CSE660
Differential Privacy

August 30, 2017

Marco Gaboardi
Room: 338-B
gaboardi@buffalo.edu
http://www.buffalo.edu/~gaboardi

Data

Aol.

NETFLIX

Google

Statistics over Data

Is this data private?

	D1	D2	D3	D4	D5	D6	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
I1	0	0	0	1	0	0	0	0	1	0	0	1	1	0	0	1
I2	1	0	1	1	1	0	1	0	1	0	1	0	0	1	0	0
I3	0	1	0	1	1	1	0	1	0	0	0	1	0	0	1	0
I4	1	0	1	0	0	1	1	0	1	1	0	0	0	0	1	1
I5	0	0	0	1	1	0	1	1	0	1	0	1	0	1	0	0
I6	0	0	1	1	0	1	1	0	1	1	0	0	1	0	1	0
I7	1	1	0	0	1	0	1	1	1	0	1	0	1	0	0	1
I8	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0
I9	0	1	0	0	1	0	1	1	0	1	1	1	0	1	1	0
I10	1	0	1	0	0	1	1	0	0	0	0	0	0	1	0	1
I11	0	1	0	1	1	0	0	1	0	1	0	1	0	1	1	0
I12	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
I13	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	0
I14	0	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0
I15	0	1	0	1	0	1	1	0	1	0	1	0	1	0	0	1

How about if we also have this data?

	ID	Name
1	I1	Alice
2	I2	Bob
3	I3	Cynthia
4	I4	Dan
5	I5	Eve
6	I6	Frank
7	I7	Guy
8	I8	Hannah
9	I9	Ivan
10	I10	Jon
11	I11	Ken
12	I12	Lou
13	I13	Mike
14	I14	Noa
15	I15	Omer

	ID	Disease
1	D1	AMAN
2	D2	Behcet
3	D3	Celiac
4	D4	Dermatitis
5	D5	Evans synd.
6	D6	Fibrosis
7	D7	Graves' dis.
8	D8	Henoch-Schonlein
9	D9	IGA Neph.
10	D10	Juv. Diabetes
11	D11	Kawasaki dis.
12	D12	Lichen planus
13	D13	Myositis
14	D14	Narcolepsy
15	D15	Optic Neuritis

How about if we also have this data?

	D2	D3	D5	D6	D8	D10	D12	D14	D15
Alice	0	1	1	0	1	1	0	0	0
Cynthia	1	0	1	1	0	0	0	1	0
Eve	0	0	1	0	0	0	0	0	0
Frank	0	1	0	1	1	0	1	1	0
Guy	1	0	1	0	1	1	1	0	1
Ivan	1	0	1	0	0	1	0	1	0
Jon	0	1	0	1	0	0	0	0	1
Lou	0	0	0	0	0	1	0	0	0
Omer	1	0	0	1	1	1	1	0	1

Database

- We can think about a database as a list of records from some universe set:

$$D \in \mathcal{X}^n$$

- Sometimes we will think to them as functions

$$D(k) \in \mathcal{X}$$

- and sometimes we will write elements explicitly

$$(d_1, \ldots, d_n) \in \mathcal{X}^n$$

Counting Queries

- A counting query $q : \mathcal{X}^n \to [0, 1]$ is a function counting the fraction of people in a dataset satisfying the predicate $q : \mathcal{X} \to \{0, 1\}$
- In symbols:

$$q(D) = \frac{1}{n} \sum_{i=1}^{n} q(d_i)$$

- Notice that we take a normalized count, which also corresponds to the average.

Example 1

Let's consider an arbitrary universe domain $\mathcal{X}$ and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

we call a point function the associated counting query

$$q_y : \mathcal{X}^n \to [0, 1]$$

Question: Suppose that we answer all the point function queries for $y \in \mathcal{X}$. What well know statistics do we obtain?

Example I

$X=\{0,1\}^3$ $\quad D \in X^{10} =$

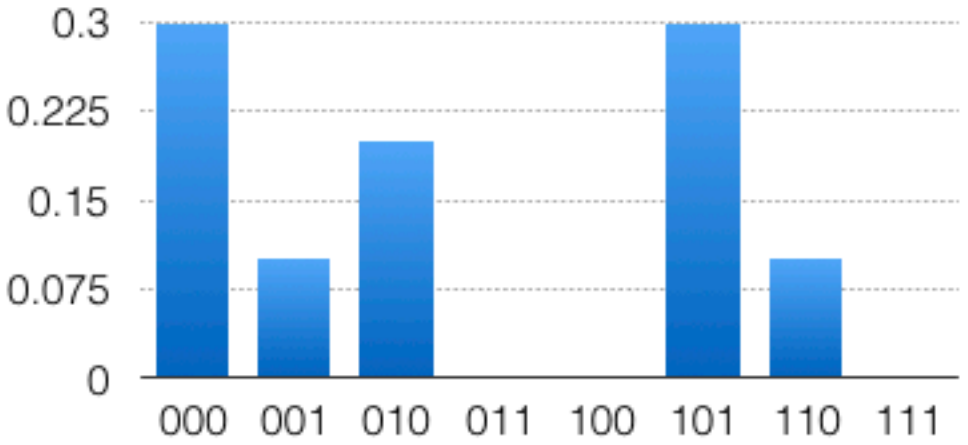
	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

$q_{000}(D) = .3$ $\quad q_{100}(D) = 0$

$q_{001}(D) = .1$ $\quad q_{101}(D) = .3$

$q_{010}(D) = .2$ $\quad q_{110}(D) = .1$

$q_{011}(D) = 0$ $\quad q_{111}(D) = 0$



Example 1

Question: Suppose that we answer all the point function queries for $y \in \mathcal{X}$. What well know statistics do we obtain?

Answer: Histogram of the universe and of the database.

Example II

Let's consider an arbitrary ordered universe domain $\mathcal{X}$ and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} 1 & \text{if } x \leq y \\ 0 & \text{otherwise} \end{cases}$$

we call a threshold function the associated counting query

$$q_y : \mathcal{X}^n \to [0, 1]$$

Question: Suppose that we answer all the threshold function queries for $y \in \mathcal{X}$. What well know statistics do we obtain?

Example II

$X=\{0,1\}^3$ with order given by the corresponding binary encoding.

$D \in X^{10} =$

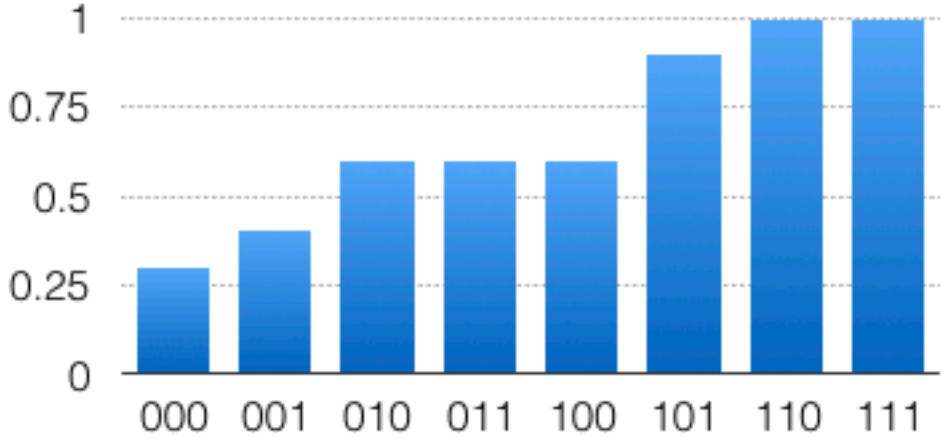
	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

$q_{000}(D) = .3$ $q_{100}(D) = .6$

$q_{001}(D) = .4$ $q_{101}(D) = .9$

$q_{010}(D) = .6$ $q_{110}(D) = 1$

$q_{011}(D) = .6$ $q_{111}(D) = 1$



Example II

Question: Suppose that we answer all the threshold function queries for $y \in \mathcal{X}$. What well know statistics do we obtain?

Answer: CDF of the universe and of the database.

Example III

Let's consider the universe domain $\mathcal{X} = \{0,1\}^d$ and let's consider the following predicate for an index $1 \leq j \leq d$

$$q_j(x) = x_j$$

we call an attribute mean function the associated counting query

$$q_j : \mathcal{X}^n \to [0,1]$$

Question: Which statistics does correspond to releasing all the attribute mean functions?

Example III

$X=\{0,1\}^3$

$D \in X^{10} =$

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1
margin	4	3	4

$q_1(D) = .4$

$q_2(D) = .3$

$q_3(D) = .4$

Example III

Question: Which statistics does correspond to releasing all the attribute mean functions?

Answer: (1-way) Marginals of the distribution

Example IV

Let's consider the universe domain $\mathcal{X} = \{0,1\}^d$ and let's consider $\vec{v} \in \{1, \bar{1}, \ldots, d, \vec{d}\}^k$ with $1 \leq k \leq d$ and

$$q_{\vec{v}}(x) = q_{v_1}(x) \wedge q_{v_1}(x) \wedge \cdots \wedge q_{v_k}(x)$$

where $q_j(x) = x_j$ and $q_{\bar{j}}(x) = \neg x_j$

We call a conjunction or k-way marginal the associated counting query

$$q_{\vec{v}} : \mathcal{X}^n \to [0,1]$$

Question: Which statistics does correspond to releasing conjunctions?

Example IV

$X=\{0,1\}^3$

$D \in X^{10} =$

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

k=2

$q_{12}(D) = .1$ $\quad q_{/12}(D) = .2$

$q_{1/2}(D) = .3$ $\quad q_{/13}(D) = .1$

$q_{13}(D) = .3$ $\quad q_{/1/2}(D) = .4$

$q_{1/3}(D) = .1$ $\quad q_{/1/3}(D) = .5$

	D1	/D1
D2	0.1	0.2
/D2	0.3	0.4

Example IV

Question: Which statistics does correspond to releasing conjunctions?

Answer: contingency tables

Linear Queries

- A linear query $q : \mathcal{X}^n \to [0, 1]$ is a function averaging the value of a function $q : \mathcal{X} \to [0, 1]$ over the elements of the dataset.
- In symbols:

$$q(D) = \frac{1}{n} \sum_{i=1}^{n} q(d_i)$$

Example 1

Let's consider the domain $\mathcal{X} = \{0,1\}^d$ and let's consider the following predicate for $y \in \mathcal{X}$

$$q_y(x) = \begin{cases} .5 * y_1 & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

Example 1

$X=\{0,1\}^3$ $\quad D \in X^{10} =$

	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

$q_{000}(D) = 0$ $\quad q_{100}(D) = 0$

$q_{001}(D) = 0$ $\quad q_{101}(D) = .15$

$q_{010}(D) = 0$ $\quad q_{110}(D) = .05$

$q_{011}(D) = 0$ $\quad q_{111}(D) = 0$

Sum queries

- Let's denote by $I \subseteq [n]$ a subset I of

$$\{0, \ldots, n\}$$

- A sum query $q_I : \{0, 1^k \rightarrow \mathbb{N}^k\}$ is defined as

$$q_I(D) = \sum_{i \in I} d_i$$

Example

$X=\{0,1\}^3$ $\quad D \in X^{10} =$

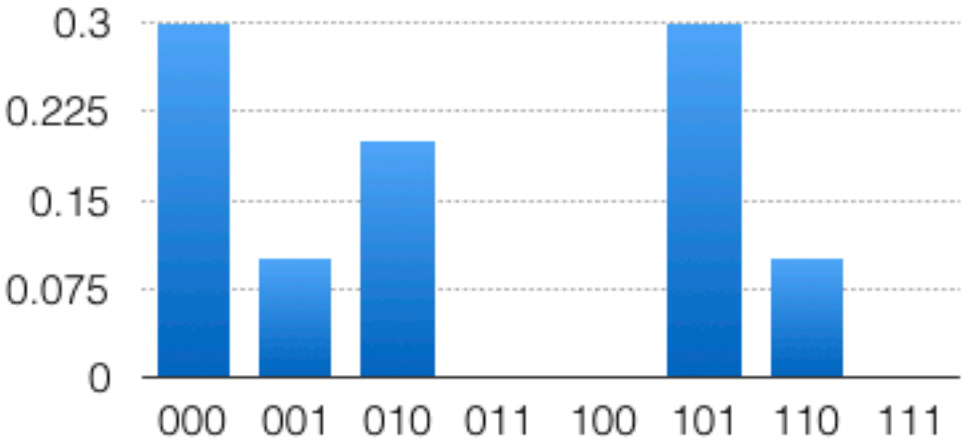
	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

$$q_{\{1,2,3\}}(D) = (1,1,1)$$
$$q_{\{1,2,4\}}(D) = (2,0,2)$$
$$q_{\{5,8\}}(D) = (0,0,0)$$
$$q_{\{2,4,7,10\}}(D) = (4,1,3)$$

Question: Is releasing the result of counting queries private?

Intuitively no, by knowing the results of the statistics we can learn a lot about individuals

Example



	D1	D2	D3
I1	0	0	0
I2	1	0	1
I3	0	1	0
I4	1	0	1
I5	0	0	0
I6	0	0	1
I7	1	1	0
I8	0	0	0
I9	0	1	0
I10	1	0	1

Question: Is releasing the result of linear or sum queries private?

Intuitively no, by knowing the results of the statistics we can learn a lot about individuals

Question: How can we make statistical queries private?

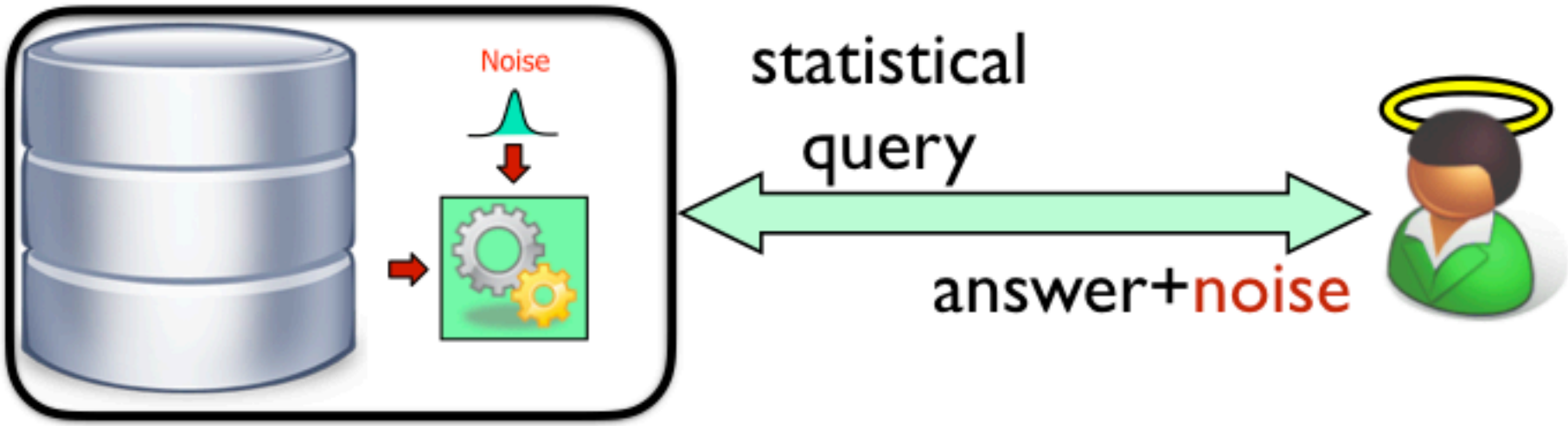
Randomized Algorithms

- Given a discrete set B the probability simplex over B, denoted $\Delta(B)$ is defined as:

$$\Delta(B) = \left\{ x \in \mathbb{R}^{|B|} : \forall i, x_i \geq 0, \text{ and } \sum_{i=1}^{|B|} x_i = 1 \right\}$$

- A randomized algorithm $\mathcal{M}$ is an algorithm associated with a total map $M : A \to \Delta(B)$
  On input $a \in A$ the algorithm outputs $\mathcal{M}(a) = b$ with probability $(M(a))_b$.
  The probability space is over the coin flips of the algorithm.

Private Statistical database



Question: What kind of noise?

Sum queries

- Let's denote by $I \subseteq [n]$ a subset I of

$$\{0, \ldots, n\}$$

- A sum query $q_I : 0, 1^k \to \mathbb{N}^k$ is defined as

$$q_I(D) = \sum_{i \in I} d_i$$

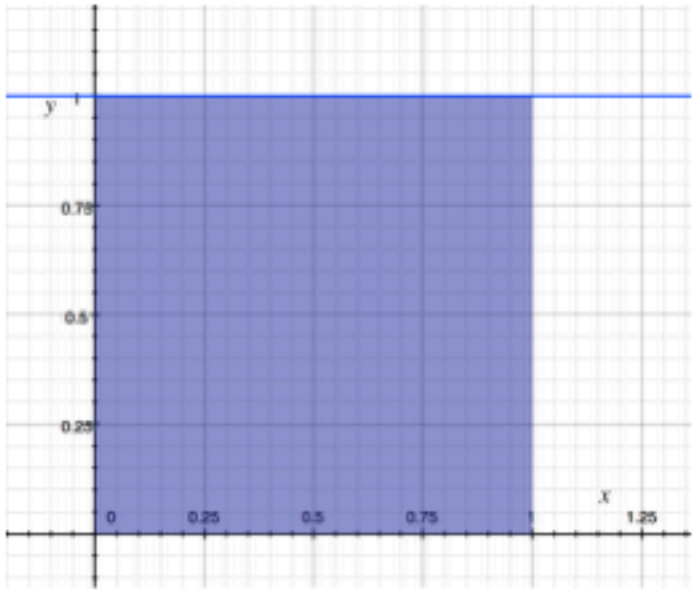
Uniform Noise

- Given a query q we want to add noise to create a new randomized query:

$$q^*(D) = q(D) + Y$$

- One way to do this is to sample Y from the uniform distribution:

$$Y \sim U[0,1]$$



Question: Does this approach protect privacy?

We first need to define what we mean by privacy.

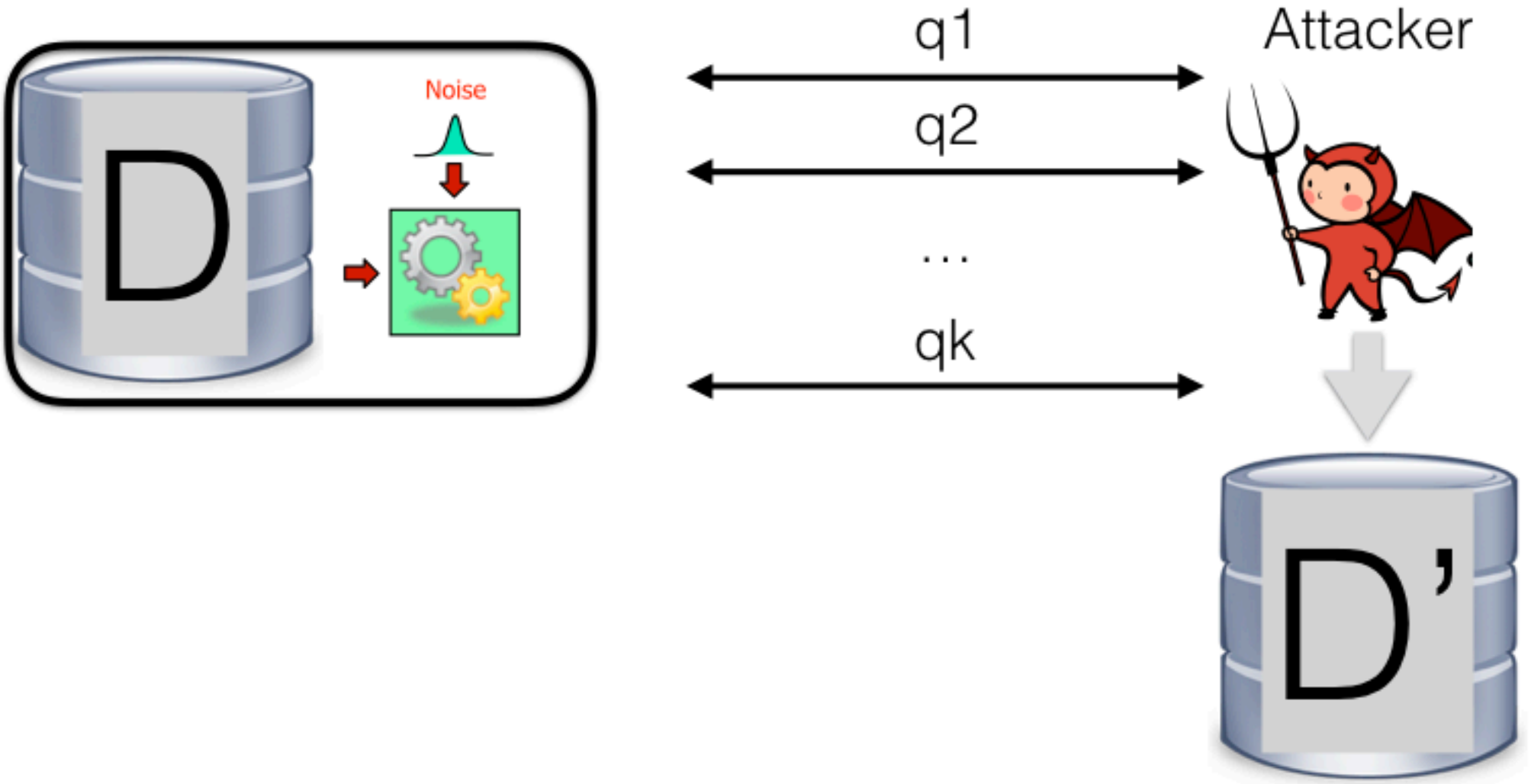
Blatantly non-privacy

- We can consider a protection mechanism as blatantly non-private if it allows to reconstruct the database.
- This may not be a complete reconstruction but an approximate one.

Reconstruction attack

- Consider an adversary A (an algorithm) that has access to some data D through a privacy mechanism q*.
- The goal of the adversary is to output some data D' that is as similar as possible to D.
- To output D' the adversary can interact several times with q*.

Reconstruction attack



Reconstruction attack

We say that the attacker wins if

$d($  ,  $) \sim 0$

In our case we can use Hamming distance

Blatantly non-privacy

The privacy mechanism $M:\{0,1\}^n \rightarrow R$ is blatantly non-private if an adversary can build a candidate database $D' \in \{0,1\}^n$, that agrees with the real database D in all but $o(n)$ entries:

$$d_H(D,D') \in o(n)$$

Additive Noise Perturbation

- We say that M is a privacy mechanism obtained by adding noise if for every query q, M creates a new randomized query:

$$q^*(D) = q(D) + Y$$

- We say that a mechanism M add noise within perturbation E iff for every q and every D:

$$|q^*(D)-q(D)| \leq E$$

Reconstruction attack with exponential adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within E perturbation. Then there is an adversary that can reconstruct the database within 4E positions.

[DinurNissim'02]

Proof

Query phase: For each $S \subseteq [n]$ let $a_S^* = q_S^*(D)$.

Rule out phase: For each $D' \in \{0,1\}^n$:
if there exists S such that $|q_S(D') - a_S^*| > E$ then rule out D'.

Output phase: Output a database D' that was not ruled out.

Notice that since for the real database we clearly have

$$|q_S(D) - q_S^*(D)| \leq E$$

the procedure clearly return a candidate output in an exponential number of steps.

We now want to show that $d_H(D,D') \leq 4E$

Proof

Let 's consider D to be the real dataset and D' to be the outputted one. Consider the sets of indices

$$S = \{ i \mid D(i)=0 \} \quad \text{and} \quad T = \{ i \mid D(i)=1 \}$$

Since D' was not ruled out we have

$$|q_S^*(D)-q_S(D')| \leq E$$

but by definition we also have

$$|q_S^*(D)-q_S(D)| \leq E$$

so by triangle inequality $|q_S(D)-q_S(D')| \leq 2E$. Since $q_S(D)=0$, we have that on the indices in S the Hamming distance between D and D' is at most 2E.

We can apply a similar reasoning toT. So overall D and D' differ in at most 4E positions.

Reconstruction attack with exponential adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within **$E=o(n)$** perturbation. Then M is blatantly non-private against an adversary A running in exponential time.

[DinurNissim'02]

Reconstruction attack with polynomial adversary

Let $M:\{0,1\}^n \rightarrow R$ be a privacy mechanism adding noise within $\mathbf{E=o(\sqrt{n})}$ perturbation. Then we can show M blatantly non-private against an adversary A running in polynomial time and **answering n queries.**

[DinurNissim'02, DworkYekhanin'08]

Number of queries

A privacy mechanism can answer with perturbation $\sqrt{n}$ at most a number of queries sublinear in n.

Question: Why error $\sqrt{n}$ is a good reference?

Sample error

- Suppose that a database contains n individuals drawn uniformly at random from a population of size N>>>n.
- Suppose we are interested in a medical condition that affects a fraction p of the population.
- Then we expect the number of individuals in the dataset with condition p is

$$np \pm \Theta(\sqrt{n})$$

- The sampling error is of the order of $\sqrt{n}$.

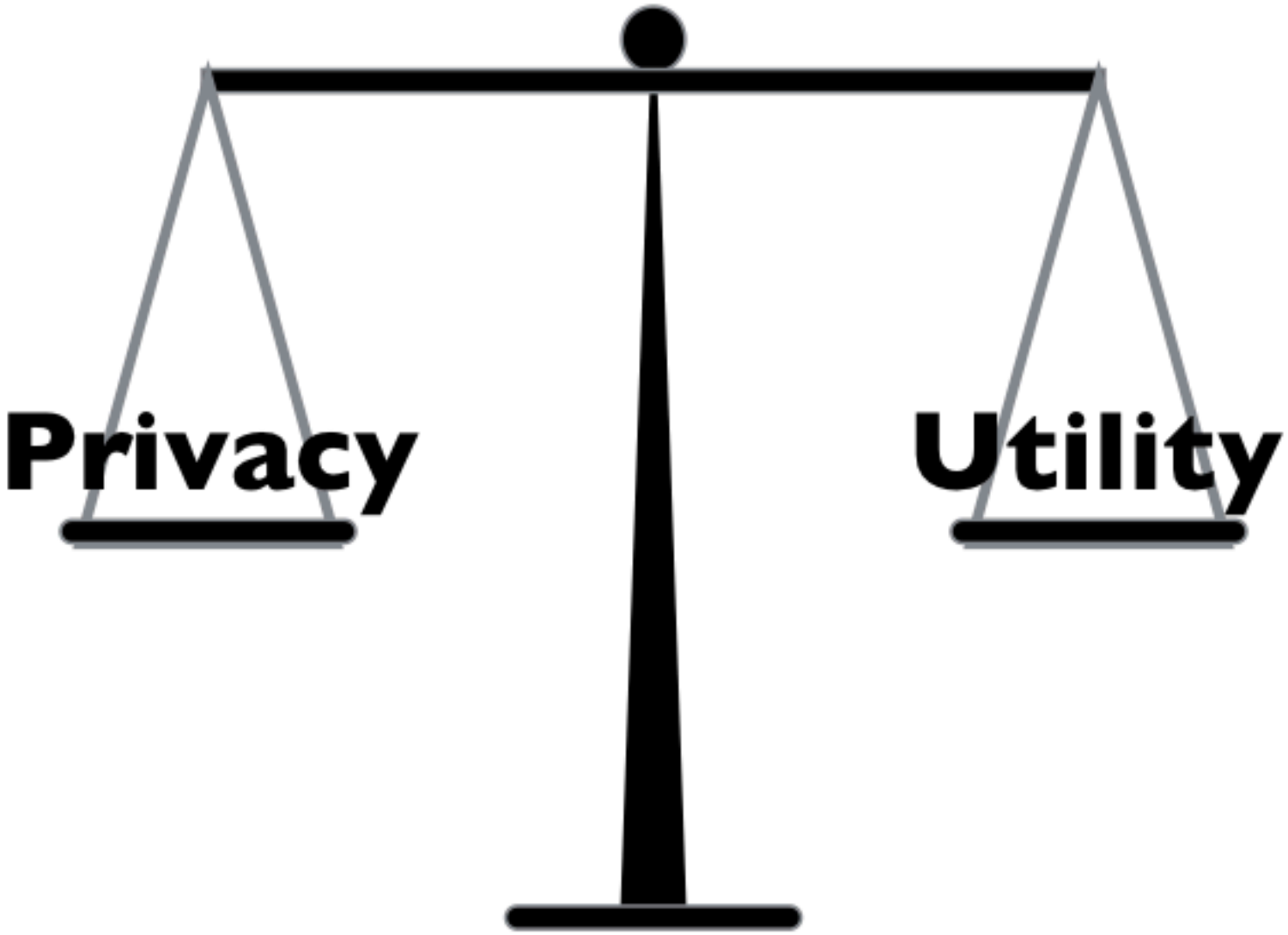
We would like the noise we introduce for privacy to be smaller than the sampling error.

Foundamental Law of Information Reconstruction

The release of too many overly accurate statistics gives privacy violations.

[DinurNissim02]

Privacy vs Utility



Privacy

Preventing blatantly non-privacy is a low bar for a privacy mechanism.
However, we should expect that the previous results apply to other stronger notions, in particular Differential Privacy.