ΥΣ13 - Computer Security

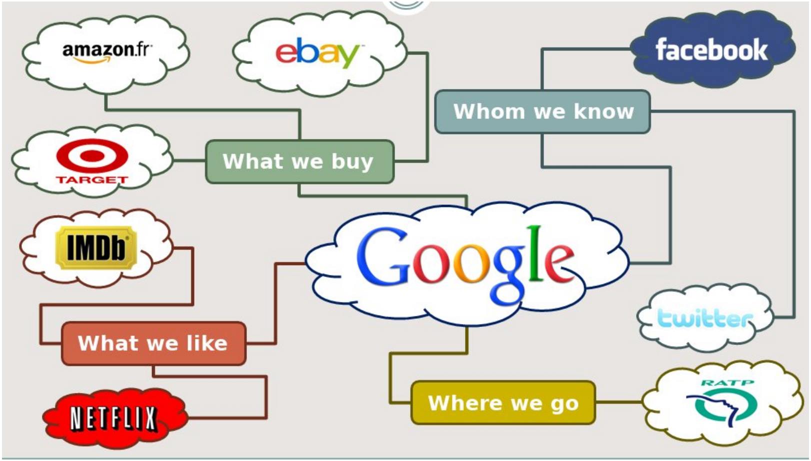
Data Privacy

Κώστας Χατζηκοκολάκης

Context

- Data is everywhere
- Can be exploited for numerous purposes:
  - Medical research
  - Transportation
  - Business insights
  - Policy, planning
  - Public safety
  - Weather prediction
  - Energy allocation
  - …
- **But** : always a privacy risk

Context


amazon.fr
ebay
facebook
Whom we know
TARGET
What we buy
IMDb
Google
twitter
What we like
RATP
Where we go
NETFLIX

The “nothing to hide” argument

Eric Schmidt on privacy:

“If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.”

Counter-arguments?

The “nothing to hide” argument

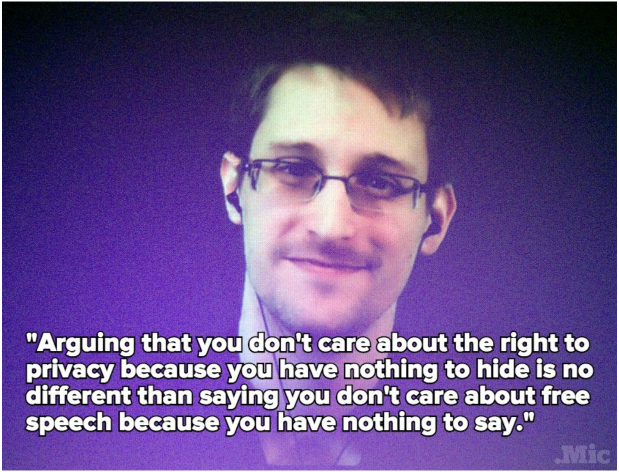
Eric Schmidt on privacy:

“If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place.”

Counter-arguments?

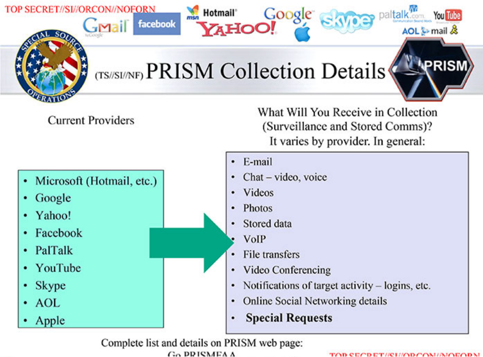
- “Obvious” things
- Surveillance
- Control, exclusion
- Errors, carelesness, guilty by association, social norms, …

The “nothing to hide” argument



"Arguing that you don't care about the right to privacy because you have nothing to hide is no different than saying you don't care about free speech because you have nothing to say."

Are these concerns relevant in the real world?



TOP SECRET//SI//ORCON//NOFORN

GMail
facebook
Hotmail
msn
YAHOO!
Google
skype
paltalk.com
YouTube
AOL mail
SPECIAL SOURCE OPERATIONS
PRISM

(TS//SI//NF) PRISM Collection Details

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

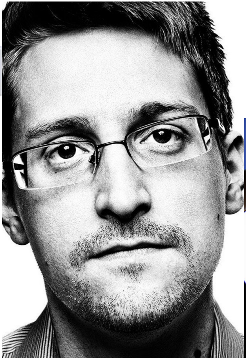
What Will You Receive in Collection (Surveillance and Stored Comms)? It varies by provider. In general:

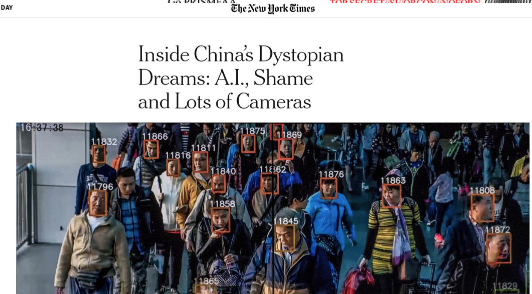
- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:

Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN





DAY

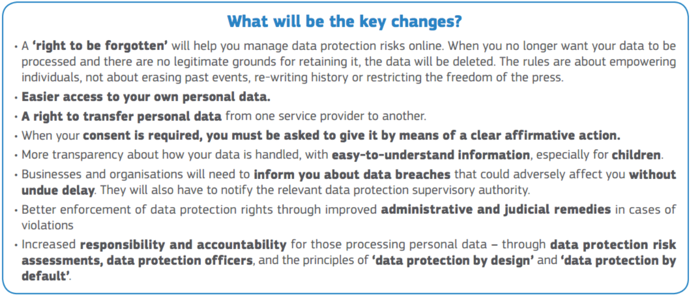
The New York Times

Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras

General Data Protection Regulation (GDPR)

About this site | Contact | Cookies | Legal notice
English (en)
European Commission
JUSTICE
Building a European Area of Justice
European Commission > Justice > Data protection > reform
HOME
ALL TOPICS
Search
DATA PROTECTION
Reform of the data protection legal framework

Reform of EU data protection rules



What will be the key changes?

- A **'right to be forgotten'** will help you manage data protection risks online. When you no longer want your data to be processed and there are no legitimate grounds for retaining it, the data will be deleted. The rules are about empowering individuals, not about erasing past events, re-writing history or restricting the freedom of the press.
- **Easier access to your own personal data.**
- **A right to transfer personal data** from one service provider to another.
- When your **consent is required, you must be asked to give it by means of a clear affirmative action.**
- More transparency about how your data is handled, with **easy-to-understand information**, especially for **children**.
- Businesses and organisations will need to **inform you about data breaches** that could adversely affect you **without undue delay**. They will also have to notify the relevant data protection supervisory authority.
- Better enforcement of data protection rights through improved **administrative and judicial remedies** in cases of violations
- Increased **responsibility and accountability** for those processing personal data – through **data protection risk assessments, data protection officers**, and the principles of **'data protection by design'** and **'data protection by default'**.

The problem of privacy

- In general, the problem of privacy is to protect the disclosure of **sensitive information** of individuals when a collection of data about these individuals (*dataset*) is made **publicly available**
- The process of transforming the dataset in order to avoid such disclosure is called **sanitization**

Privacy via anonymization

Nowadays, many institutions and companies that collect data use anonymization, i.e., they remove all personal identifiers: name, address, SSN, …

"We don't have any raw data on the identifiable individual. Everything is anonymous"
(CEO of NebuAd, a U.S. company that offers targeted advertising based on browsing histories)

Similar practices are used by Facebook, MySpace, Twitter, …

Privacy via anonymization

GDPR, Art. 5

Purpose specification (‘personal data shall be collected for **specified**, explicit and legitimate **purposes** and **not further processed** in a manner that is incompatible with those purposes’)

GDPR, Art. 6(4)

[...] the existence of appropriate **safeguards**, which may include encryption or **pseudonymisation**.’)

Privacy via anonymization

However, anonymity-based sanitization has been shown to be **highly ineffective:** Several de-anonymization attacks have been carried out in the last decade

- The quasi-identifiers allow to retrieve the identity in a large number of cases.
- More sophisticated methods (k-anonymity, ℓ-diversity, …) take care of the quasi-identifiers, but they are still prone to composition attacks

Naive anonymization

- This is the most obvious solution: remove the identity of individuals from the database, so that the sensitive information cannot be directly linked to the individual

- Example: assume that we have a medical database, where the sensitive information is disease that has been diagnosed

- For instance, Jorah Mormont may not want to reveal that he is affected by greyscale.

	Name	age	Disease
1	Jon Snow	30	cold
2	Jamie Lannister	39	amputed hand
3	Arya Stark	16	stomac ache
4	Bran Stark	14	crippled
5	Sandor Clegane	45	ignifobia
6	Jorah Mormont	48	gleyscale
7	Eddad Stark	32	headache
8	Ramsay Bolton	32	psychopath
9	Daenerys Targaryen	25	mania of grandeur

Naive anonymization

- Anonymization removes the column of the name, so that, for instance, the grayscale disease cannot be directly linked to Jorah Mormont

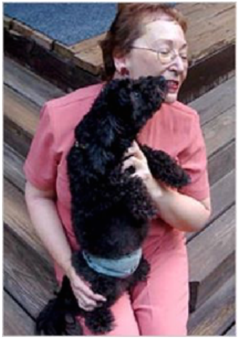
- Hystorically the first method, still used nowadays

- However, this solution has been (already several years ago) shown to be very weak and prone to de-anonymization attacks

	Name	age	Disease
1	-	30	cold
2	-	39	amputed hand
3	-	16	stomac ache
4	-	14	crippled
5	-	45	ignifobia
6	-	48	gleyscale
7	-	32	headache
8	-	32	psychopath
9	-	25	mania of grandeur

Famous deanonymization attacks : AOL

- In 2006, AOL Research released a text file containing twenty million search keywords for over 650,000 users, intended for research purposes.
- The file was anonymized (names where substituted by numbers as pseudonyms), but personally identifiable information was present in many of the queries. The NYT was able to locate an individual from the search records by cross referencing them with phonebook listings
- **From the report:** The subject conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 y.o. single men" to "dog that urinates on everything.", "landscapers in Lilburn, Ga", several people with the last name Arnold and "homes sold in shadow lake". It did not take much to identify the subject as Thelma Arnold, a 62-year-old widow with three dogs who lives in Lilburn, Ga.



Naive anonymization

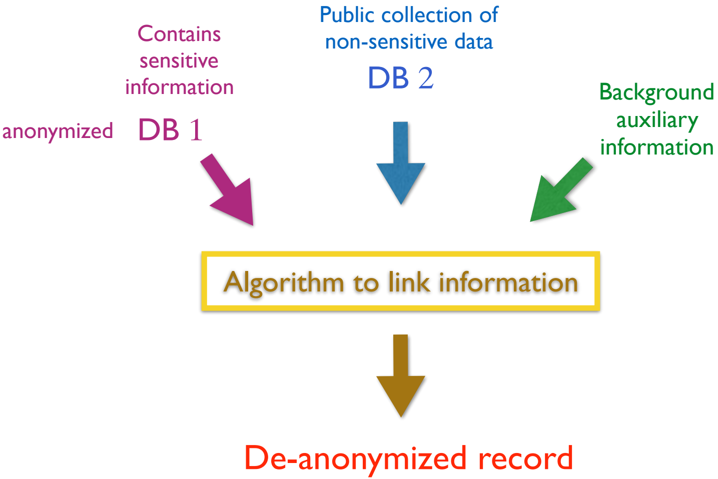
GDPR, Recital 26

"…determine **whether a natural person is identifiable**, account should be taken of **all the means reasonably likely to be used**, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."

GDPR, Art. 11(2)

"where, ..., the controller is **able to demonstrate that it is not in a position to identify the data subject**, the controller shall inform the data subject accordingly, if possible. In such cases, Articles 15 to 20 shall not apply ..."

Famous deanonymization attacks : Medical records



Contains sensitive information

anonymized DB 1

Public collection of non-sensitive data

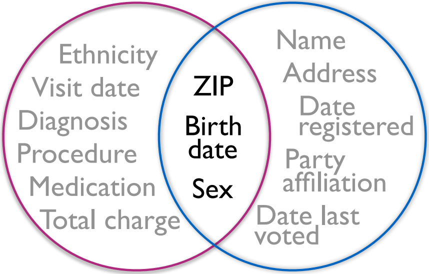
DB 2

Background auxiliary information

Algorithm to link information

De-anonymized record

Famous deanonymization attacks : Medical records



DB 1: Medical data DB 2:Voter list

87 % of US population is uniquely identifiable by 5-digit ZIP, gender, DOB

This attack has lead to the proposal of k-anonymity (that I will present later)

k-anonymity [Sweeney, Samarati, 2000]

- Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals

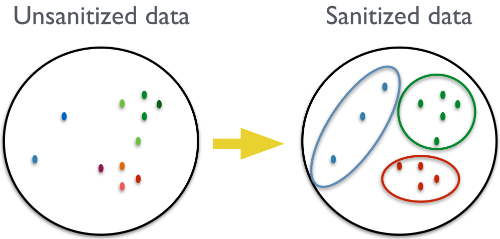
- Make every record in the table indistinguishable from a least $k-1$ other records with respect to quasi-identifiers. This can be done by:
    - suppression of attributes, and/or
    - generalization of attributes, and/or
    - addition of dummy records

- Linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

k-anonymity [Sweeney, Samarati, 2000]

Principle : group anonymity

- Ensure that each individual is indistinguishable within a group by removing individual differences

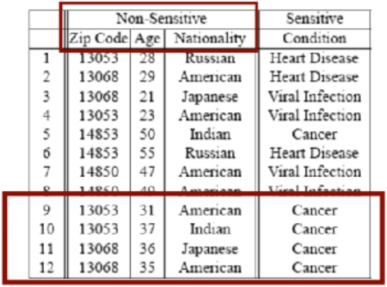


- Of course, the larger are the groups, the better the individuals are protected (within the group)
- k-anonymity ensure that the size of each group is at least k

k-anonymity [Sweeney, Samarati, 2000]

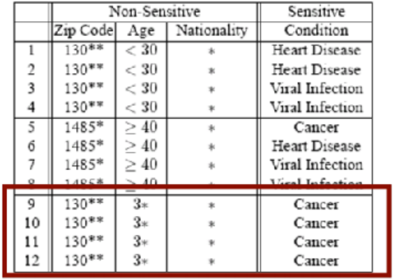
Example: 4-anonymity w.r.t. the quasi-identifiers (nationality, ZIP, age)

- achieved by suppressing the nationality and generalizing ZIP and age



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

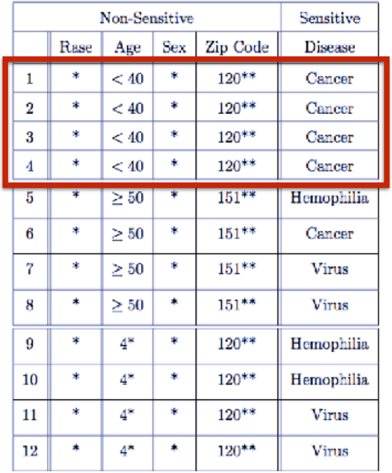


	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Problems with *k*-anonymity

- **Problem**: in the sanitized dataset, all the individual in a group may the same value for the sensitive data
- Clearly, the people in that group are not protected from the revelation of their disease
- **Example:** suppose that John's employer knows that John is less than 40, that he lives in a town with ZIP code 12032, and that he visits the hospital. He can learn that John has cancer.

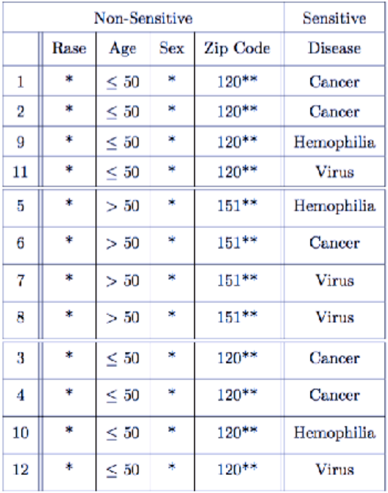


	Non-Sensitive				Sensitive
	Rase	Age	Sex	Zip Code	Disease
1	*	< 40	*	120**	Cancer
2	*	< 40	*	120**	Cancer
3	*	< 40	*	120**	Cancer
4	*	< 40	*	120**	Cancer
5	*	≥ 50	*	151**	Hemophilia
6	*	≥ 50	*	151**	Cancer
7	*	≥ 50	*	151**	Virus
8	*	≥ 50	*	151**	Virus
9	*	4*	*	120**	Hemophilia
10	*	4*	*	120**	Hemophilia
11	*	4*	*	120**	Virus
12	*	4*	*	120**	Virus

Table 2: 4-anonymous inpatient microdata.

l-diversity [Kifer et at, 2007]

- A solution: ℓ-diversity.
- The idea is to form the groups in such a way that each group contains a variety of values for the sensitive data
- It’s computationally heavy: To find the optimal solution is a combinatorial problem with exponential complexity

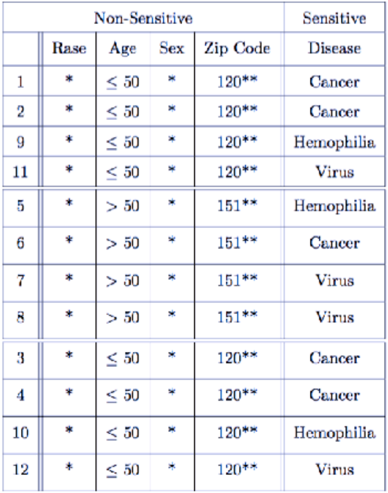


	Non-Sensitive				Sensitive
	Rase	Age	Sex	Zip Code	Disease
1	*	≤ 50	*	120**	Cancer
2	*	≤ 50	*	120**	Cancer
9	*	≤ 50	*	120**	Hemophilia
11	*	≤ 50	*	120**	Virus
5	*	> 50	*	151**	Hemophilia
6	*	> 50	*	151**	Cancer
7	*	> 50	*	151**	Virus
8	*	> 50	*	151**	Virus
3	*	≤ 50	*	120**	Cancer
4	*	≤ 50	*	120**	Cancer
10	*	≤ 50	*	120**	Hemophilia
12	*	≤ 50	*	120**	Virus

l-diversity

[Kifer et at, 2007]

- A solution: ℓ-diversity.
- The idea is to form the groups in such a way that each group contains a variety of values for the sensitive data
- It's computationally heavy: To find the optimal solution is a combinatorial problem with exponential complexity



Non-Sensitive					Sensitive
	Rase	Age	Sex	Zip Code	Disease
1	*	≤ 50	*	120**	Cancer
2	*	≤ 50	*	120**	Cancer
9	*	≤ 50	*	120**	Hemophilia
11	*	≤ 50	*	120**	Virus
5	*	> 50	*	151**	Hemophilia
6	*	> 50	*	151**	Cancer
7	*	> 50	*	151**	Virus
8	*	> 50	*	151**	Virus
3	*	≤ 50	*	120**	Cancer
4	*	≤ 50	*	120**	Cancer
10	*	≤ 50	*	120**	Hemophilia
12	*	≤ 50	*	120**	Virus

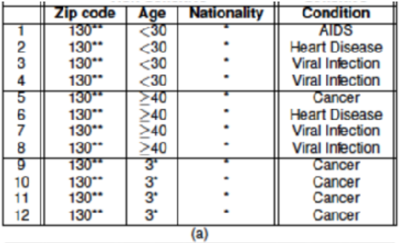
t-closeness : the distribution in each group should also be close to that of the general population

Problems with *k*-anonymity and similar methods

- Composition attacks
  - Combination of knowledge coming from different sources (linking attacks)
  - Open world: Even if present data are protected, in the future there may be some new knowledge available
- Everything can potentially be a quasi-identifier
  - Especially in high-dimensional and sparse databases

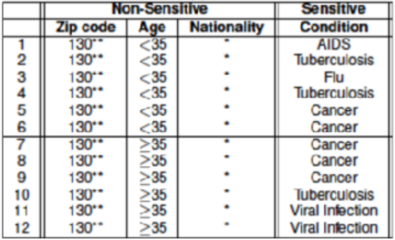
Problems with *k*-anonymity and similar methods

Alice is 28 years old, lives in 13012 and visits both hospitals.
Can we learn something about her?



	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

(a)



	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Famous deanonymization attacks : Netflix

Robust De-anonymization of Large Sparse Datasets.
Narayanan and Shmatikov, 2008.

Showed the limitations of K-anonymity

De-anonymization of the **Netflix Prize dataset** (500,000 anonymous records of movie ratings), using **IMDB** as the source of background knowledge.

They demonstrated that an adversary who knows just a few preferences about an individual subscriber can identify his record in the dataset.

IMDb®
The Internet Movie Database

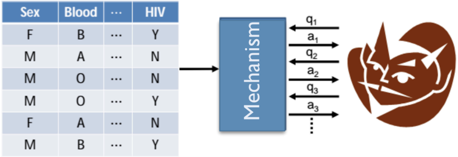
Famous deanonymization attacks : Twitter

De-anonymizing Social Networks.
Narayanan and Shmatikov, 200

By using only the network topology, they were able to show that 33% of the users who had accounts on both **Twitter** and **Flickr** could be re-identified in the anonymous Twitter graph with only a 12% error rate.

Database access via a query interface

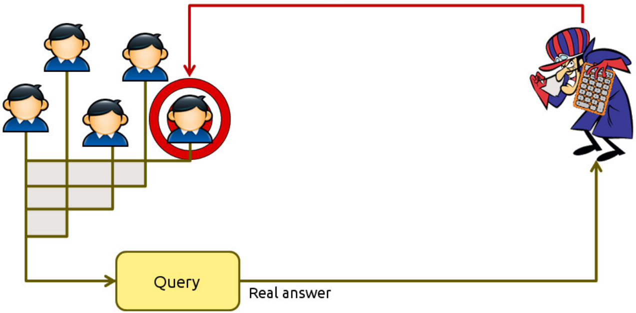
- Do not make the microdata available, but only aggregated information, by querying the interface.
- **Example:** Statistical Databases (SDB), often used for research purposes. For example, a medical SDB can be used to study the correlation between certain diseases and other attributes like: age, sex, weight, etc.



Sex	Blood	···	HIV
F	B	···	Y
M	A	···	N
M	O	···	N
M	O	···	Y
F	A	···	N
M	B	···	Y

- One can only retrieve aggregated information, not personal records
  - *"What is the average weight of people affected by the disease ?"* ✓
  - *"Does Don have the disease ?"* ✗

Problem 1 : correlation



- “what is the median age of cancer patients”
- Statistics are still correlated to personal information
- Inference could be possible

Problem 2 : composition attacks

- 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

Problem 2 : composition attacks

- 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

Problem 2 : composition attacks

Combine with the two queries: minimal weight and the minimal age of a person with the disease

Answers: 40, 100. **Unique!**

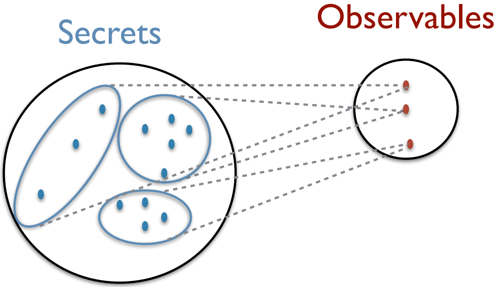
name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

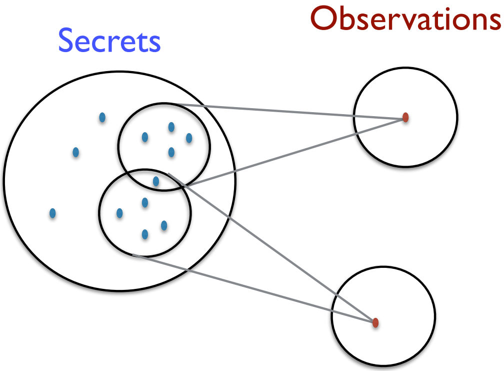
Problem 2 : composition attacks

Composition attacks are a general problem of **Deterministic approaches :** They are all based on the principle that one observation corresponds to many possible values of the secret (group anonymity)



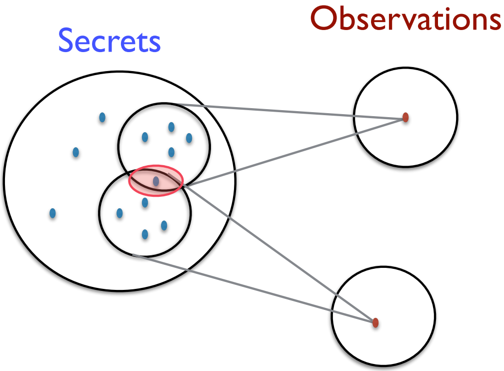
Problem 2 : composition attacks

Problem of the deterministic approaches: the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



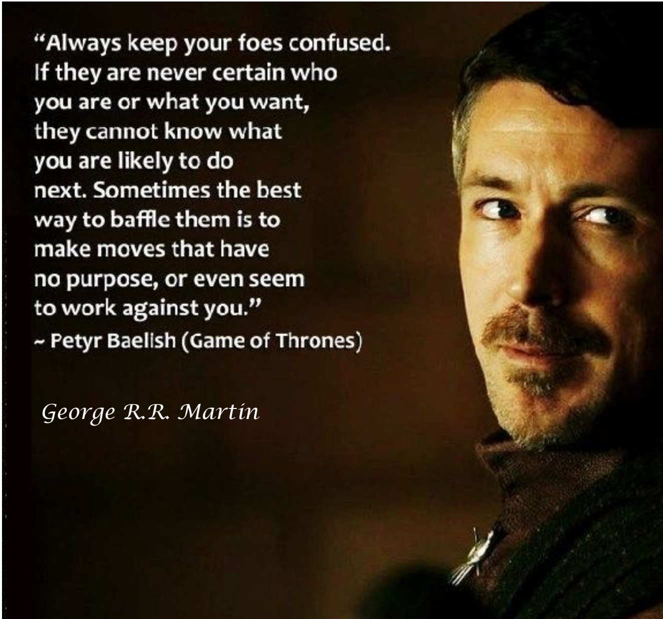
Problem 2 : composition attacks

Problem of the deterministic approaches: the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



Solution

Randomization!



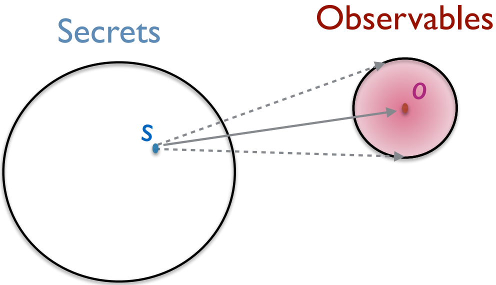
"Always keep your foes confused. If they are never certain who you are or what you want, they cannot know what you are likely to do next. Sometimes the best way to baffle them is to make moves that have no purpose, or even seem to work against you."

~ Petyr Baelish (Game of Thrones)

George R.R. Martin

Randomized approaches

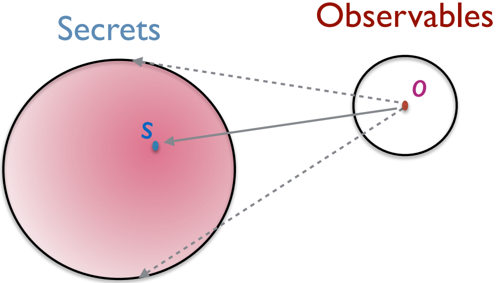
Every secret can generate any observable, according to a certain probability distribution.



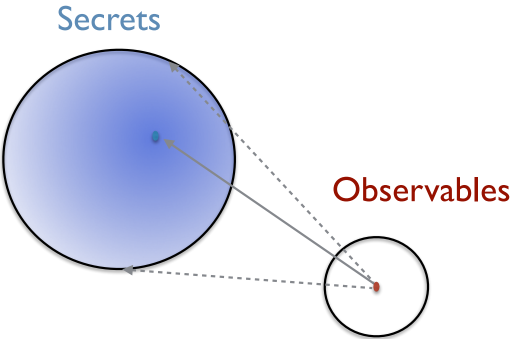
Randomized approaches

By the Bayes law

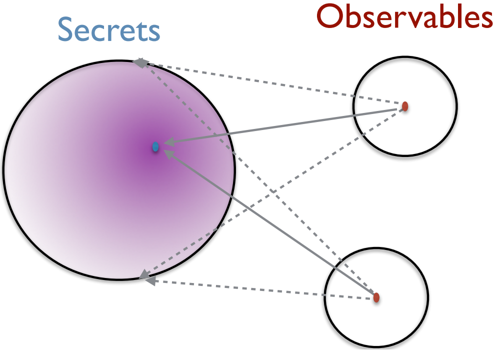
$$p(s|o) \propto p(o|s)$$



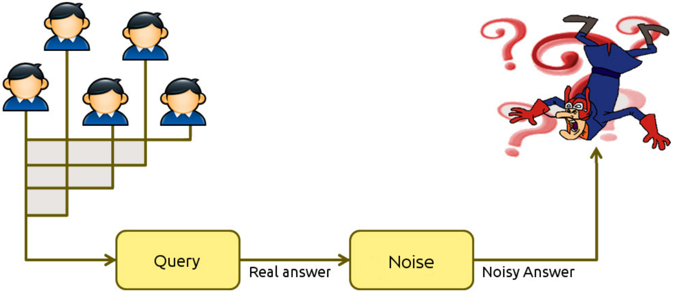
Randomized approaches



Randomized approaches



Randomization for data analysis



- Add noise to query answer before reporting

Randomization for data analysis

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

Randomization for data analysis

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

Randomization for data analysis

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

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

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

Randomization for data analysis

Questions to investigate

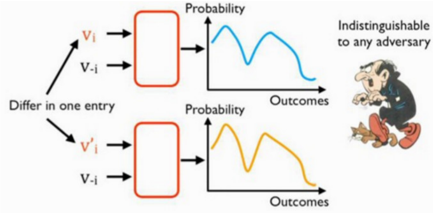
- How can we define privacy for noisy queries?
- What kind of noise do we need?

Differential Privacy

A rigorous definition of privacy for data analysis.

Main idea

Datasets differing in a single individual
should produce “similar” results
(all answers should be produced with almost the same probability).



Differential Privacy

Main idea

Datasets differing in a single individual
should produce "similar" results
(all answers should be produced with almost the same probability).

Definition

Mechanism K satisfies ϵ-differential privacy iff

$$\frac{Pr[\ K(x) = z\ ]}{Pr[\ K(x') = z\ ]} \leq e^{\epsilon} \quad \text{for all} \quad x \sim x', z$$

$x \sim x'$: differing in a single individual

Differential Privacy

Two important properties:

- Independence from the prior
- Compositionality

Independence from the prior

- Prior (initial) knowledge on the database
  - the height of Alice, etc

Independence from the prior

- Prior (initial) knowledge on the database
  - the height of Alice, etc
- The definition of DP does makes no assumptions about it
- So we can prove/disprove the privacy of K without such assumptions

Independence from the prior

- Prior (initial) knowledge on the database
  - the height of Alice, etc
- The definition of DP does makes no assumptions about it
- So we can prove/disprove the privacy of *K* without such assumptions
- Important : this does not mean that prior knowledge does not help the adversary

Compositionality

Theorem

If K_1, K_2 satisfy ϵ_1, ϵ_2-diff. privacy then their composition $K_1 \times K_2$ satisfies $\epsilon_1 + \epsilon_2$-diff. privacy.

Compositionality

Theorem

If K_1, K_2 satisfy ϵ_1, ϵ_2-diff. privacy then their composition $K_1 \times K_2$ satisfies $\epsilon_1 + \epsilon_2$-diff. privacy.

- How does this compare to *k*-anonymity?

Compositionality

Theorem

If K_1, K_2 satisfy ϵ_1, ϵ_2-diff. privacy then their composition $K_1 \times K_2$ satisfies $\epsilon_1 + \epsilon_2$-diff. privacy.

- How does this compare to *k*-anonymity?
- What about repeating the same mechanism?

Compositionality

Theorem

If K_1, K_2 satisfy ϵ_1, ϵ_2-diff. privacy then their composition $K_1 \times K_2$ satisfies $\epsilon_1 + \epsilon_2$-diff. privacy.

- How does this compare to *k*-anonymity?
- What about repeating the same mechanism?
- Privacy budget : the analyst start with an initial budget, each time he asks a question the budget is decreased by ϵ. When it is exhausted, he cannot ask more queries.

How to generate the noise?

Solution 1: **Randomized response**

- Query : “what is the average height?”. Assume integer values 50..250.

How to generate the noise?

Solution 1: **Randomized response**

- Query : “what is the average height?”. Assume integer values 50..250.
- Compute the true answer $y = f(x)$

How to generate the noise?

Solution 1: **Randomized response**

- Query : “what is the average height?”. Assume integer values 50..250.
- Compute the true answer $y = f(x)$
- Flip a (biased) coin
  - with pb $\lambda/_{200+\lambda}$ report y
  - otherwise, report some $y' \neq y$ randomly (uniform)

How to generate the noise?

Solution 1: **Randomized response**

- Query : “what is the average height?”. Assume integer values 50..250.
- Compute the true answer $y = f(x)$
- Flip a (biased) coin
  - with pb ${}^{\lambda}/_{200+\lambda}$ report y
  - otherwise, report some $y' \neq y$ randomly (uniform)
- Does this mechanism satisfy differential privacy? For which epsilon?

$$\frac{Pr[\ K(x) = 50\ ]}{Pr[\ K(x') = 50\ ]} \leq \frac{{}^{\lambda}/_{200+\lambda}}{{}^{1}/_{200+\lambda}} = e^{\ln \lambda}$$

How to generate the noise?

Solution 2: **Laplace mechanism** (the most widely used)

- Numerical queries $f : \mathcal{X} \to \mathbb{R}$
- Sensitivity (How “statistical” / “sensitive to individual data” is a query?)
  - $\Delta f = \max_{x \sim x'} |f(x) - f(x')|$
  - High : needs more noise (“what is the height of Bob?”)
  - Low : needs a less noise (“what is the average height?”)

How to generate the noise?

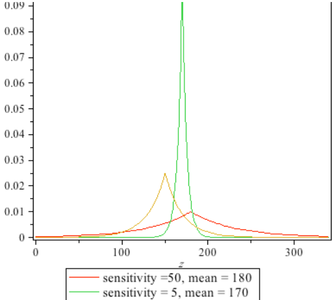
Solution 2: **Laplace mechanism** (the most widely used)

- Numerical queries $f : \mathcal{X} \to \mathbb{R}$
- Sensitivity (How “statistical” / “sensitive to individual data” is a query?)
  - $\Delta f = \max_{x \sim x'} |f(x) - f(x')|$
  - High : needs more noise (“what is the height of Bob?”)
  - Low : needs a less noise (“what is the average height?”)
- Compute $y = f(x)$
- Draw value from $\mathrm{Lap}(y, \frac{\Delta f}{\epsilon})$

How to generate the noise?

Solution 2: **Laplace mechanism** (the most widely used)

- Numerical queries $f : \mathcal{X} \to \mathbb{R}$
- Sensitivity (How "statistical" / "sensitive to individual data" is a query?)
  - $\Delta f = \max_{x \sim x'} |f(x) - f(x')|$
  - High : needs more noise ("what is the height of Bob?")
  - Low : needs a less noise ("what is the average height?")
- Compute $y = f(x)$
- Draw value from Lap$(y, \frac{\Delta f}{\epsilon})$
  - Draw a, b uniformly in $(0, 1)$
  - Report $z = y + \frac{\Delta f}{\epsilon} \log \frac{a}{b}$



References

- Why Privacy Matters Even if You Have ’Nothing to Hide
- “Anonymized” data really isn’t—and here’s why not
- *k*-Anonymity: A Model for Protecting Privacy
- Robust De-anonymization of Large Sparse Datasets
- An introduction to differential privacy