage attack

30

Sensitive attribute

Quasi-identifier: a set of attributes that don't form a key, but may become specific enough to indentify an individual

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer



Background information: e.g., from other public sources

Linkage attack demonstrated

31

- Medical data was considered anonymous when identifying attributes were removed
- Governor of Massachusetts was uniquely identified by zip, birth date, and sex
  - Just join with voter registration list (available for purchase by a small sum of money)
  - His private medical records were out there in the open!
- ☞ Think it's rare?
  - 87% of US population can be uniquely identified using the above quasi identifier!

- Ethnicity
- Visit date
- Diagnosis
- Procedure
- Medication
- Total charge
- Zip
- Birth date
- Sex
- Name
- Address
- Date registered
- Party affiliation
- Date last voted

K-anonymity

32

[Samarati et al., PODS 1998]

- Generalize/perturb quasi-identifier values so that no individual is uniquely identifiable from a group of k
  - User-specified parameter k indicates the "degree" of anonymity
- In SQL, table T is k-anonymous if each result in

```
SELECT COUNT(*)
FROM T
GROUP BY quasi-identifier;
```

is $\geq k$

Generalization: coarsening values

33

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Zip	Age	Nationality	Disease
130**	<30	Russian	Heart
130**	<30	American	Heart
130**	<30	Japanese	Viral
130**	<30	American	Viral
148**	>40	Indian	Cancer
148**	>40	Russian	Heart
148**	>40	American	Viral
148**	>40	American	Viral
130**	30-40	American	Cancer
130**	30-40	Indian	Cancer
130**	30-40	Japanese	Cancer
130**	30-40	American	Cancer

Resulting table is 1-anonymous (i.e., not much protection)

Generalization (cont'd)

34

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Viral
130**	<30	*	Viral
148**	>40	*	Cancer
148**	>40	*	Heart
148**	>40	*	Viral
148**	>40	*	Viral
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Resulting table is 4-anonymous

Homogeneity attack

35

- Alice's neighbor Bob is in the hospital
- Alice knows Bob is 35 years old and is from Dryden (13053)
- ☞ Alice learns that Bob has cancer!

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Viral
130**	<30	*	Viral
148**	>40	*	Cancer
148**	>40	*	Heart
148**	>40	*	Viral
148**	>40	*	Viral
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Background knowledge attack

36

- Alice's friend Umeko went to see a doctor
- Alice knows Umeko is 24, a Japanese, living in Freeville (13068)
- Japanese have extremely low incidence of heart disease — Background knowledge
- ☞ Alice learns Umeko has a viral infection!

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Viral
130**	<30	*	Viral
148**	>40	*	Cancer
148**	>40	*	Heart
148**	>40	*	Viral
148**	>40	*	Viral
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

So, again…

37

Turns out that K-anonymity is still imperfect!

☞ Better measures are needed

❖ E.g., L-diversity

- Intuitively, every group should have at least *L* "well represented" groups of sensitive values
  - "Well represented": roughly equal, non-negligible proportions

❖ Check literature for more

Summary

38

❖ Privacy is a pressing but difficult issue still under very active research

❖ Fundamental trade-off: utility vs. privacy

- A good solution should give you a knob to turn
- But just defining what the knob measures is challenging

❖ This lecture has barely touched the tip of the iceberg

- Leaves you with more questions than answers
- Check out the tutorial by Gehrke & Machanavajjhala (cited on the title slide), and follow literature