On Protecting Microdata in Open Data Settings: from a Data Utility Perspective

Afshin Amighi, Mortaza S. Bargh, Sunil Choenni, Alexander Latenko, Ronald Meijer Creating 010 / Hogeschool Rotterdam Research and Documentation Center Ministry of Justice and Security

> Presenter: Afshin Amighi (<u>a.amighi@hr.nl</u>) ICDS 2020





(Short) Resume

- Educational background:
 - MSc, KTH, Sweden (Software Engineering for Distributed Systems)
- Work / Research experience:
 - PhD, FMT, University of Twente, The Netherlands (Verification of Concurrent Programs)
 - Informatica Teacher: Hogeschool Rotterdam, Rotterdam, The Netherlands
 - HBO-Postdoc: Creating 010 Knowledge Center, Rotterdam, The Netherlands

Agenda

- Anonymization: Utility vs. Risk
- ARX: A Quick Look
- Our Results
- Ongoing and Future Directions

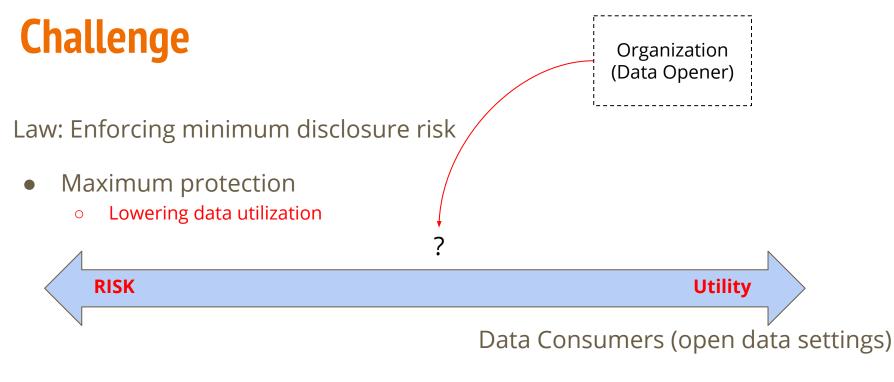
Anonymization: Utility vs. Risk

Motivation



- Opening Data is crucial in innovations and economic growth:
 - Data Collections (huge) in organizations and enterprises
 - R&D needs data analysis
- Personal data disclosure as a main threat
 - Public data must be GDPR compliant
- Huge impact on key areas of society:
 - Healthcare, e-research, e-education, e-government, ...

* GDPR: General Data Protection Regulation



- Expecting high data utility
 - Higher risk of personal disclosure

Challenge

Organizations (data openers) need guidelines

• Data intrinsics, context, environment, data usage, etc. : all can differ

We try to propose a solution



Our Approach

- Investigating several scenarios in data transformations
- Studying SDC (statistical disclosure control) techniques
- ARX as one of the available mature tools:
 - Series of experiments
 - Impact of various <u>measurements factors</u> in choosing a right policy
- Five different cases
 - What are the <u>implications</u>?

Personal Data

Personal data collection:

• In various forms such as <u>microdata (our focus)</u>, tabular data (contains aggregated data), semi-structured data as well as unstructured data.

Three personal data types (GDPR definition):

- Direct identifiable: name, address, etc.
- Indirect identifiable: special property, special feature, etc.
- Sensitive:
 - Special categories: like race, religion.
 - Criminal convictions.

Combinations of categories

Combinations of various categories may reveal individuals.

Various studies done about which combinations can be safe:

- *Example*: opening sensitive data related to persons needs strict measures.
 - "sensitive data sets can be opened to the public if they are without personal information," *

SDC techniques explore optimum combinations.

* M. S. Bargh et al., "Opening privacy sensitive microdata sets in light of GDPR," in 20th Annual International Conference on Digital Government Research, DG.O, Dubai, United Arab Emirates, June 18- 20, 2019, pp. 314–323.

Protecting Microdata

• Anonymization: This process ensures "that the risk of somebody being identified in the data is negligible" hiding the identity and/or the sensitive data of data subjects, while retaining sensitive data for the purpose of data analysis (the so-called SDC methods and tools are used).

• de-identification: This process aims at protecting a microdata set against the intrinsic threats by transforming direct identifiers (like names, social security numbers and digitized unique biometrics). This transformation is carried out via replacing direct identifiers with pseudo identifiers, masking/suppressing them or removing them.

Attribute Mapping

As a first step, the data attributes are divided into various categories.

Four disjoint sub-sets:

- Explicit Identifiers (EIDs): based on intrinsic aspects
- Quasi Identifiers (<u>QIDs</u>): subjective
- Sensitive Attributes (<u>SATs</u>): contextual
- Non-Sensitive Attributes (NATs)

Attribute Mapping

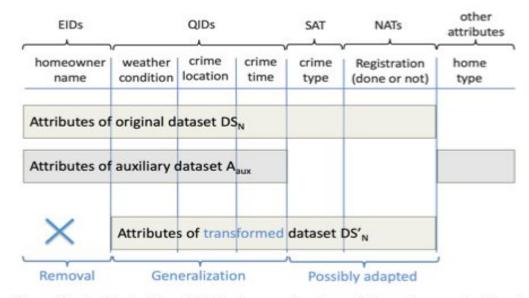


Figure 1. An illustration of attribute mapping for a data environment without the original microdata set.

Guiding Principles

Background information (auxiliary data sets) is growing (Big Data era).

It is shown that *normative privacy preserving approaches* are unable to apply the stringent definitions.

There is a need for <u>more Formal Approaches</u>:

- Mathematically proven techniques
- Regardless of different contexts, rely on properties of the data set.
- ε-differential: one of the pioneering techniques



Definition: "the presence or absence of the (personal) data of an individual in a data set must not have an observable impact on the output of an analysis/computation over that data set."

• Using ARX, we compared it with those of traditional normative approaches

Experiment: Cases

For a given data set we have defined five different cases:

- Case 1: Base (Raw data set)
- Case 2: Basic protection
 - Remove EIDs, others unchanged
- Case 3: Protection Against Data Linkage by Externs
 - Remove EIDs, generalize QIDs, do not change SATs and NATs
- Case 4: Protection against all parties
 - Remove EIDs, define all the other attributes as QID
- Case 5: Formal protection (ε-differential)
 - Remove EIDs, apply ε-differential

ARX: A Quick Look

What is it?

ARX: A comprehensive open source software for anonymizing sensitive personal data.

- privacy and risk models,
- methods for transforming data and
- methods for analyzing the usefulness of output data



Source: https://arx.deidentifier.org/

Features

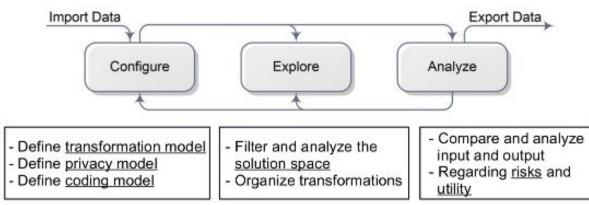
ARX:

- A graphical tool which supports: **data import**, wizards for **creating rules**, **visualizations** of data utility and risks.
- Features data handling that imports various formats (DB,Excel,CSV) and provides functionalities to **handle dirty data**.
- Is available as a **Software Library**: delivers data anonymization capabilities to any programmer.

Features

ARX as a GUI-based Tool.

• Three main (and potentially repeating) steps are: **configure**, **explore** and **analyze**.



	Anony microim T	est-Large								
							-			
-0.09	通過 推進	÷÷ √ x .	SIS DIA							
in Cor	figure transforma	files & Explorered	and a state of the	1219					1110000	19239
Input	data \	1	into pre Analyteutro	Re 🔮 Analyza i	**				Indende	re 1260 Second R.L.C.L.L.L.R. Appleo R.L.C.L.2.
Training of										1.000044423
	110		Warded states	aduation		き目り	Over	001		
1	5	Eleck	Downed	Some-callage	Vetor-Gard Detail Miles	Print I		· Maria	1. 1. 1. 1.	ters and the second
2	-	Back	Demond	Serbeloo	Derind Status	Photo:	1	9.9	Reh	Minuted and exactly in the second sec
	2	Back	Dorond	Anto adm	United Dates	PEan	1.8	59	Inci	tempts tel present tights allo attor have braine have
4	51	Flack	Dymost	Motes	United States	Salies	1993	3.0	Fick	timus of preart higher-electron hash locales. his
5	34	Back	Norer maned	Some-callege	United Street	Pilute	122	3.9	Bet	inners of prised higher election. Next images has
6	11	Black	Separated	Some-callege	United (Articl	Private	1	3.9	Tes	International present higher advances. North Annalysis I lag Million and present Higher advances. Month Annalysis I lag
7	5	Flack	Separated	Assesse	United Strict	Prus	12	9.9	Bet	
		Black	Septembel	Tone-railege	Daries of Males	Print	1	310	Red	server of pound higher effection. North Amatica Na- tamout nat present higher effection. North Amatica Nat-
.9	56	Flack	Ridowed	Some-callege	Garteel Status	Prot		. 31.01	Bel	amout not present higher advanture blattl denotes him
29	2	Minine	Diverce#	Some-callege	United-Teles	Feiter	-00	99	1944	interies of prearing an abuston block interior. See
п	54	White	Dynost	Eathelot	United-States	Feler	U.	9.9	. Phyla	times of possibility about in fast lowers in
12	51	and a second	Dearost	Mates	United States	Loop	12	99	The Real	genue net propert higher aduction. Such America . So senses net propert Higher aduction. North Jennica . So
13	57	libbe	Deprod	Some-callege	Unterstates	Look	1	3.0	-	spream of protect higher advantum family and protect higher advantum family advantum. Territorismo advantum family
34	50	Whete	Dyavasi	Sama-callege	United States	Look-	18: 15	19.0	100	space of pages higher shown have been in
15	56	and the second s	Dearcol	pour and	Canado Sanas		1.00	100		
							1			
ate					2 - 10(9 - 10 - 10		- 100 (101 (101)) (101)	Alam Par Da	nter 18	
	e himian	scan Pac Standar	last 122	Ofer	20	-	101 01-01	Alam Par Da	ade B	**
	e Indan Estano	Scale Par, Scardor	Bed M2	Ofw	20		101 01-01	_	_	
	e Indan-Estimo	scan Pachander		-	20		100 000 15-05	Asam Ancola 1900 - New York	_	
-			MI	Of the Office	9-1 U-A	_	10-51 64-6 10-03	_	_	
-		Scale Production	M2 205	-	9-1 U-A	_	100 000 15-05	_	_	
-			M2 205	-	9-1 U-A	_	10-51 64-6 10-03	_	_	
-			M2 205	-	94 04	_	10-51 64-6 10-03	_	_	
-			M2 205	-	94 04	_	10-51 64-6 10-03	_	_	
-			M2 205	-	94 04	_	10-51 64-6 10-03	_	_	

Motivation

You would like to make a serie of **experiments (our case studies)**:

- 1. One data set, various privacy modeling parameters: K, L, (ε, δ)
 - 1.1. Consider: we do not know which K,L, ... are optimum!
 - 1.2. In our study we only experimented with **K-Anonymity model**
- 2. Multiple data sets, same privacy modeling parameters
- 3. Same data attributes, different data values, ...

ARX is an open source software library:

• You can add your features and use the library code.



ARX Anonymization Tool - Exa Re Edit View Help	ample IVX da [⊒⊡]+		Tandomatoro 1999 Selected N.2.6.1.2.1.1.9 Appled (N.2.6.1.2.1.1.9	
Configure transformation			Tundromations Lotto Seecoles (C.C.L.C.L.L.C.L.C.M. appres (C.C.L.C.L.L.M.	
and the second se	- phose serves (54, working			
each suge		i f B B 0	🔹 me 🔸 age 💊 mere 🔹 merital-status 🔸 education 🔺 native-country 🔅 versicient и 🤤	• ARX as a powerful GUI-based tool:
1 - Female 52 2 - Female 54 3 - Female 51 4 - Female 52	2 White 4 White 1 White 2 White	Divorced Some-college Ur. Divorced Bachelors Ur Divorced Masters Ur Divorced Some-college Ur	Type Quasi-identifying + Data type Intriger + Format Default Transformation: Generalization + Minimum: All + Maximum: All + Level-5 Level-3 Level-3 Level-4 -	 Multiple experiments: will be tedious
5 Penule 36 6 Penule 36 1 Penule 37	6 White 6 White 7 White	Divorced Bachelori Un Divorced Some-college Un Divorced Some-college Un	1 0-4 0-9 0-15 * 2 6-4 0-9 0-13 * 2 0-4 0-9 0-13 *	
8 Promate 60 9 Promate 52 16 Promate 52	2 White	Divorced Bachelurs Dr Separated Some-college Dr Wildowed Bachelurs Dr	4 04 09 015 3 04 09 015 9 59 09 015	 ARX as a Software Library:
11 v Fonale 30 12 v Male 51 13 v Male 52	9 White 1 White 2 White	Memed-spoise-abientBachekins Uk Memed-ca-spoise Bachekins Uk Mamed-ch-spoise Mastes Uk	7 59 09 019 7 8 59 09 018 7 9 59 09 018 7	
18 ≥ Male 54 15 ≥ Male 55 16 ≥ Male 55	4 White 5 White 2 White	Merried-cliv-spouse Bachelors Ui Married-cliv-spouse Masters Ui Married-cliv-spouse Some-college Ui	10 5-9 0-9 0-19 * -	 Possibility for extensions
17 2 Male 52 18 2 Male 54 19 2 Male 56	2 White 4 White 6 White	Married-div-spoule Associato Un Married-div-spoule Bedreters Un Married-div-spoule Bedreters Un	Privacy criteria Degulation + - Type Citerion Attribute () 5 Annywity	
20 2 Male 57 21 2 Male 58 22 2 Male 56	7 White 8 White 6 White	Married-ch-spouse Bachelurs Ur Married-ch-spouse Bachelurs Ur Married-ch-spouse Bachelurs Ur		
23 2 Male 58 34 2 Male 59 25 2 Male 51	8 White 9 White 1 Black	Married-div-spouse Badhelors Un Married-div-spouse Some-college Un Married-div-spouse Some-college Un	General satings _ UMMy messare Adminute weights [Coding medial	
30 9 Male 57	7 Flack	Married-cis-spouse Masters Us *		
search suitset		C D 🗃 🖷 🛛	Approximate Cause practical manetoxicity Precomputation (*) Enable Threshold 0.8	
See 5684 / 3163	2 ± 11,84	% Selection mode Query - manual		Experiments Declaration
				ARX GUI Repetitive experiments

Core ARX: Anonymization engine

Our Results

Our Method

ARX GUI

All the cases: declared and executed

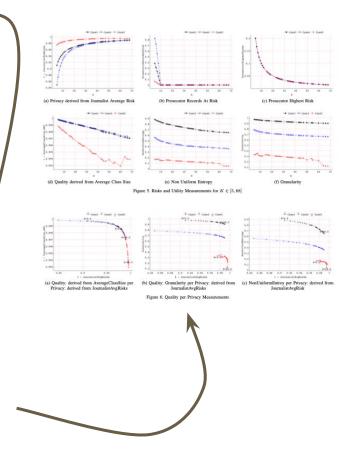
All the measurements are logged and visualized

Our main goal: behaviours of **<u>risk and utility</u>**



Repetitive experiments

Core ARX: Anonymization engine



Utility measures

General-purpose measures are studied: unpredictable data usage in Open Data settings.

- Average Equivalence Class Size
- Non-Uniform Entropy
- Granularity

Risk measures

Three risk measures:

- Prosecutor Record at Risk
- Journalist Average Risk
- Marketeer Success Rate

Results

Cases 1 and 2 are excluded in the visualization:

• They suppress the behaviour of others

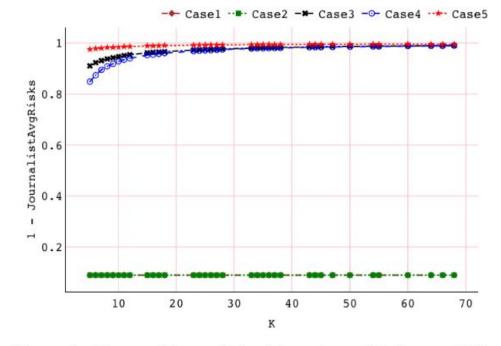


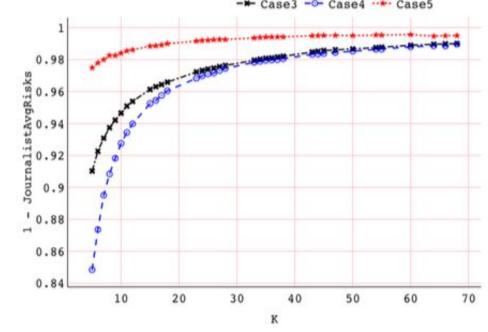
Figure 4. All cases Privacy derived from Journalist Average Risks

Results: Privacy

Higher k, results in higher Privacy

Case 5 (formal) is outperforming (cases 3 and 4 are pretty close)

- Case 3: Protection Against Data Linkage by Externs
- Case 4: Protection against all parties
- Case 5: Formal protection (ε-differential)

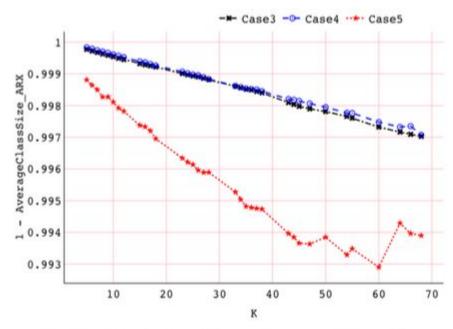


(a) Privacy derived from Journalist Average Risk

Results: Quality

ACS as an indication of information loss

- Quality is visualized
- Higher privacy, has the price of lower quality
- Cases 3, 4: very close

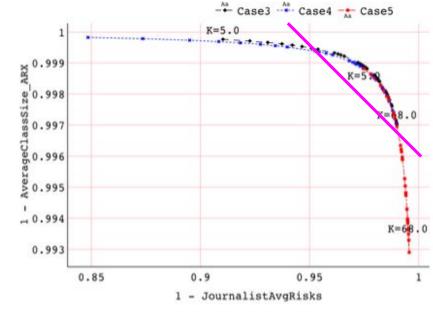


(d) Quality derived from Average Class Size

Results: Quality-Privacy

Putting all together:

- Case 5, all in a higher privacy
- Optimum area: can be helpful



(a) Quality: derived from AverageClassSize per Privacy: derived from JournalistAvgRisks

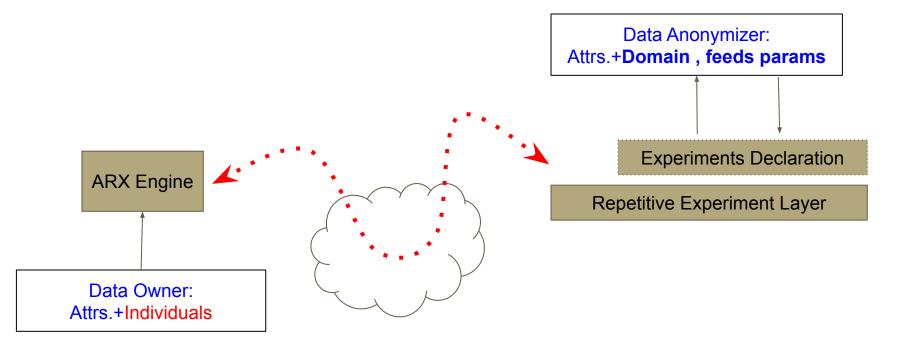
Conclusion

In our study:

- We analyzed various cases: considering external parties and data controllers.
- We applied SDC techniques with a range of analysis: Quality and Privacy.
- Results provide first steps as road map for data openers:
 - How to keep a balance between Quality and Privacy
- We plan for more data
- We expect graphs can be helpful to improve usability

Future Work: Reducing Risk of data anonymizer

Research Question: Is is possible to detach the tool expert from data content?



More experiments ...

To have more clear picture

- There is a need for more experiments on real data
 - Realistic environment, real context

