Workshop on privacy for IOT

Héber HWANG ARCOLEZI

Université de Franche-Comté, UFR-STGI

Quick recap: TD and TP

- ~ 2h
- Group of two (if wish)
- Send to → heber.hwang_arcolezi [at] univ-fcomte [point] fr
- Questions in the discord group

Quick recap: Data Protection[1]

- *What?* → Law designed to protect our personal data
- *Why?* → Every time we use a service, go to the doctor, pay taxes, online shopping, make mobile phone calls, …
  - We transfer personal/sensitive data
  - Companies gather knowledge without consent (profiling, targeting, …)
  - Citizens can '*only*' hold up on data protection regulations (e.g., GDPR)
- *Privacy* → internationally recognized human right

Quick recap: Privacy-utility trade-off

- Given a dataset with personal (sensitive) data:
  - Health information
  - Social network activity
  - Location
  - Census data
- How can one:
  - ‘Learn’ (data mining) patterns and basic statistics
  - Without compromising the ‘privacy’ of users (?)

Research
Urban planning
Business development
Identify threats
...

Quick recap: Pseudonymization

- Identifier fields: deleted and replaced by an id.
- Advantage: calculations identical to those on the initial database
- *Problems?*

	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

Quick recap: AOL Data Release[2]

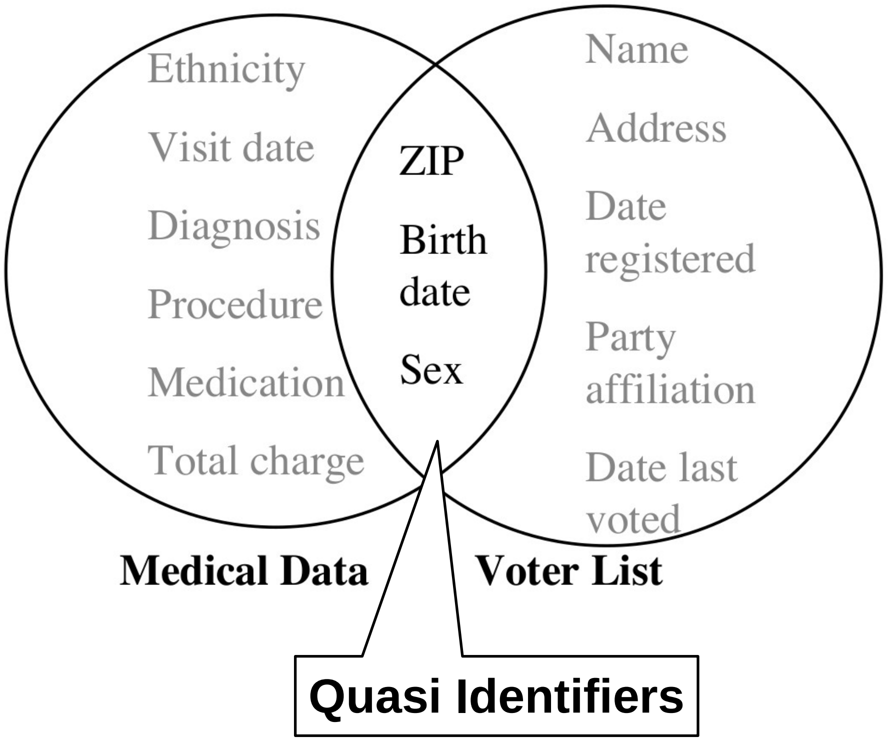
AnonID	Query
4417749	“people with last name 'Arnold"'
4417749	“landscapers in Lilburn, Ga”
4417749	“60 single men”
4417749	“dog that urinates on everything”
4417749	dog-related queries

⇝ Early supression of data from the AOL site.

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

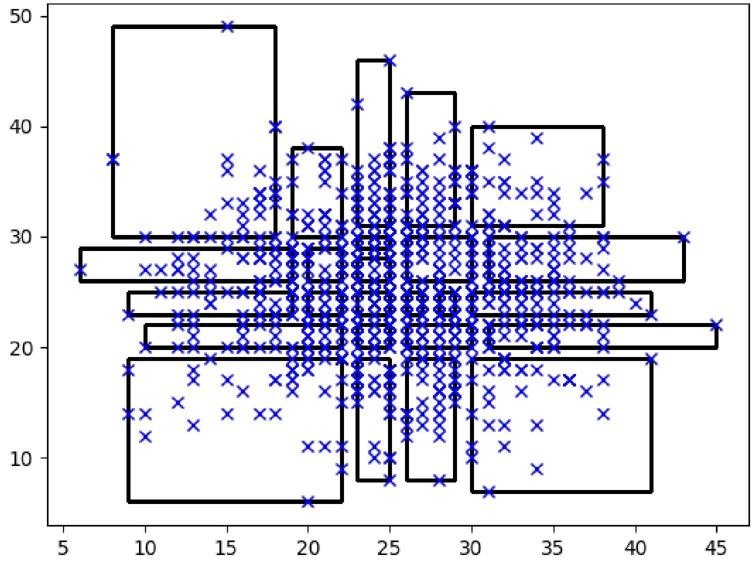
Quick recap: Massachusetts Gov[3]

- Pseudonymized and public medical database + Public voters list, USA census, 1990
- Sweeney's research: ~ 87% of the US population uniquely identifiable (Zip, DoB, Sex)
- Massachusetts Governor's medical data identification



Quick recap: *k*-anonymity[3]

- Make every record in the dataset *indistinguishable* from at least k-1 others.
- “Safety in a group”
- How?
  - Clustering,
  - Suppression,
  - Generalization,
  - Dummy records, ...



Quick recap: *k*-anonymity

- Generalization, Mondrian algorithm, Incognito algorithm, …
- *Some* problems?
  - Homogeneity
  - Background knowledge

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

Quick recap

- Questions?
- Discussion
  - Personal data, why care?
  - Big data (IoT) and data mining

Plan

- Extensions of *k*-anonymity
  - ***l*-diversity**
  - *t*-closeness

l-diversity[4]

- Main idea → Each equivalent class (EQ) contains at least *l* well-represented sensitive values
- Database is *l*-diverse *iff* all its EQs are *l*-diverse

4-anonymity and 3-diversity

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

3 sensitive val. $\neq$

3 sensitive val. $\neq$

Limitation of *l*-diversity

- Example → Original DB:
  - One sensitive value: HIV test
  - Two outcomes: positive (1 %) and negative (99 %)
- Values with degrees of sensitivity very different:
  - Little opposition for the ones whose test is negative (like 99% of the population)
  - Strong reluctance to be known tested positive
- EQs with only negative outcomes do not need *l*-diversity

Limitation of *l*-diversity

- *l*-diversity is difficult to achieve: |DB| = 10,000, pos. (1 %) and neg. (99 %)
  - To achieve 2-diversity, there can be at most 10,000*1% = 100 EQs
- The overall distribution of sensitive values matters:
  - An EQ with equal number of positive and negative records
  - Diversity does not differentiate among:
    - EQ1: 49 positives and 1 negative
    - EQ2: 1 positive and 49 negatives

Limitation of *l*-diversity

Bob	
ZIP	Age
47688	29

3-diverse with sensitive : salary and disease

ZIP	Age	Salary	Disease
476**	2*	3K	gastric ulcer
476**	2*	4K	gastritis
476**	2*	5K	stomach cancer
4790*	≥ 40	6K	gastritis
4790*	≥ 40	11K	grippe
4790*	≥ 40	8K	bronchitis
476**	3*	7K	bronchitis
476**	3*	9K	pneumonia
476**	3*	10K	stomach cancer

- Possible deductions from knowing that Bob is in EQ1
  - his Salary ([3K-5K]) is relatively low
  - suffers from stomach related diseases (semantic meaning matters...)

Plan

- Extensions of *k*-anonymity
  - *l*-diversity
  - ***t*-closeness**

t-closeness[5]

- Main idea → Distribution of sensitive attribute values in each EQ should be close to that of the original dataset (distance ≤ t)
- Measure distance between two distributions so that semantic relationship among sensitive attribute values is captured.
  - Earth Move Distance
- A DB is said to have *t*-closeness if all EQs have *t*-closeness
- Limitations:
  - Utility may suffer too much
  - Distinction between QIDs and sensitive attributes

t-closeness: example[5]

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

Summary

- *k*-anonymity, *l*-diversity, *t*-closeness:
  - Require*:* Difference between quasi-identifiers and sensitive attributes
  - Require*:* Model (or at least try to) background knowledge of adversaries
  - Not compositional
  - Syntactic privacy models: Privacy is a property of only the final output
    - Generalize the database entries until some syntactic condition is met
- Next session → From syntactical privacy notions to Differential Privacy[6]
  - Privacy is a property of the algorithm

References

1. Privacy International. The Keys to Data Protection: A Guide for Policy Engagement on Data Protection. 2018
2. 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.
3. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.
4. Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity." ACM Transactions on Knowledge Discovery from Data (TKDD) 1.1 (2007): 3-es.
5. Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007.
6. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science (3–4), 211–407 (2014)

Thanks for your attention!

Further questions??

Feedback most welcome :D (email me)