Foundations of Privacy

Lecture 2

Resume of previous lecture

- Problem of statistical databases: we want to make available aggregate information, but without compromising the private data of the individual participating in the database

- This is not so easy to do. Naive deterministic methods, such as k-anonymity, are vulnerable to combination attacks

Example

- A medical database D1 containing correlation between a certain disease and age.
- Query: "what is the minimal age of a person with the disease"

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

D1 is **2-anonymous with respect to the query**. Namely, every possible answer partitions the records in groups of at least 2 elements

Alice	Bob
Carl	Don
Ellie	Frank

- A medical database D2 containing correlation between the disease and weight.
- Query: "what is the minimal weight of a person with the disease"

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Also D2 is 2-**anonymous**

Alice	Bob
Carl	Don
Ellie	Frank

k-anonymity is not compositional

Combine with the two queries:

minimal weight and the minimal age of a person with the disease

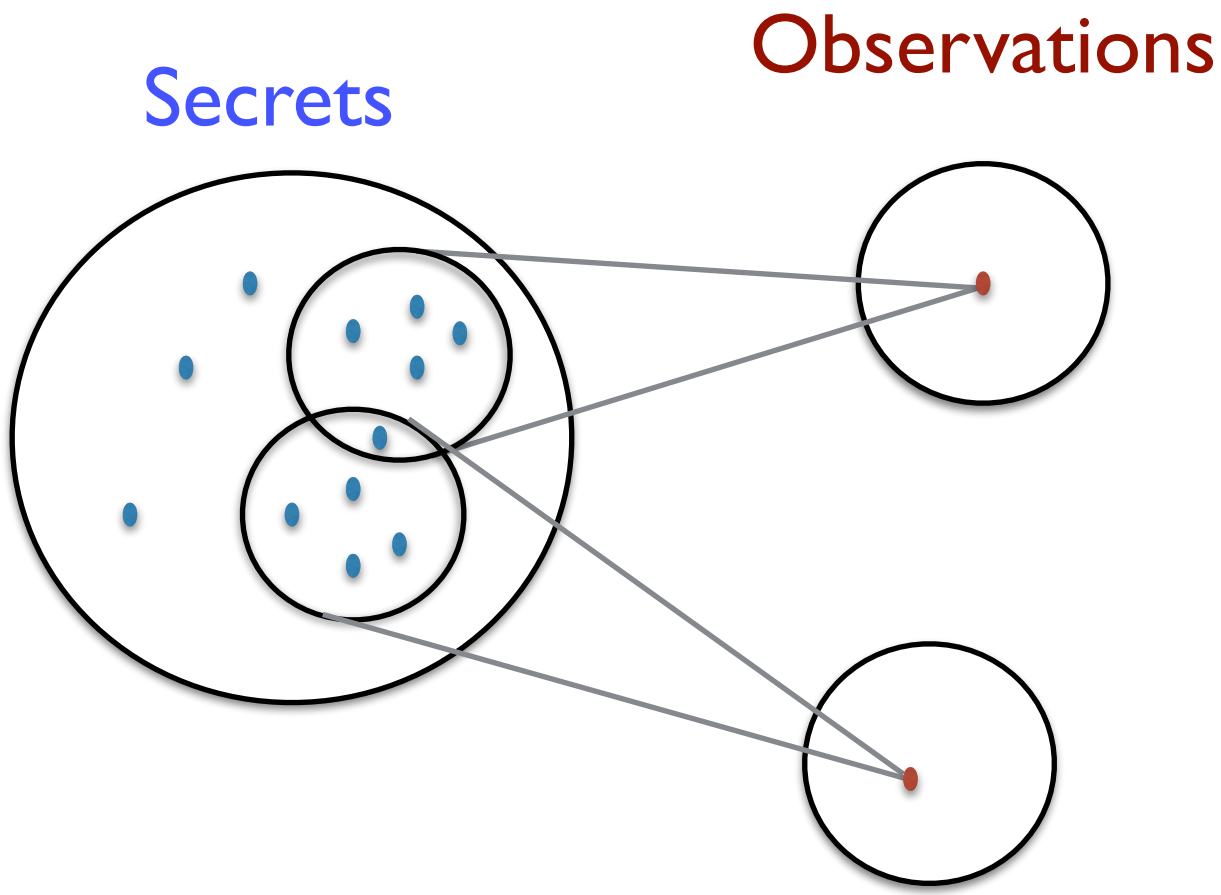
Answers: 40, 100

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

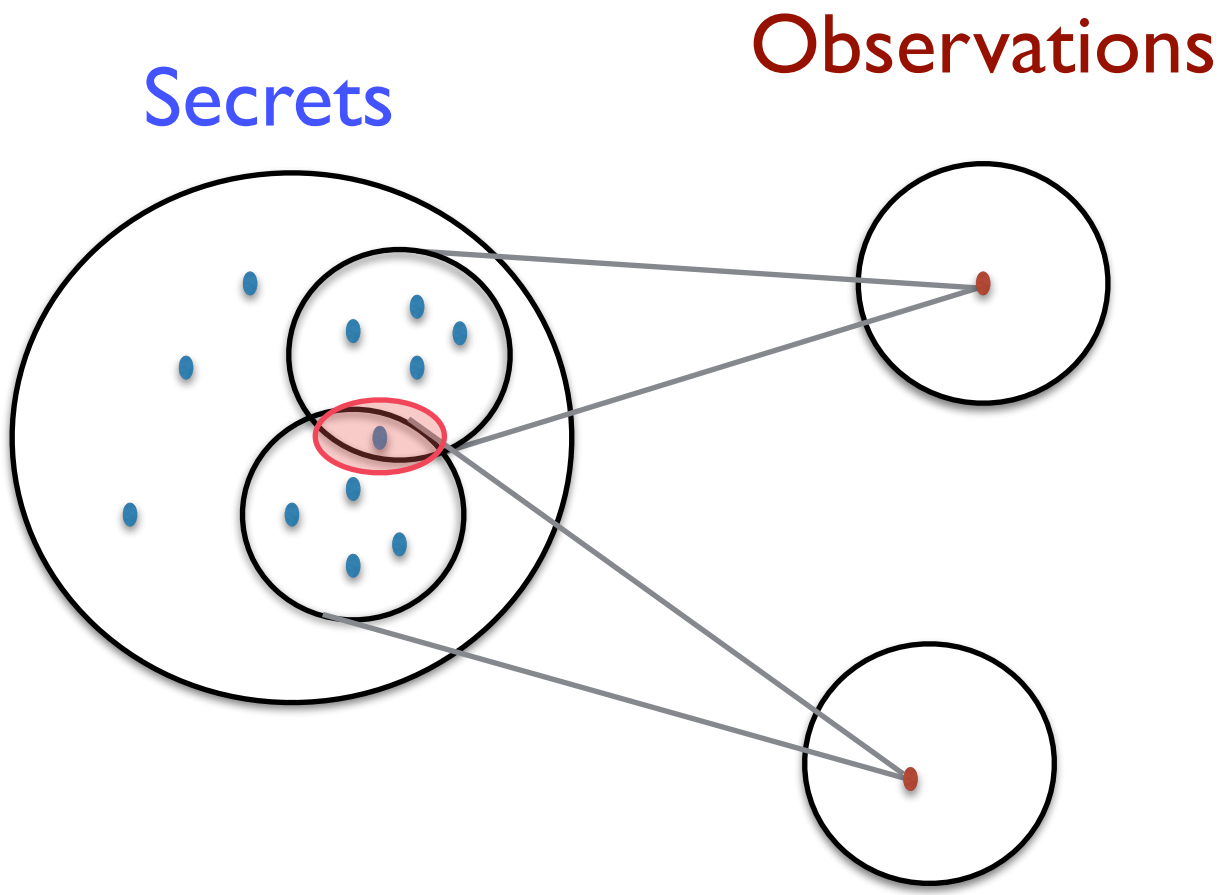
name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

This is a general problem of the deterministic approaches (based on the principle of many-to-one): the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



This is a general problem of the deterministic approaches (based on the principle of many-to-one): the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



A better solution

Introduce some probabilistic noise on the answer, so that the answers of minimal age and minimal weight can be given also by other people with different age and weight

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal age:

40 with probability 1/2

30 with probability 1/4

50 with probability 1/4

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal weight:
100 with prob. 4/7
90 with prob. 2/7
60 with prob. 1/7

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

Even if he combines the answers, the adversary cannot tell for sure whether a certain person has the disease

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Randomized mechanisms

- A randomized mechanism (for a certain query) reports an answer which is an approximation of the true answer and is generated randomly according to some probability distribution
- Randomized mechanisms are more robust to combination attacks than the deterministic ones
- However, we need to choose carefully the probability distribution, in order to get the desired degree of privacy, and in order to maintain a certain degree of utility for the query
- There is a trade-off between utility and privacy, but it is not strict: for a certain degree of privacy, one mechanism can give a better utility than another. It is therefore interesting to try to find the optimal mechanism (the mechanism with highest utility), among those that offer the desired degree of privacy.
- To solve the above problem, and more in general to reason about privacy and utility, we need formal, rigorous definitions of these notions.
- A definition of privacy that has become very popular: Differential Privacy [Cynthia Dwork, ICALP 2006]

Plan of the lecture

- The standard definition of Differential Privacy
- The Bayesian interpretation of DP
- Compositionality of DP
- The privacy budget
- Implementation of DP: Laplacian noise
- Examples and exercises

Databases

- V is a set whose elements represent all possible **values of the records** ($v \in V$ can be a tuple, i.e. it can be composed by various fields). We assume that V contains a special element $\perp$ representing a dummy record, or the absence of the corresponding record.
- A **database** of n records is an element of V^n. We will represent the databases by x, x_1, x_2, ...
- We assume a probability distribution π on the databases. We will indicate by X the corresponding random variable.
- Two databases x_1, x_2 are **adjacent** if they differ for exactly one record. We will indicate this property with the notation $x_1 \sim x_2$
  - $x_1 \sim x_2$ represent the fact that x_1 and x_2 differ for the information relative to an individual. Either this individual has been added to x_2, or he has been removed from x_2, or has changed value.
- The number of records in which two databases x_1, x_2 differ from each other is called "Hamming distance" between x_1, x_2.

Queries

- (The answer to) a query f can be seen as a function from the set of databases $\mathcal{X} = V^n$ to a set of values $\mathcal{Y}$. Namely,

$$f : \mathcal{X} \to \mathcal{Y}$$

- $y = f(x)$ is the **true answer** of the query f on the database x.
- For a given f, the distribution π on $\mathcal{X}$ also induces a distribution on $\mathcal{Y}$. We will denote by Y the random variable associated to the distribution on $\mathcal{Y}$.

Randomized mechanisms

- A randomized mechanism for the query f is any probabilistic function $\mathcal{K}$ from $\mathcal{X}$ to a set of values $\mathcal{Z}$. Namely,

$$\mathcal{K} : \mathcal{X} \to \mathcal{D}\mathcal{Z}$$

  where $\mathcal{D}\mathcal{Z}$ represents the set of probability distributions on $\mathcal{Z}$.

- $\mathcal{Z}$ does not necessarily coincide with $\mathcal{Y}$.

- z drawn from $\mathcal{D}(x)$ is a **reported answer** of the query $\mathcal{K}$ on the database x.

- Note that π and $\mathcal{K}$ induce a probability distribution also on $\mathcal{Z}$. We will denote by Z the random variable associated to this probability distribution

Differential Privacy

- We are now ready to define **Differential Privacy** for a randomized mechanism $\mathcal{K}$.
- Let us first consider the discrete case. Namely, $\mathcal{K}(x)$ is discrete, for every database x.
- **Definition (Differential Privacy)** $\mathcal{K}$ is ε-differentially private if for every pair of databases $x_1, x_2 \in \mathcal{X}$ such that $x_1 \sim x_2$, and for every $z \in Z$, we have:

$$p(Z = z|X = x_1) \leq e^{\varepsilon} p(Z = z|X = x_2)$$

  where $p(Z = z|X = x)$ represents the conditional probability of z given x, namely the probability that on the database x the mechanism reports the answer z
- This definition therefore means that the value (or the presence) of an individual does not affect significantly the probability of getting a certain reported value.

Differential Privacy

- Let us now consider the continuous case. Namely, $\mathcal{K}(x)$ is a probability density function on $\mathcal{Z}$. The only thing that changes is that we consider a measurable subset $\mathcal{S}$ of $\mathcal{Z}$ instead than a single z:
- **Definition (Differential Privacy)** $\mathcal{K}$ is ε-differentially private if for every pair of databases $x_1, x_2 \in \mathcal{X}$ such that $x_1 \sim x_2$, and for every measurable $\mathcal{S} \subseteq Z$, we have:

$$p(Z \in \mathcal{S} | X = x_1) \leq e^{\varepsilon} p(Z \in \mathcal{S} | X = x_2)$$

  where $p(Z \in \mathcal{S} | X = x)$ represents the probability that on the database x the mechanism reports an answer in $\mathcal{S}$

- This definition therefore means that the value (or the presence) of an individual does not affect significantly the probability that the reported value satisfy a certain property.

Independence from the prior

- The distribution π on the databases is called prior, meaning: *before* the reported answer
- π represents the knowledge that a potential adversary (aka user, in the case of DP) has about the database (before knowing the answer of $\mathcal{K}$)
- We note that the definition of DP does not depend on π. This is a very good property, because it means that we can design mechanisms that satisfy DP without taking the knowledge of the adversary into account: the same mechanism will be good for all adversaries.

Compositionality

- Differential privacy is compositional, namely: given two mechanisms $\mathcal{K}_1$ and $\mathcal{K}_2$ on $\mathcal{X}$ that are respectively ε_1 and ε_2-differentially private, their composition $\mathcal{K}_1 \times \mathcal{K}_2$ is $(\varepsilon_1 + \varepsilon_2)$-differentially private.

  Note: $\mathcal{K}_1 \times \mathcal{K}_2$ is defined by the following property: if $\mathcal{K}_1(x)$ reports z_1 and $\mathcal{K}_2(x)$ reports z_2, then $(\mathcal{K}_1 \times \mathcal{K}_2)(x)$ reports (z_1, z_2).

  Proof: exercise

- **Privacy budget**: An user is given an initial budget α. Each time he asks a query, answered by ε-differentially private mechanism, his budget is decreased by ε. When his budget is exhausted, he is not allowed to ask queries anymore.

Bayesian interpretation

- Let X_i be the random variable representing the value of the individual i, and let X_{others} be the random variable representing the value of all the other individuals in the database.
  Similarly, let x_i and x_{others} represent possible values for X_i and X_{others}.
  Note that (x_i, x_{others}) represents and element in $\mathcal{X}$.
  Analogously, let π_i represent the component of the prior distribution that concerns the value of the individual i.

- ε-differential privacy in the discrete case is equivalently characterized by the following property: For all $(x_i, x_{others}) \in \mathcal{X}$, for all $z \in Z$, and for all π_i,

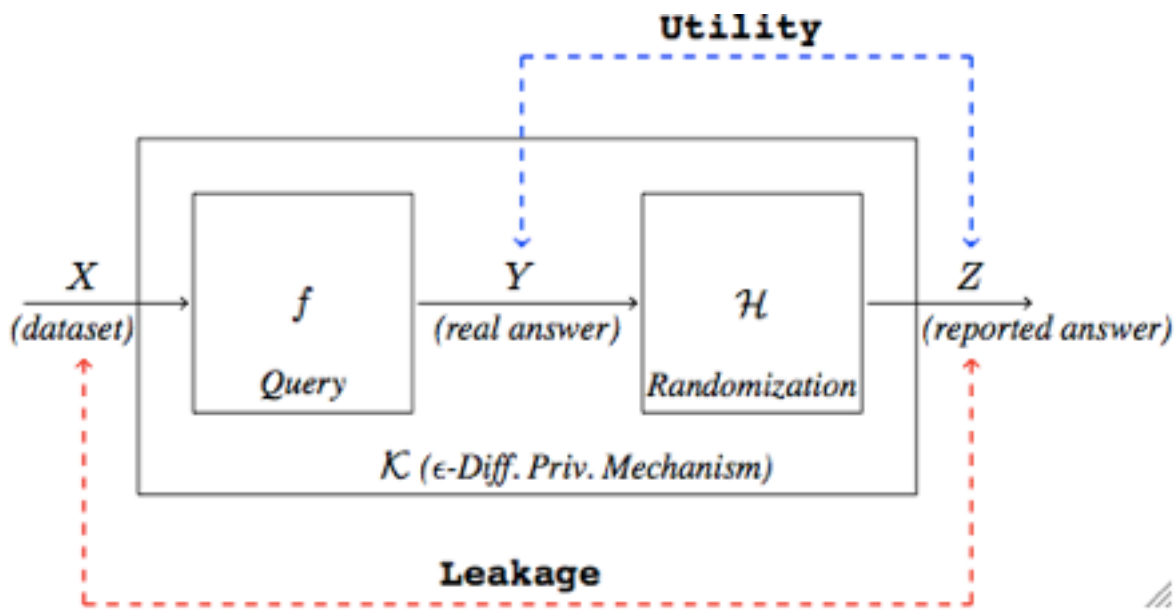
$$p(X_i = x_i | X_{others} = x_{others}, Z = z) \leq e^{\varepsilon} p(X_i = x_i | X_{others} = x_{others})$$

  Namely: assuming that the adversary knows the value of all the other individuals in the database, the reported answer does not increase significantly his probabilistic knowledge of the value of i, with respect to his prior knowledge

  Note: $p(X_i = x_i | X_{others} = x_{others})$ is called *prior* of x_i, and $p(X_i = x_i | X_{others} = x_{others}, Z = z)$ is called *posterior* of x_i.

Oblivious Mechanisms

- Given $f: \mathcal{X} \to \mathcal{Y}$ and $\mathcal{K}: \mathcal{X} \to \mathcal{Z}$, we say that $\mathcal{K}$ is oblivious if it depends only on $\mathcal{Y}$ (not on $\mathcal{X}$)
- If $\mathcal{K}$ is oblivious, it can be seen as the composition of f and a randomized mechanism $\mathcal{H}$ (noise) defined on the exact answers $\mathcal{K} = f \times \mathcal{H}$



- Privacy concerns the information flow between the databases and the reported answers, while utility concerns the information flow between the correct answer and the reported answer

A typical oblivious differentially private mechanism: Laplacian noise

- Randomized mechanism for a query $f: \mathcal{X} \rightarrow \mathcal{Y}$.
- A typical randomized method: **add Laplacian noise.** If the exact answer is y, the reported answer is z, with a probability density function defined as:

$$dP_y(z) = c\, e^{-\frac{|z-y|}{\Delta f}\varepsilon}$$

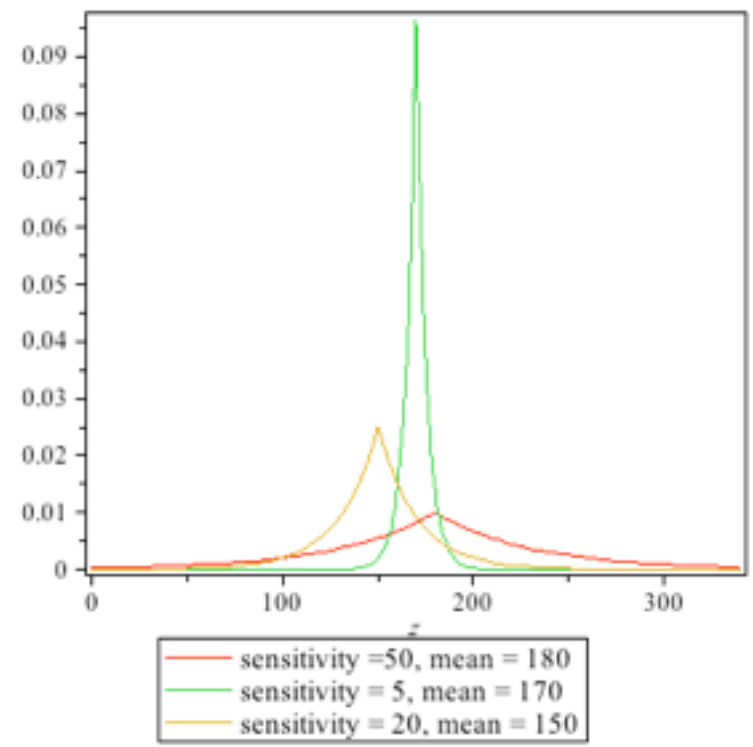
where Δf is the *sensitivity* of f:

$$\Delta f = \max_{x \sim x' \in \mathcal{X}} |f(x) - f(x')|$$

($x \sim x'$ means x and x' are adjacent, i.e., they differ only for one record)

and c is a normalization factor:

$$c = \frac{\varepsilon}{2\,\Delta f}$$



The geometric mechanism

- The geometric mechanism is a sort of discrete Laplacian.
- Assume that $\mathcal{Y}$ and $\mathcal{Z}$ are sets of integers. In the geometric mechanism, the probability distribution of the noise is:

$$p(z|y) = c\, e^{-\frac{|z-y|}{\Delta f}\varepsilon}$$

- where c is a normalization factor, defined so to obtain a probability distribution, and Δf is the sensitivity of query f

- Note that it does not make much sense to report answers outside $\mathcal{Y}$. If $\mathcal{Y}$ is an interval [a,b], we can truncate the mechanism, i.e., set $\mathcal{Z} = \mathcal{Y}$, and transfer on the extremes a and b all the probability that (according to the formula above) would fall outside the interval, to the left or to the right, respectively.

Counting Queries

- A counting query is a query of the form: How many individuals (tuples) in the database satisfy the property $\mathcal{P}$?

- The sensitivity of a counting query is 1

Some applications of DP

- The Census Bureau project *OnTheMap*, which allows to give researchers access to the data of the agency while protecting the privacy of the citizens
  http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data/

- Google' RAPPOR: Randomized Aggregatable Privacy Preserving Ordinal Response.
  Used for collecting statistics from end-user
  http://www.computerworld.com/article/2841954/googles-rappor-aims-to-preserve-privacy-while-snaring-software-stats.html

Exercises

1. Define the noise probability distribution for the geometric mechanism for a counting query when $\mathcal{Y}$ is the interval [0,n].
2. Define the truncated geometric mechanism for a counting query when $\mathcal{Y}$ and $\mathcal{Z}$ are the the interval [0,n].
3. Define the noise density for the Laplacian mechanism for a query "average height", assuming that the height of the population varies from 100 to 200 cm and that the database contains at least 10 elements.
4. Prove that the laplacian mechanism is ε-differentially private.
5. Prove that the geometric mechanism is ε-differentially private.
6. Prove that the truncated geometric mechanism is ε-differentially private.

Exercises

7. Prove that ε-differential privacy can be equivalently defined as follows

$\mathcal{K}$ is ε-differentially private if for every pair of databases $x_1, x_2 \in \mathcal{X}$ (not necessarily adjacents), and for every $z \in \mathcal{Z}$, we have:

$$p(Z = z|X = x_1) \leq e^{\varepsilon h(x_1,x_2)} p(Z = z|X = x_2)$$

where $h(x_1, x_2)$ represents the Hamming distance between x_1 and x_2