

CS 6431

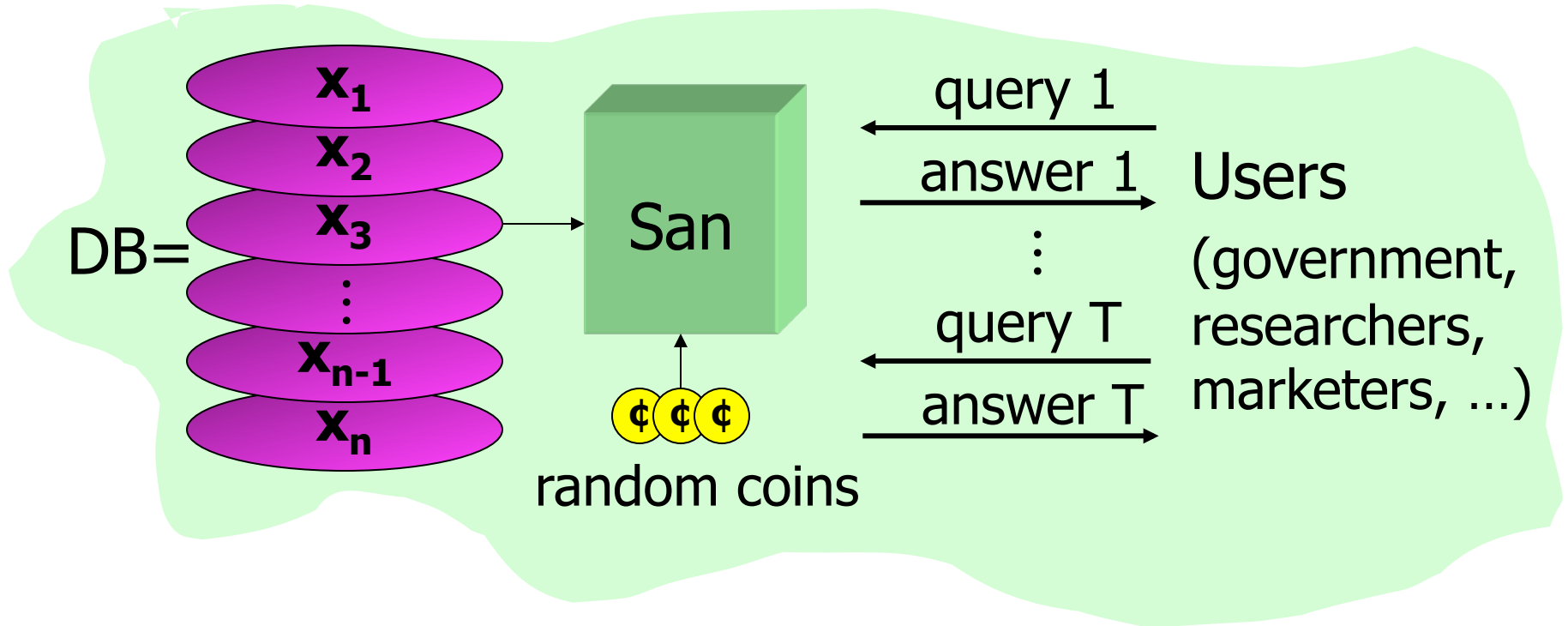
Data Privacy

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



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

Linkage Attack

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	City	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party
.....
Sue J. Carlson	1459 Main St.	Cambridge	02142	9/15/61	female	democrat
.....

Figure 1 Re-identifying anonymous data by linking to external data

Public voter dataset

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 '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

Identifiers vs. Sensitive Attributes

◆ Sensitive attributes

- Medical records, salaries, etc.
- These attributes is what the researchers need, so they are released unmodified

Key Attribute		Quasi-identifier		Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

- ◆ The information for each person contained in the released table cannot be distinguished from at least $k-1$ individuals whose information also appears in the release
 - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- ◆ Any quasi-identifier present in the released table must appear in at least k records

K-Anonymity Protection Model

- ◆ Private table \rightarrow Released table RT
- ◆ Attributes: A_1, A_2, \dots, A_n
- ◆ Quasi-identifier subset: A_i, \dots, A_j

Let $RT(A_1, \dots, A_n)$ be a table, $QI_{RT} = (A_i, \dots, A_j)$ be the quasi-identifier associated with RT, $A_i, \dots, A_j \subseteq A_1, \dots, A_n$, and RT satisfy k -anonymity. Then, each sequence of values in $RT[A_x]$ appears with at least k occurrences in $RT[QI_{RT}]$ for $x=i, \dots, j$.

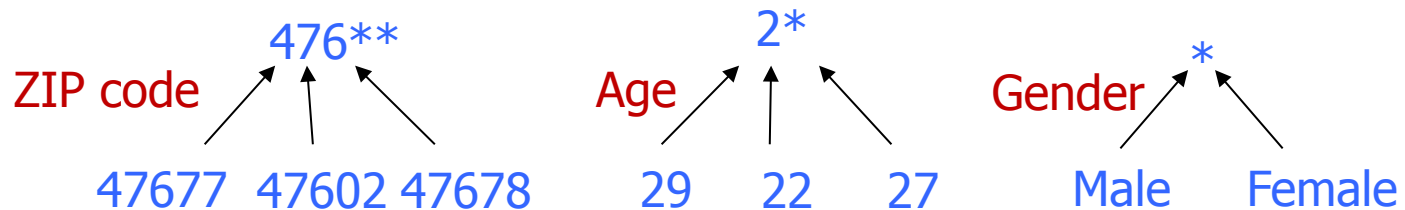
Goal: each record is indistinguishable from at least $k-1$ other records ("equivalence class")

Achieving k-Anonymity

Lots of algorithms in the literature aiming to produce “useful” anonymizations, usually without any clear notion of utility

◆ Generalization

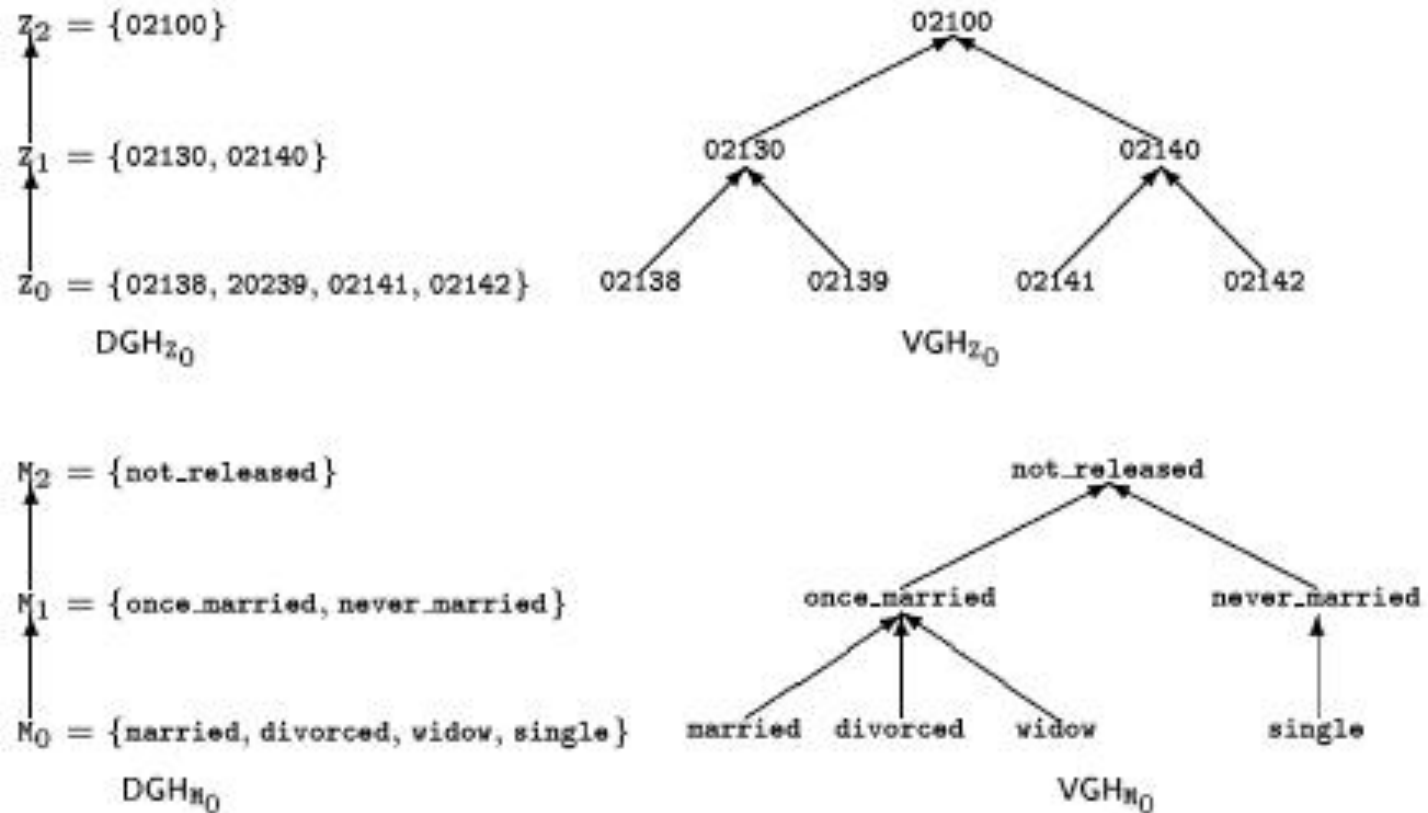
- Replace quasi-identifiers with less specific but semantically consistent values until get k identical
- Partition ordered-value domains into intervals



◆ Suppression

- When generalization causes too much information loss (this often happens with “outliers”)

Generalization in Action



Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of k -anonymity, where $k=2$ and $Q=\{Race, Birth, Gender, ZIP\}$

Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data source

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

Figure 2 Example of k -anonymity, where $k=2$ and $Q=\{Race, Birth, Gender, ZIP\}$

By linking these two tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

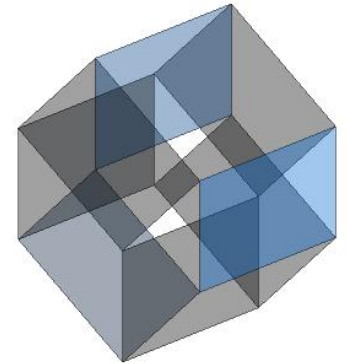
QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- ◆ Released table is 3-anonymous
- ◆ If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

Curse of Dimensionality

[Aggarwal VLDB '05]

- ◆ 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 \Rightarrow k -anonymized datasets are useless



What Does k-Anonymity Prevent?

- ◆ **Membership disclosure:** Attacker cannot tell that a given person is 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+

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!

Complementary Release Attack

Ganta et al. (KDD 2008)

- ◆ Different releases of the same private table can be linked to compromise k-anonymity

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Linking Independent Releases

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Exploiting Distributions

- ◆ k-Anonymity does not provide privacy if
 - Sensitive values in an equivalence class lack diversity
 - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	Age
47678	27

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Background knowledge attack

Carl	
Zipcode	Age
47673	36

I-Diversity

Machanavajjhala et al. (ICDE 2006)

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be
“diverse” within each
quasi-identifier equivalence class

Distinct I-Diversity

- ◆ Each equivalence class has at least I well-represented sensitive values
- ◆ Doesn't prevent probabilistic inference attacks

...	Disease
	...
	HIV
	HIV
	...
	HIV
	pneumonia
	bronchitis
	...

10 records {

8 records have HIV

2 records have other values

Other Versions of l -Diversity

◆ Probabilistic l -diversity

- The frequency of the most frequent value in an equivalence class is bounded by $1/l$

◆ Entropy l -diversity

- The entropy of the distribution of sensitive values in each equivalence class is at least $\log(l)$

◆ Recursive (c, l) -diversity

- $r_1 < c(r_1 + r_{l+1} + \dots + r_m)$ where r_i is the frequency of the i^{th} most frequent value
 - Most frequent value does not appear too frequently

My Favorite Charts

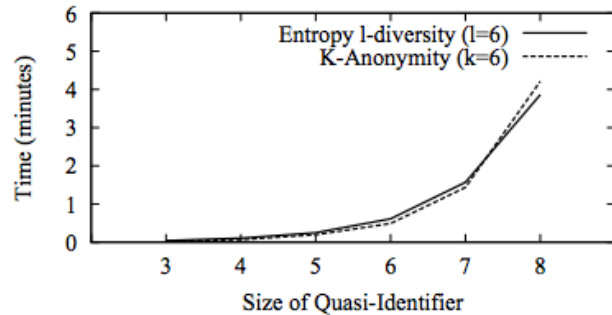


Figure 5. Adults Database

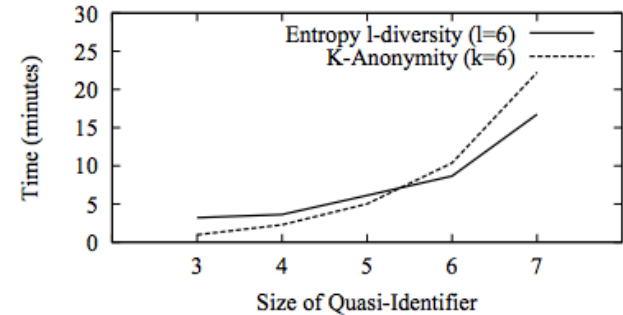


Figure 6. Lands End Database

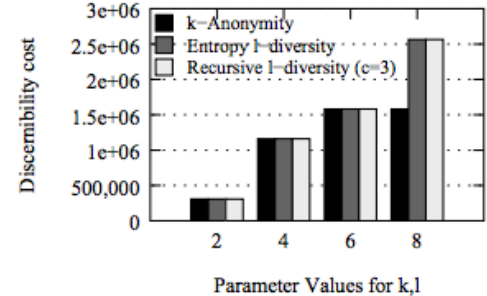
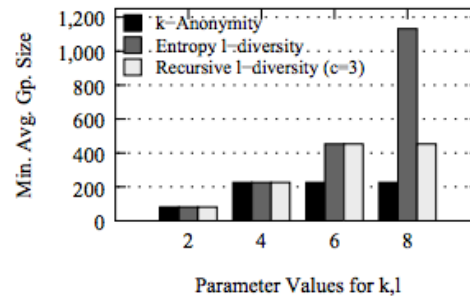
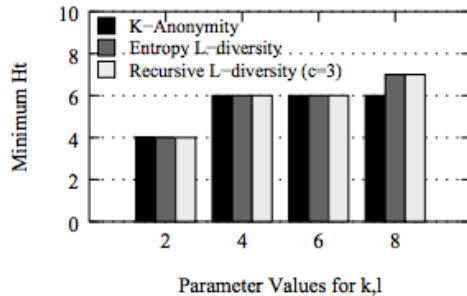


Figure 7. Adults Database. $Q = \{\text{age, gender, race, marital_status}\}$

Limitations of l-Diversity

- ◆ Example: sensitive attribute is HIV+ (1%) or HIV- (99%) – very different sensitivity!
- ◆ l-diversity is unnecessary
 - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- ◆ l-diversity is difficult to achieve
 - Suppose there are 10000 records in total
 - To have distinct 2-diversity, there can be at most $10000 * 1\% = 100$ equivalence classes

Skewness Attack

- ◆ Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- ◆ Consider an equivalence class that contains an equal number of HIV+ and HIV- records
 - Diverse, but potentially violates privacy!
- ◆ l-diversity does not differentiate:
 - Equivalence class 1: 49 HIV+ and 1 HIV-
 - Equivalence class 2: 1 HIV+ and 49 HIV-

Does not consider overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	Age
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥ 40	50K	Gastritis
4790*	≥ 40	100K	Flu
4790*	≥ 40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

Does not consider the semantics of sensitive values!

Try Again: t-Closeness

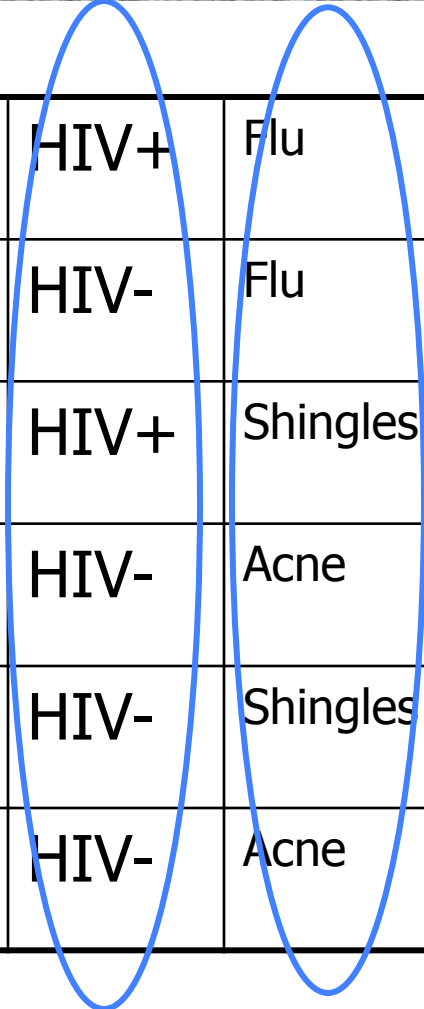
[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

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

Trick question: Why publish quasi-identifiers at all??

Anonymized “t-Close” Database



Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous,
l-diverse and t-close...

...so secure, right?

What Does Attacker Know?

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

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

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

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
		HIV+	Shingles
Caucas	787XX	HIV-	Acne
		HIV-	Shingles
Caucas	787XX	HIV-	Acne

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

Strawman Definition

- ◆ Assume x_1, \dots, x_n are drawn i.i.d. from unknown distribution
- ◆ Candidate definition: sanitization is safe if it only reveals the distribution
- ◆ Implied approach:
 - Learn the distribution
 - Release description of distribution or re-sample points
- ◆ This definition is tautological!
 - Estimate of distribution depends on data... why is it safe?

Blending into a Crowd

Frequency in DB or frequency in underlying population?

◆ Intuition: “I am safe in a group of k or more”

- k varies (3... 6... 100... 10,000?)

◆ Many variations on theme

- Adversary wants predicate g such that $0 < \#\{i \mid g(x_i)=\text{true}\} < k$

◆ Why?

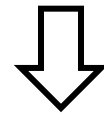
- Privacy is “protection from being brought to the attention of others” [Gavison]
- Rare property helps re-identify someone
- Implicit: information about a large group is public
 - E.g., liver problems more prevalent among diabetics



Clustering-Based Definitions

- ◆ Given sanitization S , look at all databases consistent with S
- ◆ Safe if no predicate is true for all consistent databases
- ◆ k-anonymity
 - Partition D into bins
 - Safe if each bin is either empty, or contains at least k elements
- ◆ Cell bound methods
 - Release marginal sums

	brown	blue	Σ
blond	2	10	12
brown	12	6	18
Σ	14	16	



	brown	blue	Σ
blond	[0,12]	[0,12]	12
brown	[0,14]	[0,16]	18
Σ	14	16	

Issues with Clustering

- ◆ Purely syntactic definition of privacy
- ◆ What adversary does this apply to?
 - Does not consider adversaries with side information
 - Does not consider probability
 - Does not consider adversarial algorithm for making decisions (inference)

“Bayesian” Adversaries

- ◆ Adversary outputs point $z \in D$
- ◆ Score = $1/f_z$ if $f_z > 0$, 0 otherwise
 - f_z is the number of matching points in D
- ◆ Sanitization is safe if $E(\text{score}) \leq \varepsilon$
- ◆ Procedure:
 - Assume you know adversary's prior distribution over databases
 - Given a candidate output, update prior conditioned on output (via Bayes' rule)
 - If $\max_z E(\text{score} \mid \text{output}) < \varepsilon$, then safe to release

Issues with “Bayesian” Privacy

- ◆ Restricts the type of predicates adversary can choose
- ◆ Must know prior distribution
 - Can one scheme work for many distributions?
 - Sanitizer works harder than adversary
- ◆ Conditional probabilities don't consider previous iterations
 - Notorious problem in query auditing

Problems with Classical 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

[Dwork]

- ◆ Privacy: for some definition of “privacy breach,”
 \forall distribution on databases, \forall adversaries A , $\exists A'$
such that $\Pr(A(\text{San})=\text{breach}) - \Pr(A'()=\text{breach}) \leq \epsilon$
 - For reasonable “breach”, if $\text{San}(\text{DB})$ 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!

(Very Informal) Proof Sketch

- ◆ Suppose DB is uniformly random
 - Entropy $I(\text{DB} ; \text{San}(\text{DB})) > 0$
- ◆ “Breach” is predicting a predicate $g(\text{DB})$
- ◆ Adversary knows $r, H(r ; \text{San}(\text{DB})) \oplus g(\text{DB})$
 - H is a suitable hash function, $r = H(\text{DB})$
- ◆ By itself, does not leak anything about DB (why?)
- ◆ Together with $\text{San}(\text{DB})$, reveals $g(\text{DB})$ (why?)

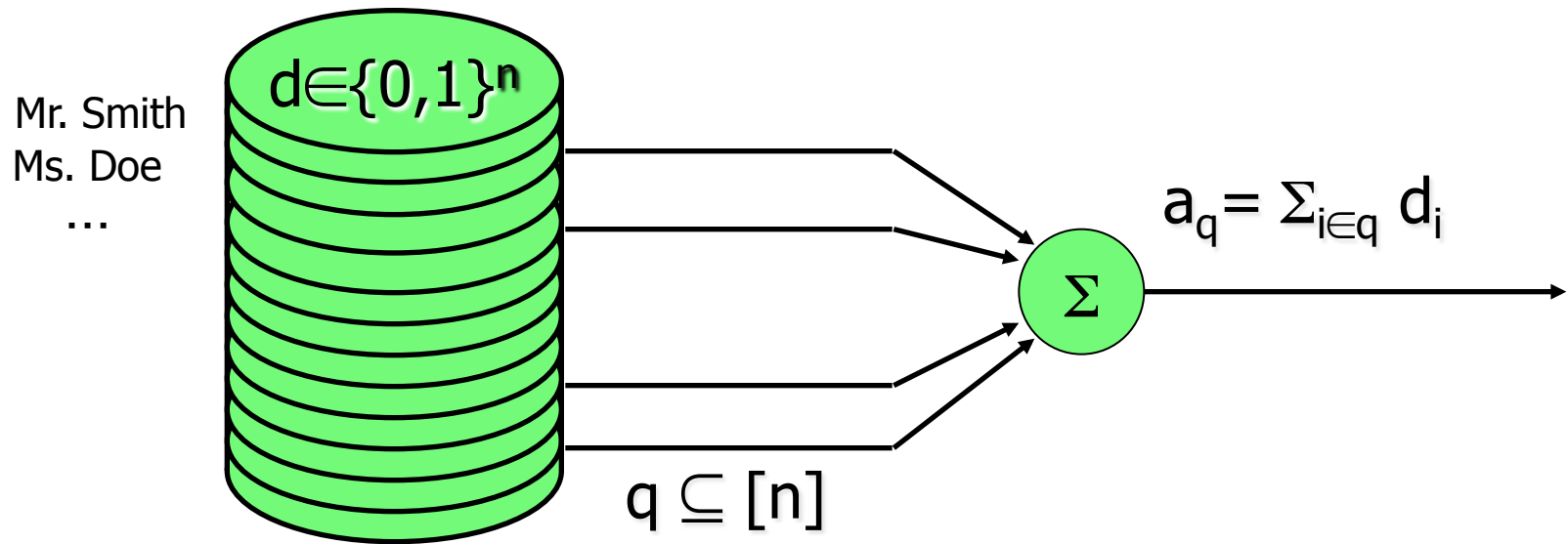
Dinur-Nissim Impossibility Results

[PODS 2003]



The following slides shamelessly jacked from Kobbi Nissim

Statistical Database (SDB)



Information-Privacy Tradeoff

◆ Private function:

- want to hide $\pi_i(d_1, \dots, d_n) = d_i$

◆ Information functions

- want to reveal $f_q(d_1, \dots, d_n) = \sum_{i \in q} d_i$

◆ Explicit definition of private functions

• Crypto: secure function evaluation

- want to reveal $f()$
- want to hide all functions $\pi()$ not computable from $f()$
- Implicit definition of private functions

Approaches to SDB Privacy

[Adam and Wortmann 1989]

- ◆ Query restriction
 - Require queries to obey some structure
- ◆ Perturbation
 - Give “noisy” or “approximate” answers

Perturbation

- ◆ Database: $d = d_1, \dots, d_n$
- ◆ Query: $q \subseteq [n]$
- ◆ Exact answer: $a_q = \sum_{i \in q} d_i$
- ◆ Perturbed answer: \hat{a}_q

Perturbation E :

For all q : $|\hat{a}_q - a_q| \leq E$

General perturbation:

$$\begin{aligned} \Pr_q [|\hat{a}_q - a_q| \leq E] &= 1 - \text{neg}(n) \\ &= 99\%, 51\% \end{aligned}$$

Perturbation Techniques

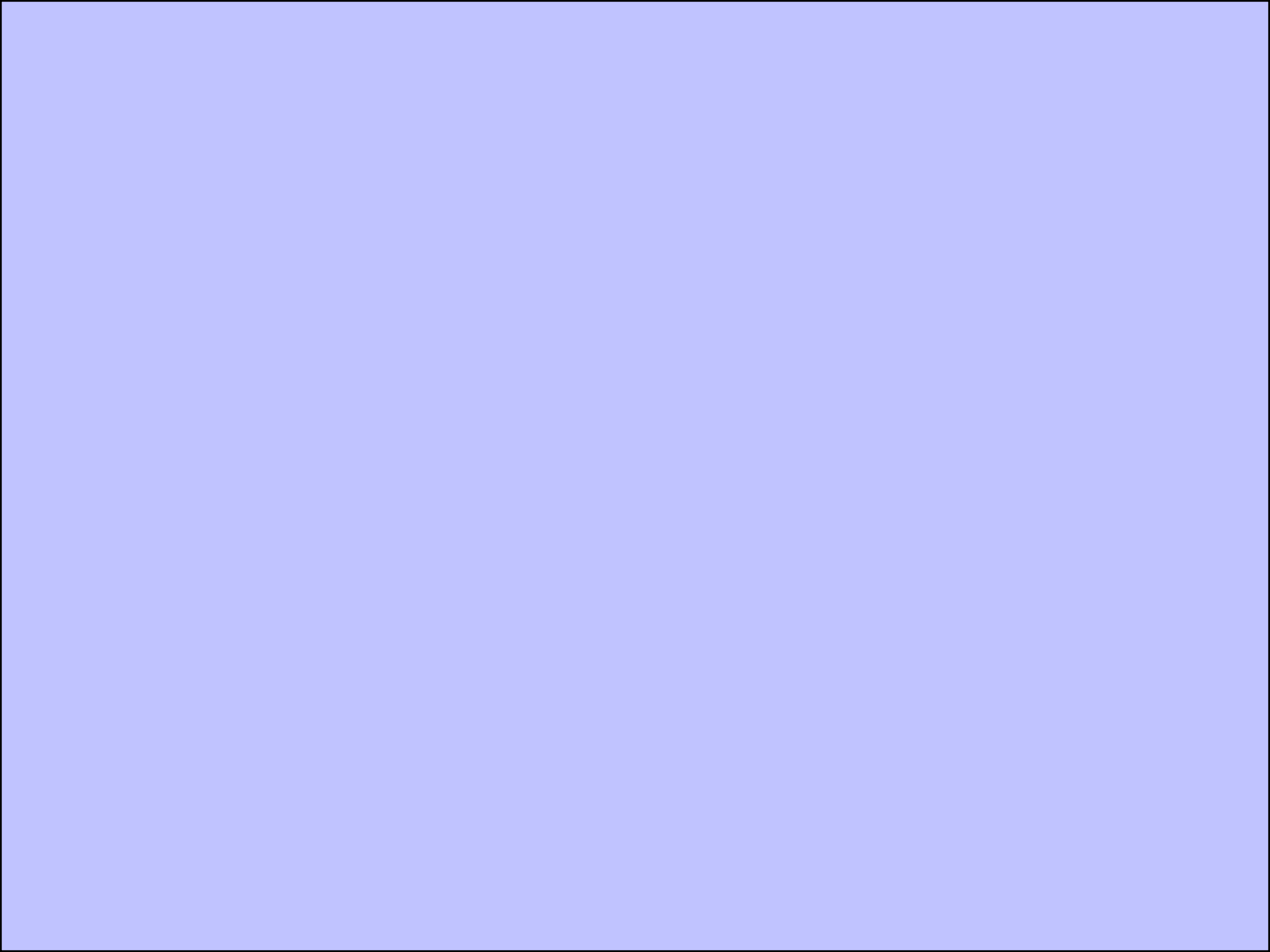
[Adam and Wortmann 1989]

Data perturbation:

- Swapping [Reiss 84][Liew, Choi, Liew 85]
- Fixed perturbations [Traub, Yemini, Wozniakowski 84] [Agrawal, Srikant 00] [Agrawal, Aggarwal 01]
 - Additive perturbation $d'_i = d_i + E_i$

Output perturbation:

- Random sample queries [Denning 80]
 - Sample drawn from query set
- Varying perturbations [Beck 80]
 - Perturbation variance grows with number of queries
- Rounding [Achugbue, Chin 79] Randomized [Fellegi, Phillips 74] ...



Privacy from $\approx\sqrt{n}$ Perturbation

- Database: $d \in_R \{0,1\}^n$
- On query q :
 1. Let $a_q = \sum_{i \in q} d_i$
 2. If $|a_q - |q|/2| > E$ return $\hat{a}_q = a_q$
 3. Otherwise return $\hat{a}_q = |q|/2$
- Privacy is preserved
 - If $E \cong \sqrt{n} (\lg n)^2$, whp always
 - No information about d if
- USELESS!

Can we do better?

- Smaller E ?
- Usability ???

Main Theorem

Given a DB response algorithm with **perturbation** $E \ll \sqrt{n}$,
there is a poly-time **reconstruction** algorithm that outputs a
database d' s.t. **$\text{dist}(d, d') < o(n)$**

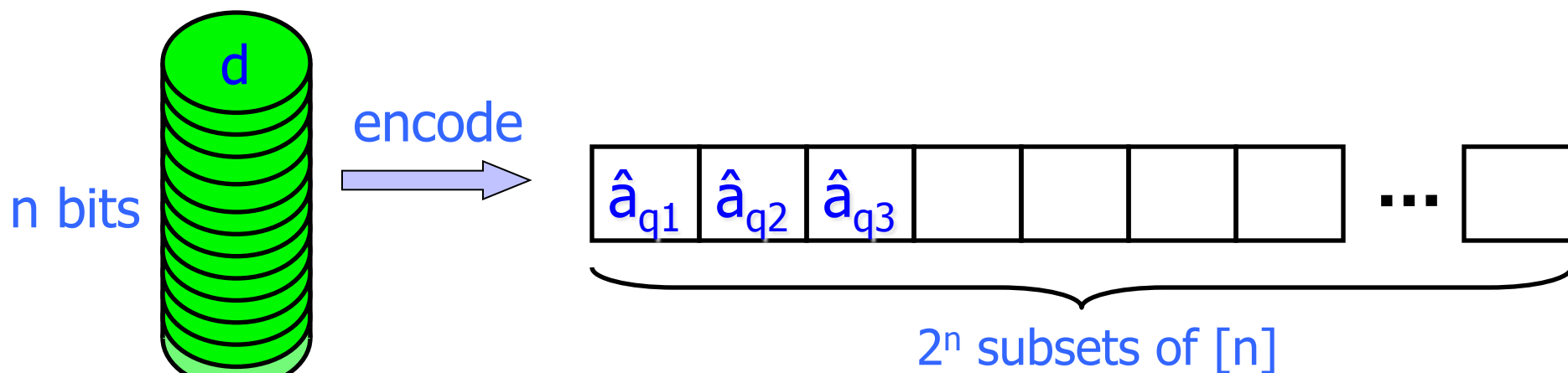
... yet can reconstruct
the entire database

So much perturbation,
responses are useless

Conclusion:

privacy in statistical databases cannot be achieved

Decoding Adversary



(where $\hat{a}_q = \sum_{i \in q} d_i + \text{pert}_q$)

Decoding problem:

given access to $\hat{a}_{q1}, \dots, \hat{a}_{q_{2^n}}$ reconstruct d in time $\text{poly}(n)$

Reconstruction Algorithm

- **Query phase:** Get \hat{a}_{q_j} for t random subsets q_1, \dots, q_t of $[n]$
- **Weeding phase:** Solve the linear program:
$$0 \leq x_i \leq 1$$
$$|\sum_{i \in q_j} x_i - \hat{a}_{q_j}| \leq E$$
- **Rounding:** Let $c_i = \text{round}(x_i)$, output c

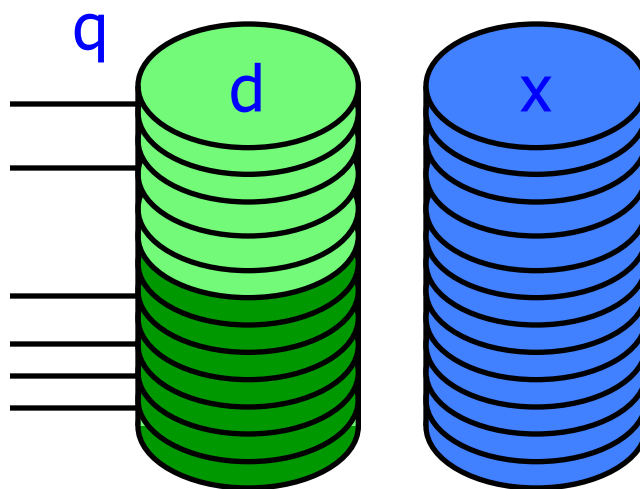
Observation: An LP solution always exists, e.g. $x=d$

Proof of Correctness

Consider $x=(0.5,\dots,0.5)$ as a solution for the LP

Observation: A random q often shows a \sqrt{n} advantage either to 0's or to 1's.

- Such a q disqualifies x as a solution for the LP
- We prove that if $\text{dist}(x,d) > \epsilon \cdot n$, then whp there will be a q among q_1,\dots,q_t that disqualifies x



Extensions of the Main Theorem

◆ “Imperfect” perturbation

- Can approximate the original bit string even if database answer is within perturbation only for 99% of the queries

◆ Other information functions

- Given access to “noisy majority” of subsets, can approximate the original bit string

Adversaries

- ◆ Exponential adversary
 - Strong breaking of privacy if $E \ll n$
- ◆ Polynomial adversary
 - Non-adaptive queries
 - Oblivious of perturbation method and database distribution
 - Tight threshold $E \approx \sqrt{n}$
- ◆ What if adversary is more restricted?