Identifiers and guesswork

Cédric Lauradoux

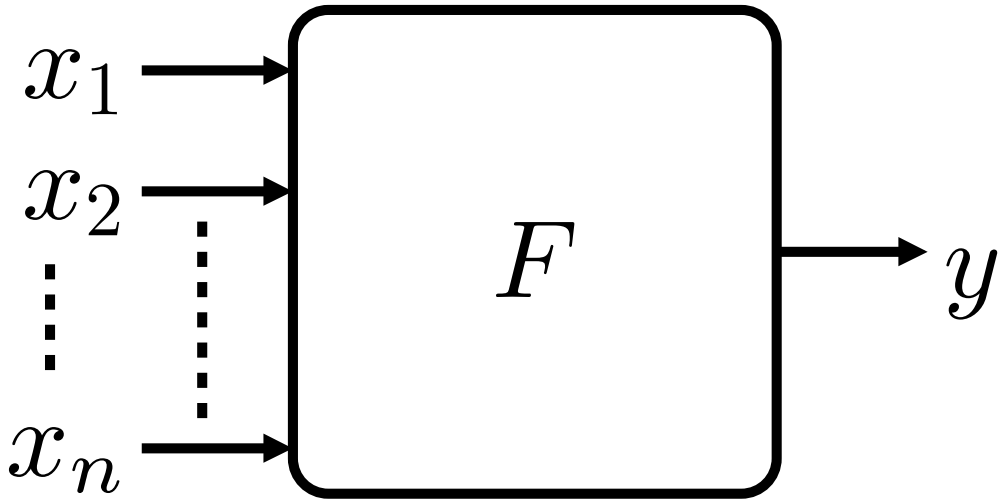
Identifiers and guesswork

Outline

- **Motivations**
  - *Identifiers are everywhere...*
  - *Tracking is a new business.*
- **Guesswork**
  - *Uniformity versus non-uniformity*
  - *Exhaustive search, dictionary.*
- **Examples**
  - *Age of the captain,*
  - *MAC addresses.*

Identifiers

- Let e be an entity with n attributes $x_1, x_2, \cdots, x_n$. y is an identifier of e computed using H.



- Widely used in computer science and everyday life !

Unique identifiers

- Let $e_1 = (x_1, \cdots, x_n)$ and $e_2 = (x'_1, \cdots, x'_n)$ be 2 *distinguishable entities*, *i.e.* $\exists i$ such that $x_i \neq x'_i$.

- Let y_1 and y_2 be the corresponding identifier of e_1 and e_2. y_1 and y_2 are unique identifiers if $y_1 \neq y_2$ for any distinguishable entities.

- Sometimes, *pseudo-uniqueness* : $\Pr(y_1 = y_2) = \epsilon$.

Private identifiers

- Knowing an identifier y, it is **not computationally feasible** to recover $(x_1, \cdots, x_n)$

- **No information leakage from an identifier.**
- We will focus on this problem in the talk !

Accountable identifiers

- Knowing an identifier y and **some trapdoor**, it is possible to recover $(x_1, \cdots, x_n)$.
- Recovering attributes (name and last name) may be required for $legal\ issues$.
- Privacy and accountability are clearly contradictory.

Breaking bad

- **Tracking** : learning your habits and mobility patterns.
- **Dedicated malwares :** identifiers can be used as a payload's trigger.
- **Data anonimization :** private identifiers are always thought as a great way to anonimized database. . .
- **Example :** *Mobilitics project*

Mobilitics project

Looking for information leakage

- INRIA-CNIL project on privacy on smarphone :
  – iOS,
  – Android.

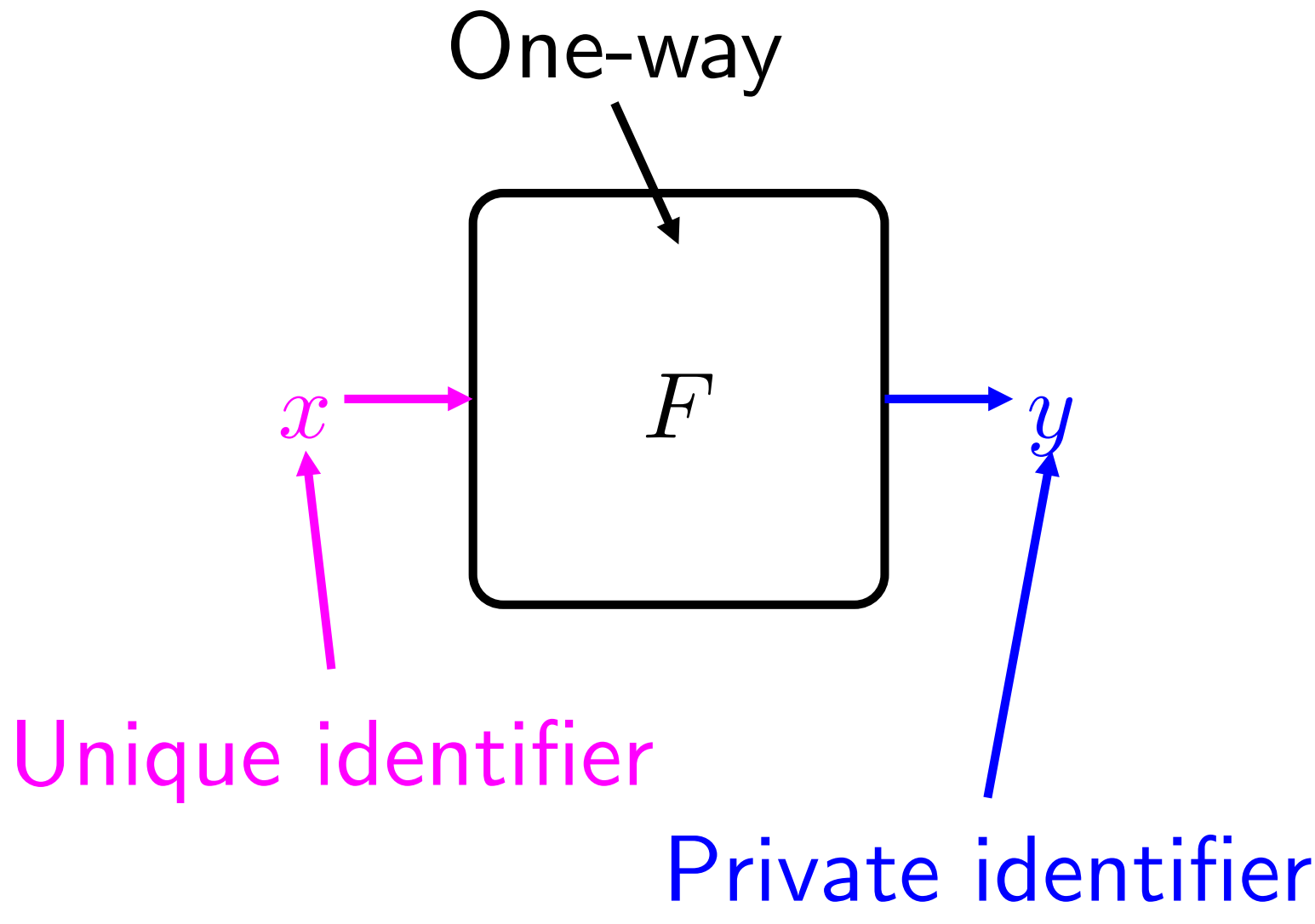
- **Goal :** *quantify the amount of information leaked by your smarphones apps.*

- `https://team.inria.fr/privatics/mobilitics/`

Private identifiers

- F must be **a one-way function**.
- F is $often$ a **cryptographic hash function** :
  – MD5
  – SHA-1
  – SHA-3. . .
- Is it enough for ensure privacy ?

Private identifiers



Point of view of Navizon

Hashed data can not be reverse engineered by a third party to reveal a devices MAC address. This means that anyone who gains access to the database directly from Amazon authorized or unauthorized ***will only see long strings of numbers and letters****. They would not be able to get any information that could be linked to a back to a particular mobile device owner.*

Problem

- **Given a private identifier recover the input !**

- **Attack strategies :**
  1. Invert the one-way function $= \mathit{full\ cryptanalysis}$.
  2. **Guess the input !**

Guesswork

- **Problem** – Guess the value of a discrete random variable X in one trial of a random experiment.

- **Idea** – Ask questions of the form :
  "Did X take on its i-th possible value ?"
  until the answer is "Yes !". (n choices)

- Let G be the **number of guesses needed** to recover the value. We want to minimize $E(G)$.

Cryptography

- **Values to guess :**
  - secret keys,
  - plaintext,
  - source of RNGs.

- **Uniformly distributed :** worst case for guesswork.

- **Metrics :** Shannon entropy, min-entropy.

Security

- **Values to guess :** **passwords !**
- Non-uniformly distributed.
- **Metrics :** many.

Cryptographic guesswork

- **Renyi entropy :**

$$H_\alpha(X) = -\frac{1}{1-\alpha}\sum_x \mathbf{Pr}[X = x]^\alpha.$$

- **Shannon's entropy :**

$$H(X) = -\sum_x \mathbf{Pr}[X = x]\log_2 \mathbf{Pr}[X = x].$$

- **Min-entropy :**

$$H_\infty(X) = -\log_2(\max_x \ \mathbf{Pr}[X = x]).$$

Cryptographic guesswork

Exhaustive search

- **Precomputation** – None
- **Memory** – None
- **Online search** – $E(G) = \frac{n-1}{2}$

Cryptographic guesswork

Dictionary

- **Precomputation** – n
- **Memory** – n
- **Online search** – $E(G) = 1$
- Let's forget about dictionary attack and TMTO. . .

$$\mathbf{W} = G \times M.$$

Information Theory Interlude

$$
\begin{aligned}
H(Y) &\leq H(X_1, X_2, \cdots X_n) \\
&\leq \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) \\
&\leq \sum_{i=1}^{n} H(X_i) \text{ if the } X_i \text{ are independent.}
\end{aligned}
$$

Non-uniform case

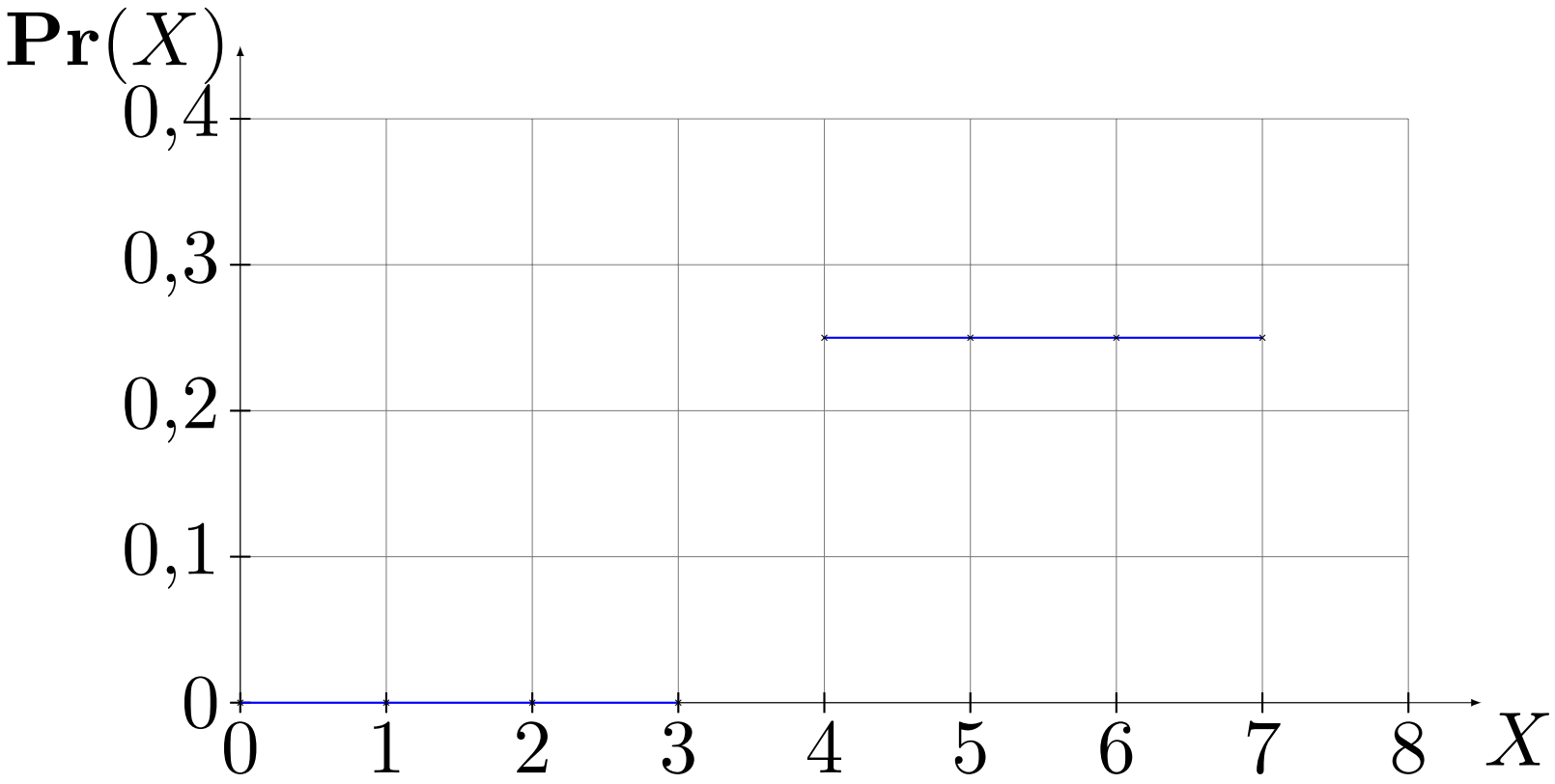
- Let us assume that the values are not uniformly distributed.

- **Question :** can we improve cryptographic methods ?

- In other words, can we achieve

$$E(G) \geq \frac{2^{H(X)} + 1}{2}$$

Uninteresting non-uniform case



- In this case, $n = 8$:

$$E(G) = \frac{2^{H(X)} + 1}{2} = \frac{5}{2}.$$

Non-uniform case

Alternative

- **Enumerate** the possible values of X in order of decreasing probability.

- Let $\mathbf{p} = (p_1, p_2, p_3, \cdots, p_n)$ be a monotone distribution :

$$p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_n$$

- $E(G) = \sum_{i=1}^{n} i \cdot p_i$.

Non-uniform case

- **Lower bound :** [Massey ISIT'94]

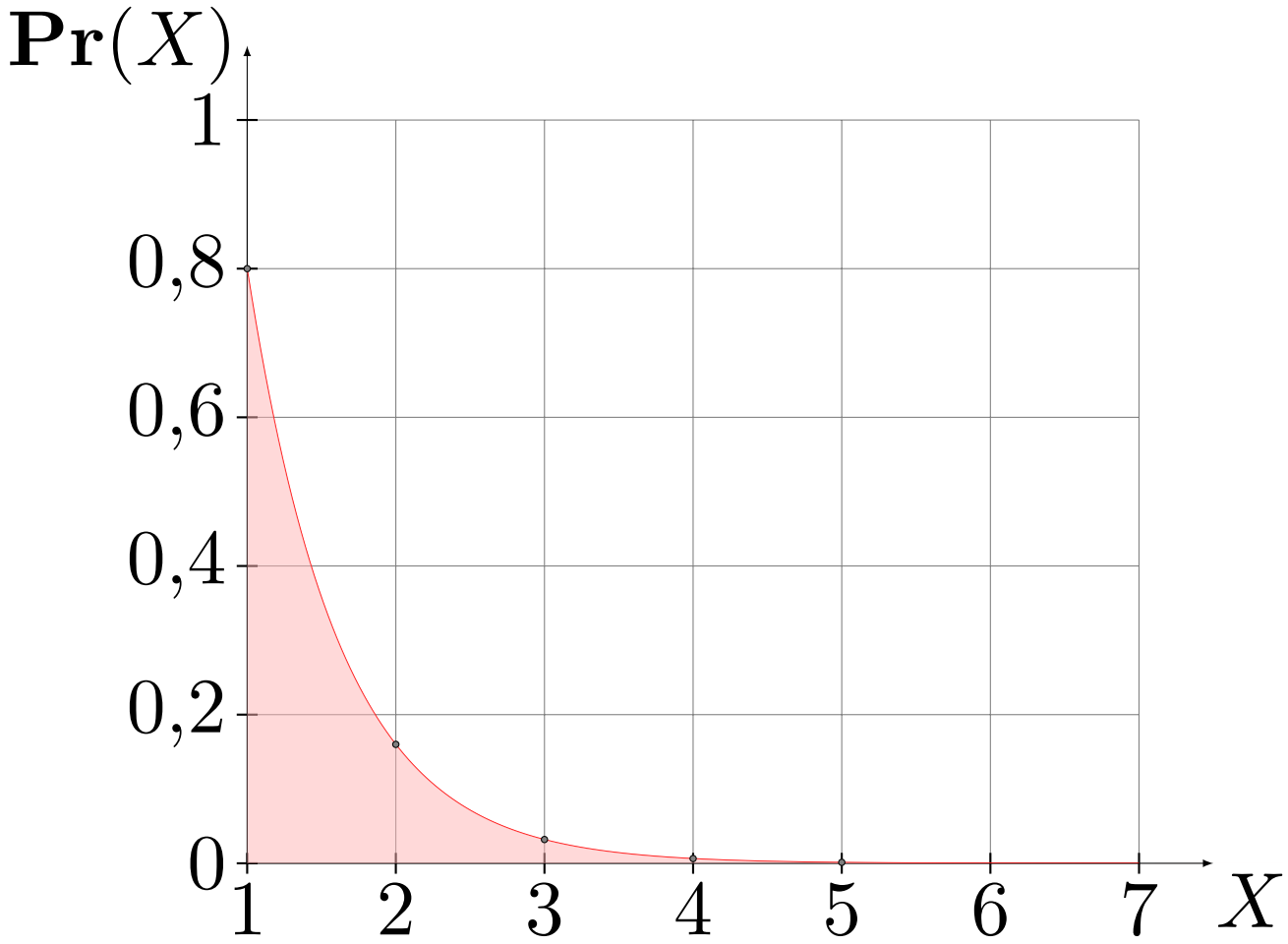
$$E(G) \geq \frac{2^{H(X)}}{4} + 1.$$

Holds for $n \to \infty$, $H(X) > 2$.

- **No trivial upper bound.** [Massey ISIT'94]

Geometric distribution

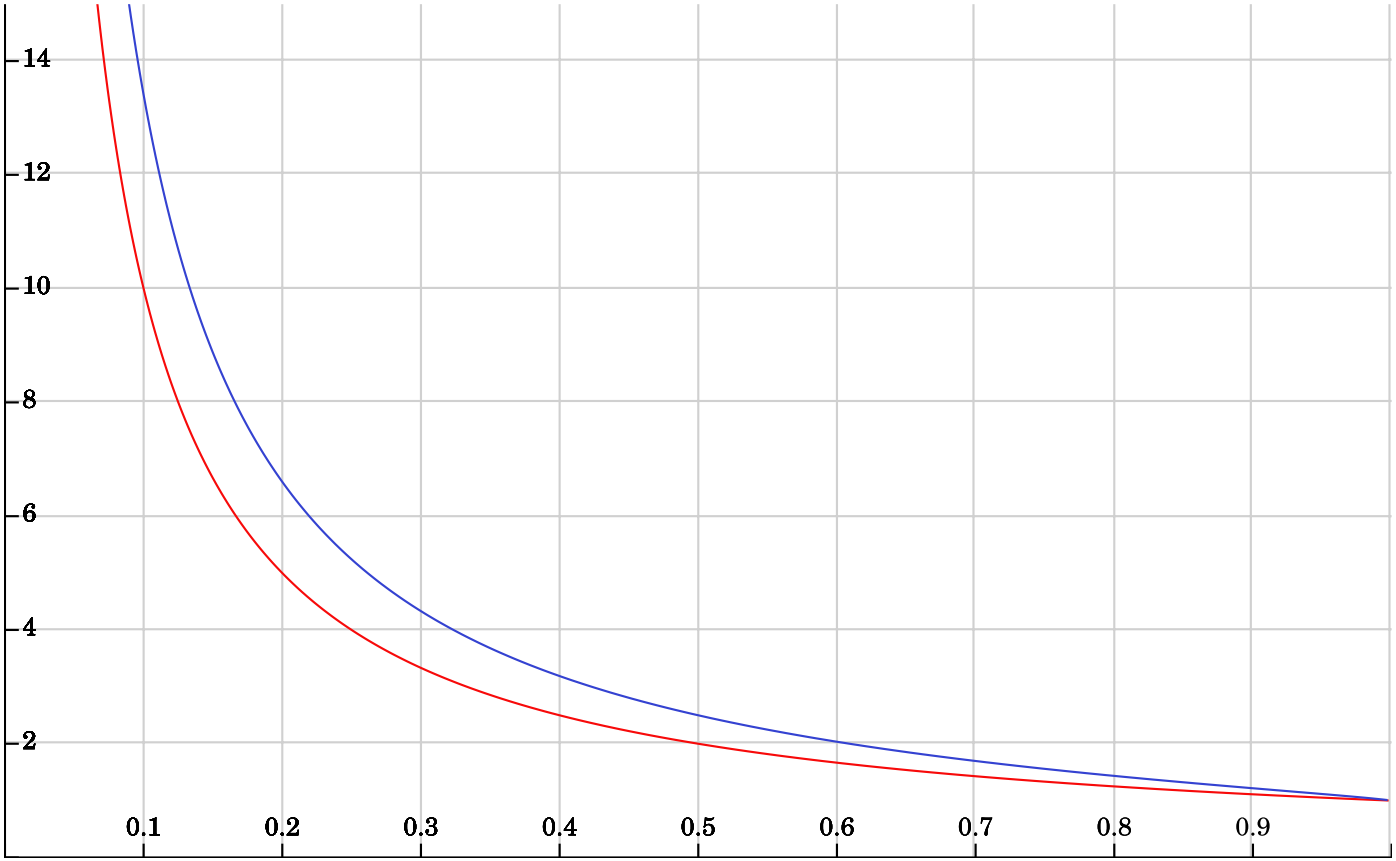
- $\mathbf{Pr}(X = k) = (1 - p)^{k-1}p$



Geometric distribution

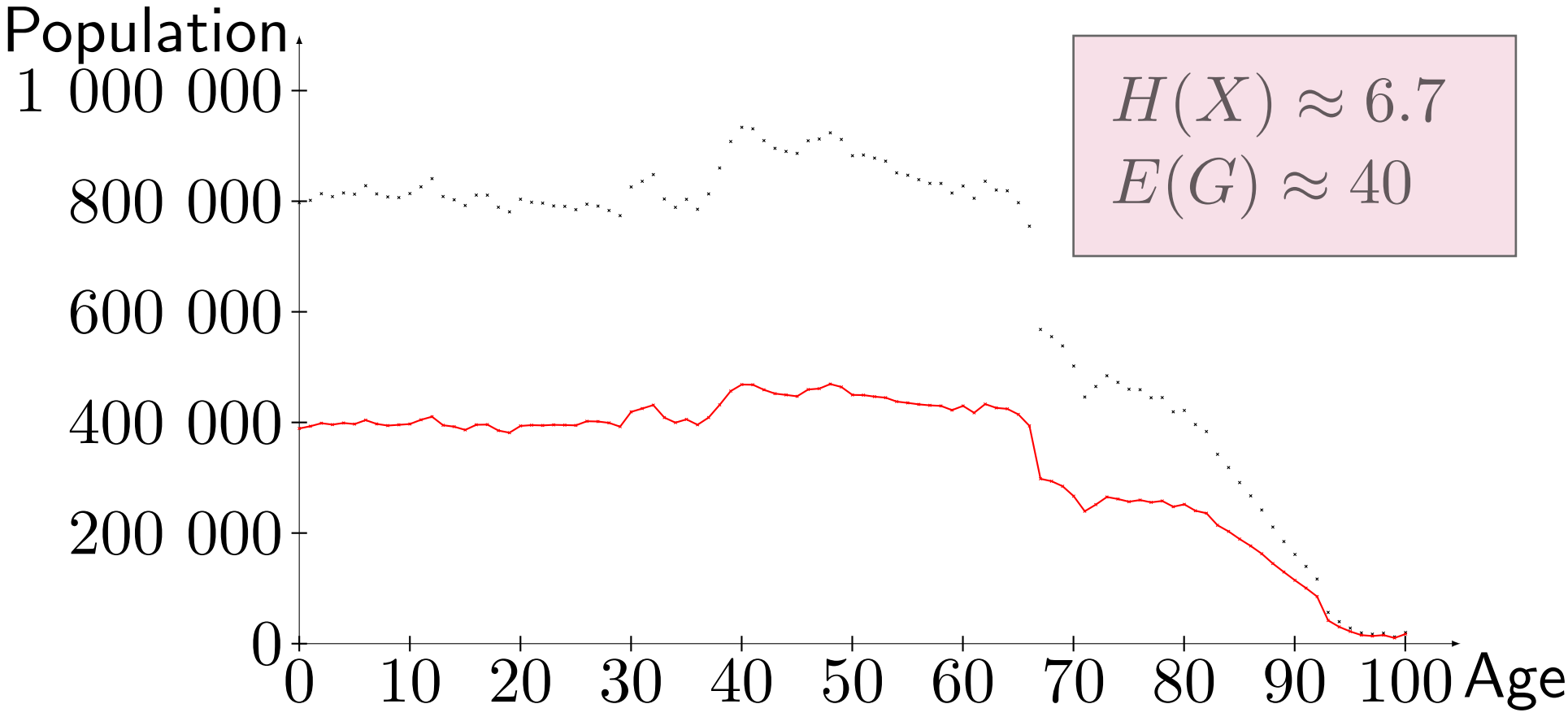
- $E(G) = 1/p$
- $H(X) = \frac{-(1-p)\log_2(1-p) - p\log_2(p)}{p}$

Geometric distribution



Guessing the age of French citizens

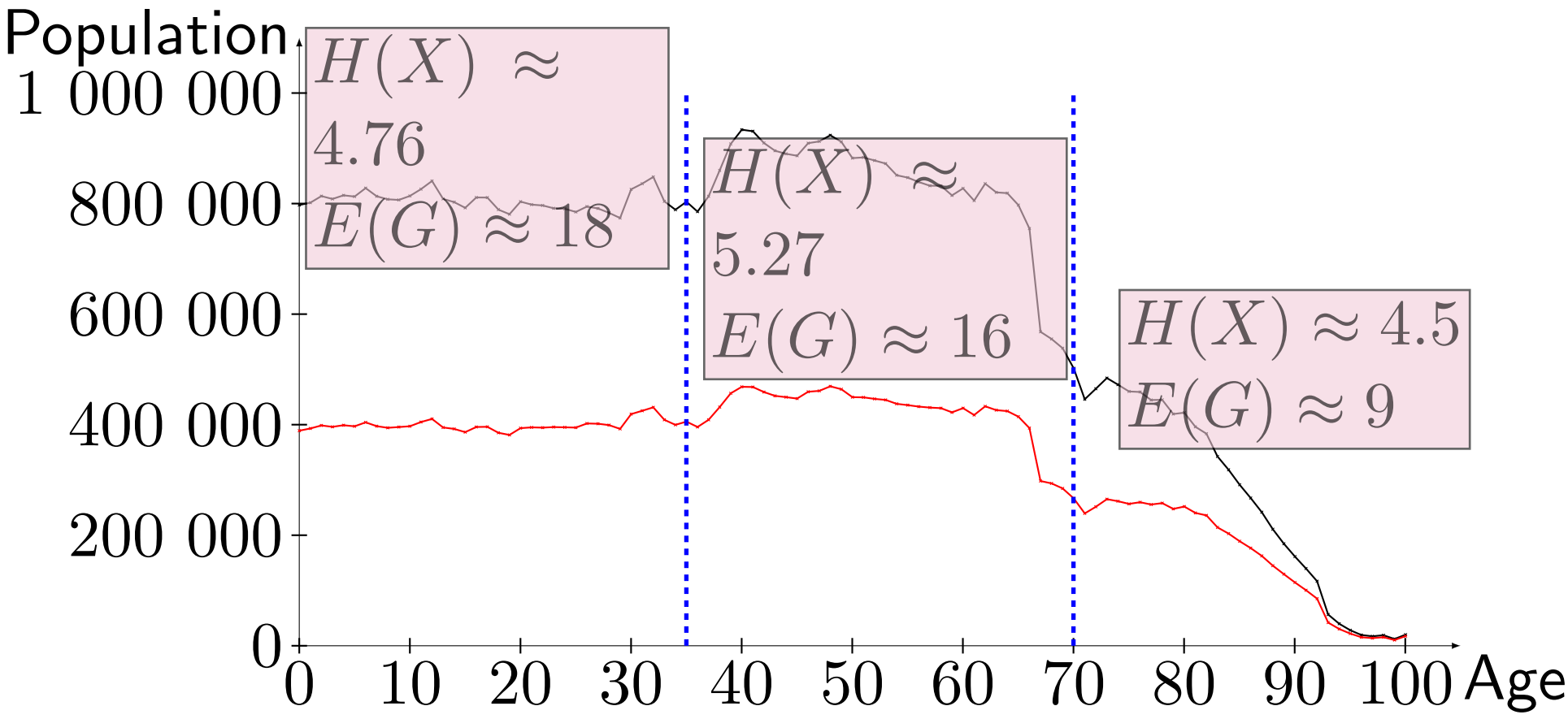
- Source INSEE `http://tinyurl.com/qhplomp`



Guessing the age of French citizens

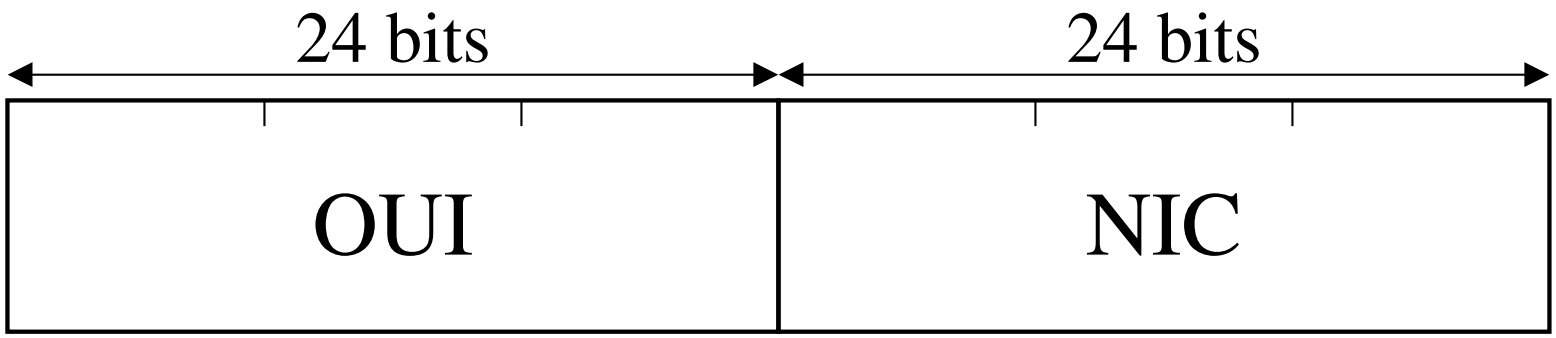
with a side information

- Side information : age belongs to 1 of the 3 sets



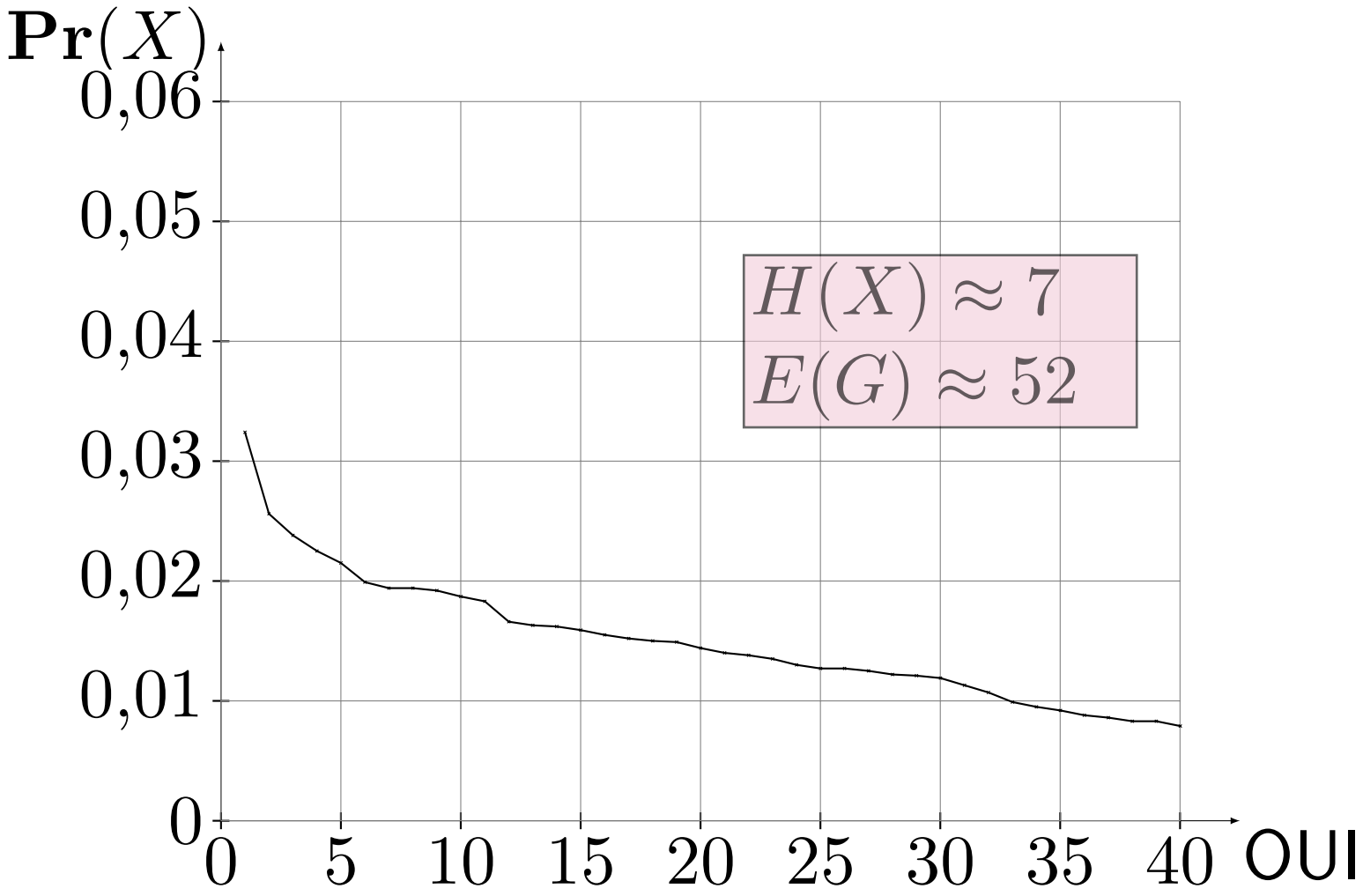
Guessing MAC address

- **Exhaustive search :** $2^{48} \approx 1$ day using $Hashcat$
- **Structure of a MAC address :**

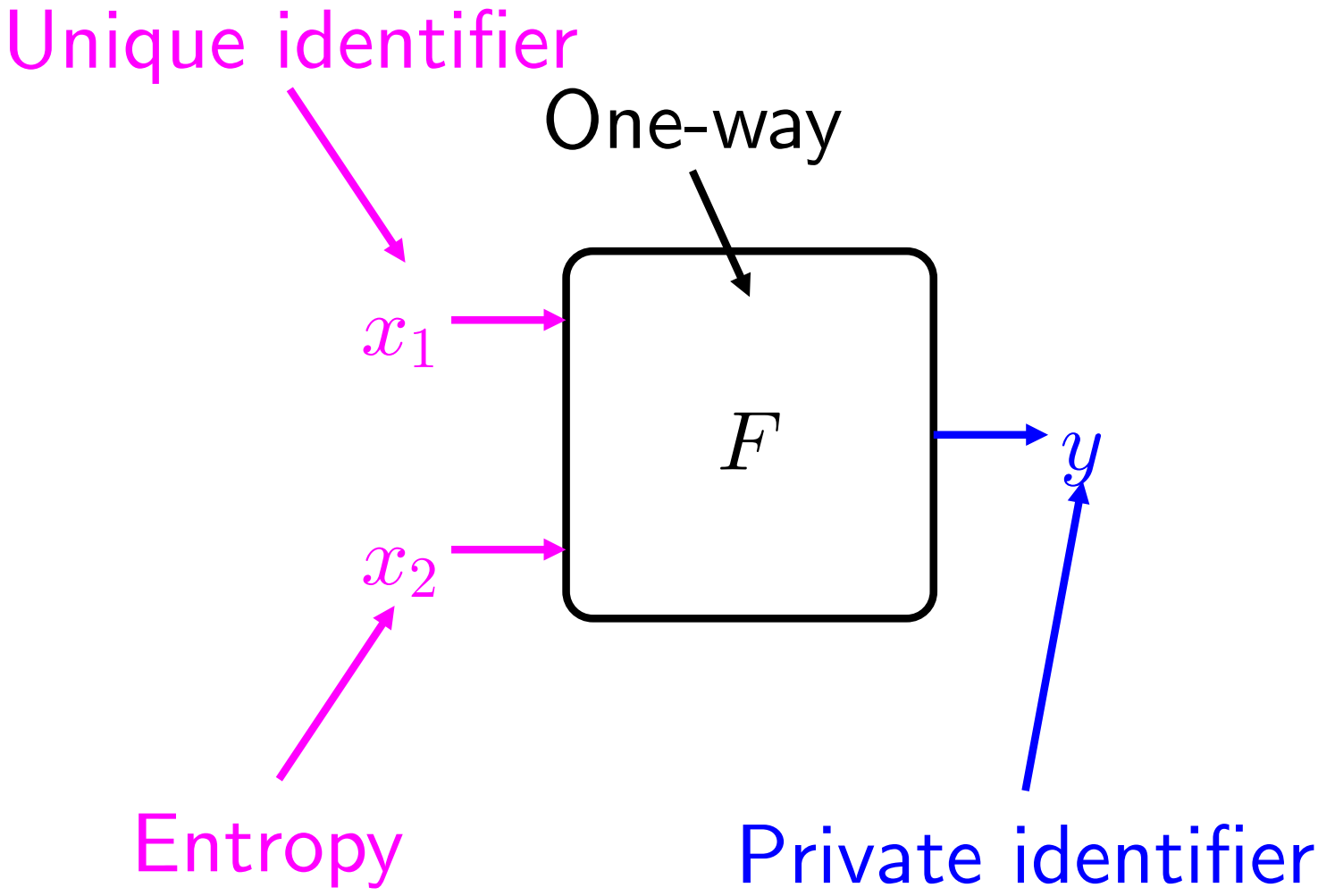


  - **OUI :** Organizationally Unique Identifier
  - **NICS :** Network Interface Controller Specific

OUI prefix


$\mathbf{Pr}(X)$
0,06
0,05
0,04
0,03
0,02
0,01
0
$H(X) \approx 7$
$E(G) \approx 52$
0
5
10
15
20
25
30
35
40
OUI

How to make good private identifiers ?



Timestamp ?

- **Using timestamp is not a great idea :**
  - Posix time : 32 bits.
  - In a year, you have 2^{25} seconds : we can conclude yourself !

- We need true random numbers !