Workshop on privacy for IOT

Héber Hwang Arcolezi

Université de Franche-Comté, UFR-STGI

Some info

Distribution of hours (expected)

- 5 Thursdays with 4h programmed each
- About $\sim$ 1.5h of theory with discussion about the subject and intervals
- About $\sim$ 2h for practical and didactic exercises

Evaluation

i Presence

ii Didactic and practical work (questions in discord group)
   - Group of two (if wish)
   - Send to $\rightarrow$ heber.hwang_arcolezi [at] univ-fcomte [point] fr
   - Discussion on a paper : understanding of the paper main ideas, organization, experiments, what could be improved (?), ...

Personal data, why care ?

Risks on privacy breach

- Sensitive information can be used for *fraudulent* purposes.
- Some issues :
  - i Inference attacks (use of background knowledge)
    ↳ Social media can reveal political opinions, religion, ...
  - ii Privacy-utility trade-off
    ↳ Minimize effects of anonymization on the data utility
  - iii Who can we trust ?
    ↳ Enterprises ? Service provider ? Data analyst ?

Example : mobile phone data

- Green party politician Malte Spitz asks 6 months of mobile phone data
- He asks for a company to treat the data and infer knowledge
- Result : → `https://www.zeit.de/datenschutz/malte-spitz-data-retention`

Motivation [1]

IoT data collection's power

- Cheap sensors and computing power, miniaturization, location positioning technology, ...
- Ability to sense, analyze, and communicate.
- Comprises a wide range of products (growing list) :
  - Individual : fitness and wearable devices, smart devices (TVs, locks, lighting), indoor security system, ...
  - Enterprise : optimization by sensors, face recognition systems, worker productivity tracking devices, drones, ...

Security VS privacy (?)

- Security concerns about *illegitimate* uses of data.
  - Cybersecurity risks
  - IoT devices shall meet basic security protocols (encryption, authentication, security updates)
- Privacy concerns about numerous *legitimate* but harmful use of data :
  - Legitimate $\neq$ positive (profiling, targeting, ...) ;
  - IoT devices often lack screens (control privacy settings)
    ↳ links to websites, where the privacy policies are often difficult to find or understand

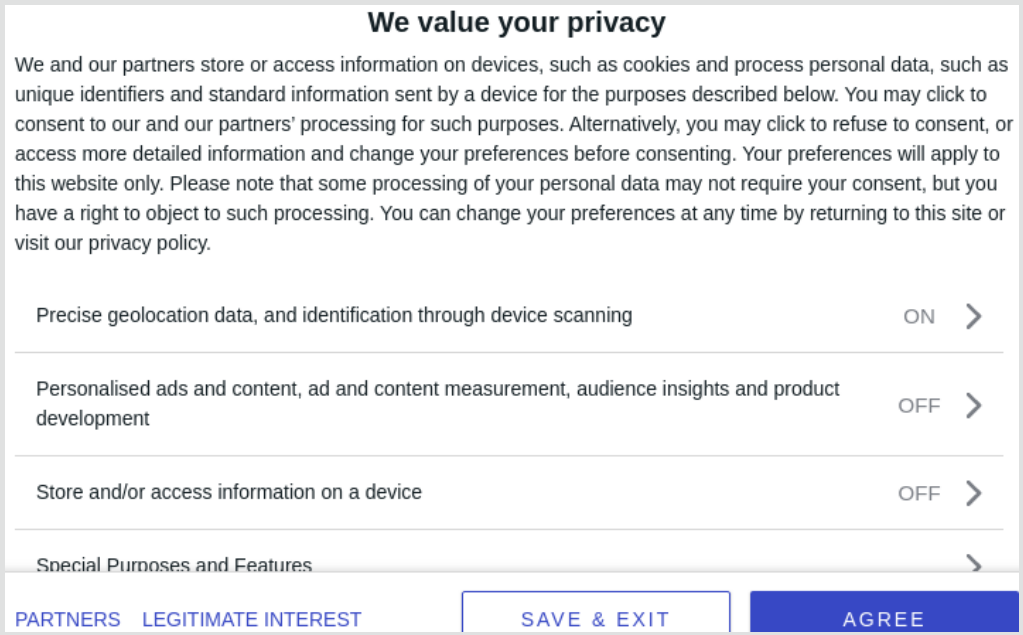
1. Rosner, G and Kenneally, E, Privacy and the Internet of Things : Emerging Frameworks for Policy and Design (June 7, 2018). UC Berkeley Center for Long-Term Cybersecurity/Internet of Things Privacy Forum.

Motivation

Some examples [2]

- What does google stores from you ? → `https://myactivity.google.com/myactivity`
- What does google recommends for you ? → `https://adssettings.google.com/authenticated`
- Google explicitly says that no personal data is sold to other companies → yet, they own and exploit them.

Cookies : `https://www.powerthesaurus.org/`



2. https ://openclassrooms.com/fr/courses/5280946-protegez-les-donnees-personnelles

Plan

Big data and privacy protection

Unsafe publication of data

A first model of privacy protection : k-anonymity

Plan

Plan

Privacy : Aggravated by Big Data

Big Data & Data mining

- Data Mining : inference of interesting knowledge from large amounts of data (Big Data)
- General Trends : Exploring growing Data // Big Data
- Analyze/extraction of knowledge : technically feasible today

Big Data : application and volume

- Areas : IOT, scientific discovery, health, human mobility, profiling
- Big Data Health Market : $\approx$ 67.82 G$ in 2025 (Globe News Wire)

- 90% of all data : created within the last two years (IBM)
- 97.2% of organizations : invest in Big Data and AI (New Vantage)
- Jobs in the field : $\approx$ 2.7M in 2020 (Forbes)

Plan

Privacy ? [5]

History

- Expression of “the right to be let alone" [3]
- “No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, [. . .]. Everyone has the right to the protection of the law against such interference or attacks." [4]
- In the age of the Internet and (personal) data transmitted by : smartphones, messaging services, GPS, fitness equipment, search engine . . .

Examples of problematic inferences

- Refrigerator controlling consumed products : $\rightsquigarrow$ number of present/absent people at home, sanitary risk ? insurance ?
- Health tracking application : positions, heart rates shared with Apple / Google only ?

3. Warren, S. D., & Brandeis, L. D. (1890). The right to privacy. Harvard law review, 193-220.
4. United Nations, (1949), Universal Declaration of Human Rights.
5. https ://openclassrooms.com/fr/courses/5280946-protegez-les-donnees-personnelles

Scandals of *non*-privacy protection

National Security Agency (NSA) PRISM (2007-2013) [6]

- In 2007, the PRISM program created by the NSA aimed to monitor communications exchanged on Big Tech's online services, in particular
- Transmission of raw data to third countries
- Spying on international political leaders (including A. Merkel's cell phone)

Cambridge Analytica (CA) (2014-2018) [7]

- "This is your digital life" : application for Facebook, which allows the extraction of personal data present on the network
- 87M Facebook user accounts extracted in from 2014
- Political targeting to convince to vote for D. Trump in 2016
- "Without CA, there would have been no Brexit" according to C. Wylie

6. https ://en.wikipedia.org/wiki/PRISM_(surveillance_program)
7. https ://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal

Plan

General Data Protection Regulation (GDPR)[8]

Oriented towards individuals' rights :

- A. 6 : "Minimum data collected and processed in a fair and lawful manner"
- A. 12 : "Transparent information, communication and modalities for the exercise of the rights of the data subject"
- A. 16 : "Right to rectification"
- A. 17 : "Right to erasure ('right to be forgotten')"
- A. 20 : "Right to data portability" : recoverable data (readable format)

Data collectors

- Pay attention to $\rightarrow$ Economic model of the service : information extraction
- DB administrators : often honest but at risk
- Data encryption : often a solution to *illegitimate* use of data

8. https ://gdpr-info.eu/

Plan

Big data and privacy protection

Unsafe publication of data

- Anonymization by pseudonymization : attackable

A first model of privacy protection : k-anonymity

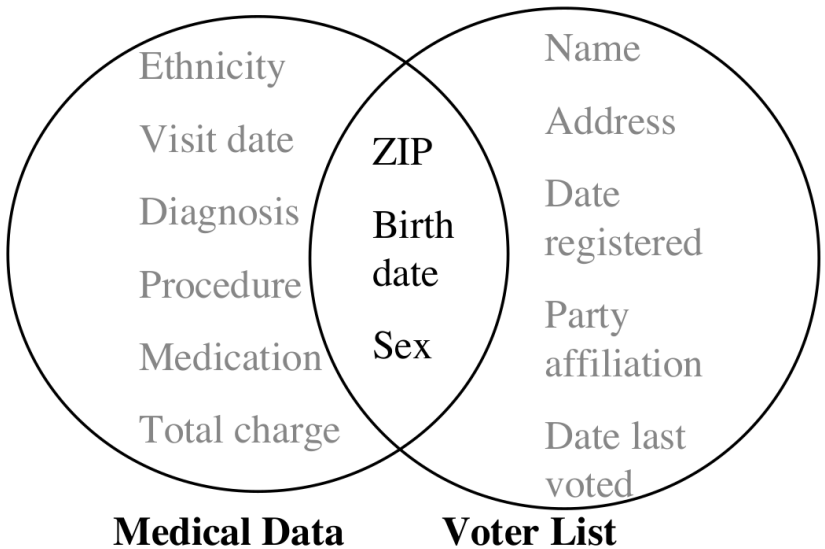
Pseudonymization

- Identifier fields : deleted and replaced by an id.
- Advantage : calculations identical to those on the initial database (Average age/cancer = 37.8)

	Non-sensitive				Sensitive
Id	Zip	Age	Gender	Nationality	Disease
1	13053	28	M	russian	heart
2	13068	29	M	american	heart
3	13068	21	F	japanese	viral
4	13053	23	M	american	viral
5	14853	49	M	indian	cancer
6	14853	48	F	russian	heart
7	14850	47	M	american	viral
8	14850	49	F	american	viral
9	13053	31	M	american	cancer
10	13053	37	M	indian	cancer
11	13068	36	F	japanese	cancer
12	13068	35	F	american	cancer

Pseudonymization : attack by QID, Sweeney [9]

- Pseudonymized and public medical database



- Notion of Quasi IDentifiers (QID) : zip, date of birth, gender
- Public voters list, USA census, 1990 : "87% of the population in the US had characteristics that likely made them unique based only on 5-digit Zip, gender, date of birth"
- Massachusetts Governor's medical data identification

9. Sweeney, L. (2002). k-anonymity : A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Pseudonymization : intersection attack [10]

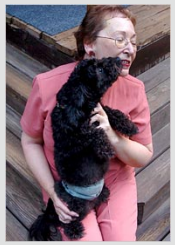
2006, 20M web search queries collected by AOL, 658K unnamed user

AnonID	Query	QueryTime
1326	“holiday mansion houseboat”	2006-03-29
1326	“back to the future”	2006-04-01
591476	“english spanish translator”	2006-03-20
591476	“panama vacations”	2006-03-20
591476	“breast reduction”	2006-03-23
591476	“volunteer work at hospitals in brooklyn”	2006-05-24
591476	...	...
591476	“how to secretly poison your ex”	2006-03-12

Thelma Arnold, 62 years, widow living in Lilburn, Ga., reidentified in 3 days

AnonID	Query
4417749	“people with last name ’Arnold”’
4417749	“landscapers in Lilburn, Ga”
4417749	“60 single men”
4417749	“dog that urinates on everything”
4417749	dog-related queries

$\rightsquigarrow$ Early supression of data from the AOL site.



10. BARBARO, Michael, ZELLER, Tom, et HANSELL, Saul. A face is exposed for AOL searcher no. 4417749. New York Times, 2006, vol. 9, no 2008, p. 8.

Plan

Big data and privacy protection

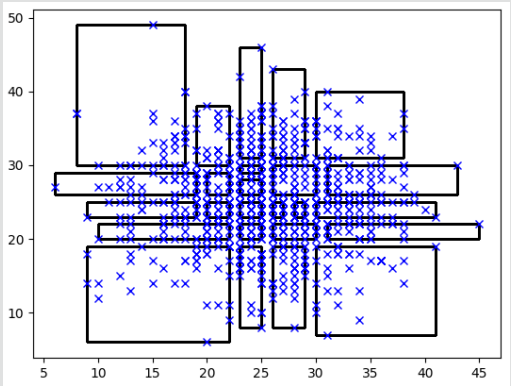
Unsafe publication of data

A first model of privacy protection : k-anonymity

- Two algorithms to achieve it
- Utility measurements
- k-anonymity attacks

k-anonymity[9]

Graphic intuition : regroup



Principle : equivalence classes sized k at least

- QID : attributes whose combinations lead to an identification
- Level of detail of QID values : to be reduced so that there are at least k different individuals whose QIDs are equal
- Individuals with the same QIDs : are part of the same equivalence class
- To solve :
  - value for k ?
  - loss of information ?

Plan

Big data and privacy protection

Unsafe publication of data

A first model of privacy protection : k-anonymity

- Two algorithms to achieve it
- Utility measurements
- k-anonymity attacks

k-anonymity by generalization

Reduce detail levels

- Zip : suppress numbers D$\rightarrow$G : XXXX*, XXX**,XX***, X****, total suppression.
- Age : by intervals of growing amplitude : 10, 20, 40, total suppression.
- Gender : total suppression
- Nationality : by continent, total suppression.
- $\rightsquigarrow 6 \times 5 \times 2 \times 3 = 180$ combinations of generalization !

Grouping by equivalence classes of card. $\geq$ to k

	Quasi-IDentifiers				Sensitive
Id	Zip	Age	Gender	Nationality	Disease
1	130**	[21; 31[	*	*	heart
2	130**	[21; 31[	*	*	heart
3	130**	[21; 31[	*	*	viral
4	130**	[21; 31[	*	*	viral
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	heart
7	148**	[41; 50[	*	*	viral
8	148**	[41; 50[	*	*	viral
9	130**	[31; 41[	*	*	cancer
10	130**	[31; 41[	*	*	cancer
11	130**	[31; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer

4 individuals

4 individuals

4 individuals

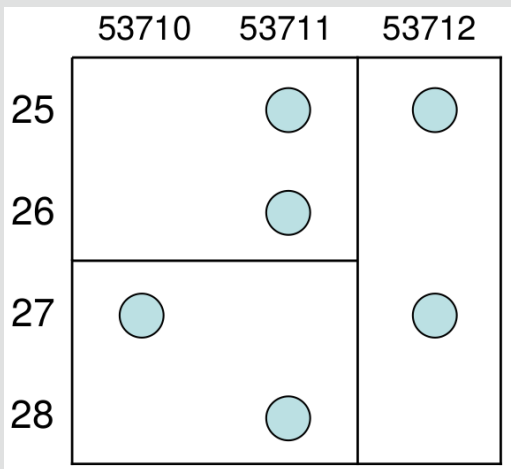
k-anonymity by Mondrian algorithm [11]

General algorithm

1. Partition data into groups of size $\geq k$, with kd-trees
2. Abstract attribute values : a group $\equiv$ a value

Example of 2-anonymity, partitioning of (Age, zip)

Age	Gender	Zip	Disease
25	M	53711	Flu
25	F	53712	Hepatitis
26	M	53711	Bronchitis
27	M	53710	Broken arm
27	F	53712	HIV
28	M	53711	Lost nail



Age	Gender	Zip	Disease
[25-26]	M	53711	Flu
[25-27]	F	53712	Hepatitis
[25-26]	M	53711	Bronchitis
[27-28]	M	[53710-53711]	Broken arm
[25-27]	F	53712	HIV
[27-28]	M	[53710-53711]	Lost nail

11. LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006, April). Mondrian multidimensional k-anonymity. In 22nd International conference on data engineering (ICDE'06) (pp. 25-25). IEEE.

Plan

C_{AVG} : avg nb of elements per standardized class

Definition for $|T|$ records and properties

$$C_{AVG} = \frac{|T|}{|EQs|} \times \frac{1}{k}$$

- and $|EQs|$: number of equivalence classes.
- $\frac{|T|}{|EQs|}$ average number of elements per class ($\geq k$) $\rightsquigarrow$ $C_{AVG} \geq 1$
- Utility decreases as C_{AVG} increases.

Example with 4-anonymity

	Quasi-IDentifiers				Sensitive
Id	Zip	Age	Gender	Nationality	Disease
1	130**	[21; 31[	*	*	heart
2	130**	[21; 31[	*	*	heart
3	130**	[21; 31[	*	*	viral
4	130**	[21; 31[	*	*	viral
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	heart
7	148**	[41; 50[	*	*	viral
8	148**	[41; 50[	*	*	viral
9	130**	[31; 41[	*	*	cancer
10	130**	[31; 41[	*	*	cancer
11	130**	[31; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer

- $C_{AVG} = \frac{12}{3} \times \frac{1}{4} = 1$
- Optimal for this metric

Discernability

Definition for $|T|$ records, and properties

$$Disc = \sum_{EQ, |EQ| \geq k} |EQ|^2 + \sum_{EQ, |EQ| < k} |T| \times |EQ|$$

- EQ, an equivalence class
- Utility decreases as $Disc$ increases.

Example with 4-anonymity

	Quasi-IDentifiers				Sensitive
Id	Zip	Age	Gender	Nationality	Disease
1	130**	[21; 31[	*	*	heart
2	130**	[21; 31[	*	*	heart
3	130**	[21; 31[	*	*	viral
4	130**	[21; 31[	*	*	viral
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	heart
7	148**	[41; 50[	*	*	viral
8	148**	[41; 50[	*	*	viral
9	130**	[31; 41[	*	*	cancer
10	130**	[31; 41[	*	*	cancer
11	130**	[31; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer

- $Disc = 4^2 + 4^2 + 4^2 = 48$

Plan

Big data and privacy protection

Unsafe publication of data

A first model of privacy protection : k-anonymity

- Two algorithms to achieve it
- Utility measurements
- k-anonymity attacks

k-anonymity : attacks

Example

	Quasi-IDentifiers				Sensitive
Id	Zip	Age	Gender	Nationality	Disease
1	130**	[21; 31[	*	*	heart
2	130**	[21; 31[	*	*	heart
3	130**	[21; 31[	*	*	viral
4	130**	[21; 31[	*	*	viral
5	148**	[41; 50[	*	*	cancer
6	148**	[41; 50[	*	*	heart
7	148**	[41; 50[	*	*	viral
8	148**	[41; 50[	*	*	viral
9	130**	[31; 41[	*	*	cancer
10	130**	[31; 41[	*	*	cancer
11	130**	[31; 41[	*	*	cancer
12	130**	[31; 41[	*	*	cancer

$\}$ 4 individuals

$\}$ 4 individuals

$\}$ 4 individuals

Attacks

- Homogeneity :
  - $\oplus$ 35-year-old patient known to be $\rightsquigarrow$ cancer.
  - $\ominus$ Known 29-year-old patient $\rightsquigarrow$ ~~cancer~~.
- Additional knowledge : a *21* year old Japanese, P(heart|Japanese) = weak $\rightsquigarrow$ viral.

Thanks for your attention !

Further questions ? ?
Feedback most welcome :D (email me)