BANQUE NATIONALE
Réalisons vos idées

CRM WORKSHOP 2022

Measuring the performance, security, and confidentiality of an anonymization approach

Pascal JUTRAS DUBE
Patrick MESANA

Supervisor : Gilles CAPOROSSI

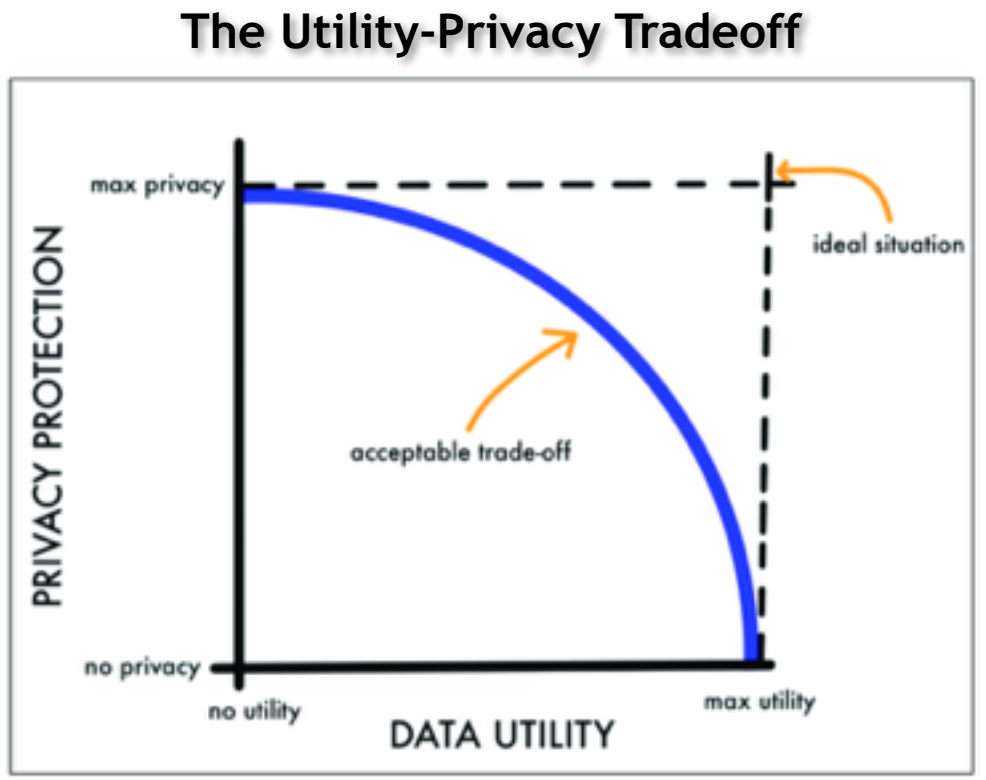
2022-08-22

Plan

- ❑Context
- ❑What do we expect from the workshop
- ❑A head start on the privacy literature
- ❑Workshop evaluation criteria
- ❑Public datasets
- ❑Questions

Context

- The Bank is a business that runs on trust and data is one our strongest assets
- Cybersecurity and Data Governance put a lot of constraints on organizations.
- How do you get the best tradeoff between privacy risks and data utility?
- Our approach: Internal Privacy-Preserving Data Publishing (Anonymization for now)



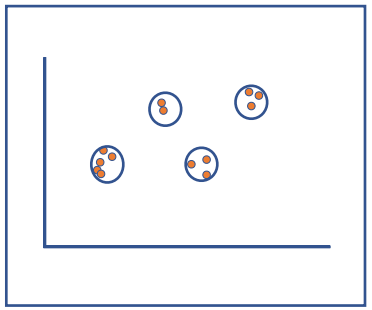
DOI:10.3390/inventions6030045

Privacy-Preserving Data Publishing (PPDP)

Example: anonymisation

Data Source

DB token	date	merchant	$
1234	20-01-01	grocery inc	20
1234	20-01-12	airline inc	200
1234	20-01-18	grocery inc	80
1235	20-01-22	clothes inc	34
…	…	…	…



Generalisation
Noise addition
Deletion
…

Anonymised Dataset

DB token	date	merchant	$
1234	20-01-01	grocery inc	10
1234	20-01-12	airline inc	250
1234	20-01-18	grocery2 inc	50
1235	20-01-22	clothes inc	34
…	…	…	…

1. Can you still re-identify someone in the anonymized dataset?
2. Can you infer information on someone in the anonymized dataset?

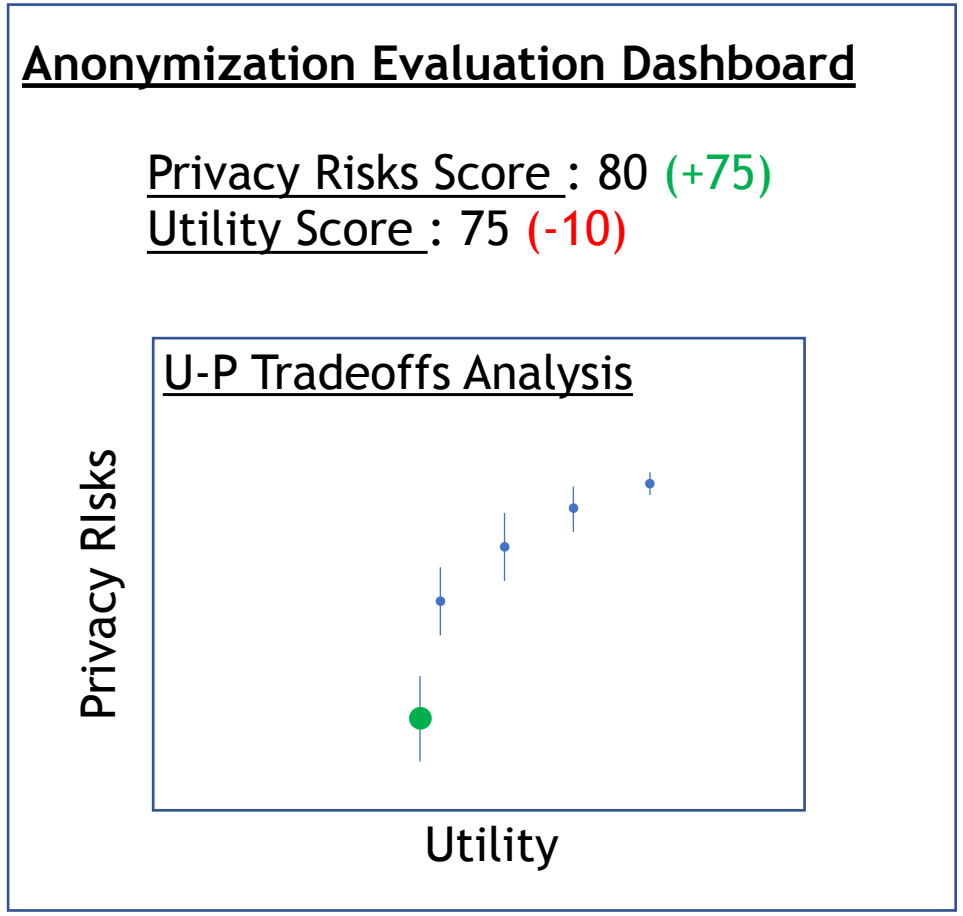
CRM Workshop Challenge

Our Objective: Accelerate access to personal information by using anonymized or synthetic data within the bank's environments, with the confidence we have risks under control.

Workshop Challenge: Participants will have to come up with metrics and insights related to the privacy risks identified in the scientific literature and contextualized with the bank's needs. Utility tradeoff insights are a plus.

Results should be agnostic of the privacy-preserving method used.

An example …



A head start on the privacy literature

Goals	Techniques	Privacy Models
Privacy-preserving data publishing (PPDP)	Anonymization / Data transformations - Generalization - Deletion - Noise addition - Aggregation - Permutation	K-anonymity (2002) L-diversity (2007) T-closeness (2007)
	Data Synthesis - Agent-based modelling - Sampling - ML	Domain Specific (2000+) Differential Privacy (2008)
Privacy-preserving data mining (PPDM) *Now includes ML and Advanced Analytics*	Noisy queries Noisy ML training and predictions	Differential Privacy (2008)

Privacy Models != Privacy Risks

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

Linkage attack

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K

Quasi Identifiers (QIDs)

Can you associate exactly one record to *John Smith*?

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K
4	M	33	MSc	DS	74K

Equivalence Class

Singling out depends on John's uniqueness!

Identity Disclosure Risk

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

Linkage attack

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K

Quasi Identifiers (QIDs)

Sensitive Information

Can you infer John's income?

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K
4	M	33	MSc	DS	74K

It depends on the diversity of incomes of similar QIDs!
Attribute Disclosure Risk

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

Linkage attack

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K

Quasi Identifiers (QIDs)

Can you tell if John was included in the dataset?

What is a privacy attack on a published dataset?

Name: John Smith
Sex: Male
Age: 33
Education: MSc
Occupation: Data Scientist

ID	Sex	Age	Education	Occupation	Income
3	M	33	MSc	DS	75K
4	M	33	MSc	DS	74K

It depends on John's *likelihood*!
Membership Disclosure Risk

PPDP Privacy Risks

Identity disclosure (Legal Privacy)	Can you associate a record to an individual that you know?	Anonymization
Attribute disclosure (Confidentiality)	Can you infer a sensitive information from an individual that you know?	Anonymization Data Synthesis
Membership disclosure (Differential Privacy)	Can you deduce that a record of an individual you know is present/ absent in the dataset?	Data Synthesis

References

- Technical Privacy Metrics: A Systematic Survey (acm.org)
- Data anonymisation and synthesis
- Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey
- Exposed! A Survey of Attacks on Private Data
- Privacy-preserving data publishing: A survey of recent developments (acm.org)
- Flexible Data Anonymization Using ARX — Current Status and Challenges Ahead
- Protecting privacy using k-anonymity

Workshop Evaluation Criteria

1. Are the metrics explainable to decision makers who are not experts in privacy preservation techniques?
2. Are the risk metrics reflecting plausible attacker scenarios in the present context?
3. Are the metrics covering all the risks? What are the risks they are covering and why these risks in particular?
4. Is your methodology robust? Is your methodology agnostic of the dataset?
5. Do you have tradeoff insights on what happens when you change the anonymization parameters?

Public Datasets

LINK to Dropbox

Questions