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 (a.amighi@hr.nl)
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

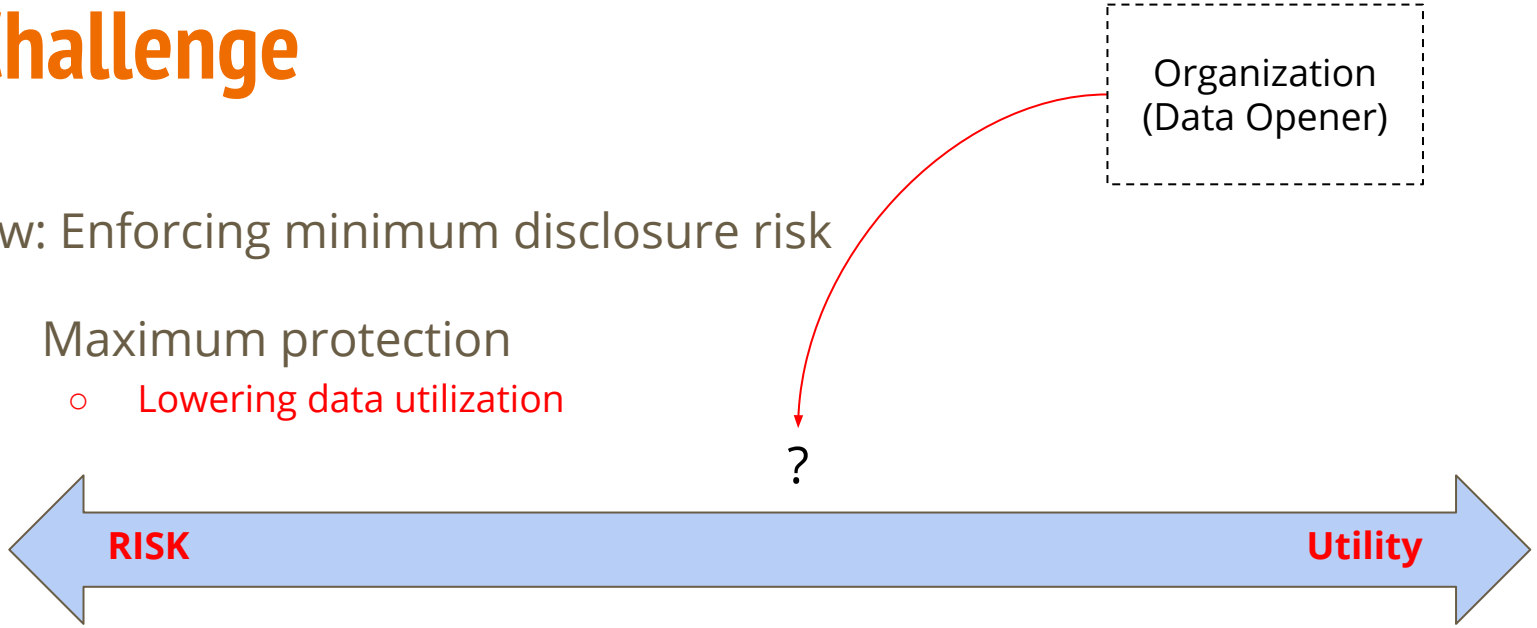
- Opening Data is crucial in innovations and economic growth:
  - Data Collections (huge) in organizations and enterprises
  - R&D needs data analysis
- Personal data disclosure as a main threat
  - Public data must be GDPR compliant
- Huge impact on key areas of society:
  - Healthcare, e-research,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

* GDPR: General Data Protection Regulation

Challenge

Law: Enforcing minimum disclosure risk

- Maximum protection
  - Lowering data utilization



Data Consumers (open data settings)

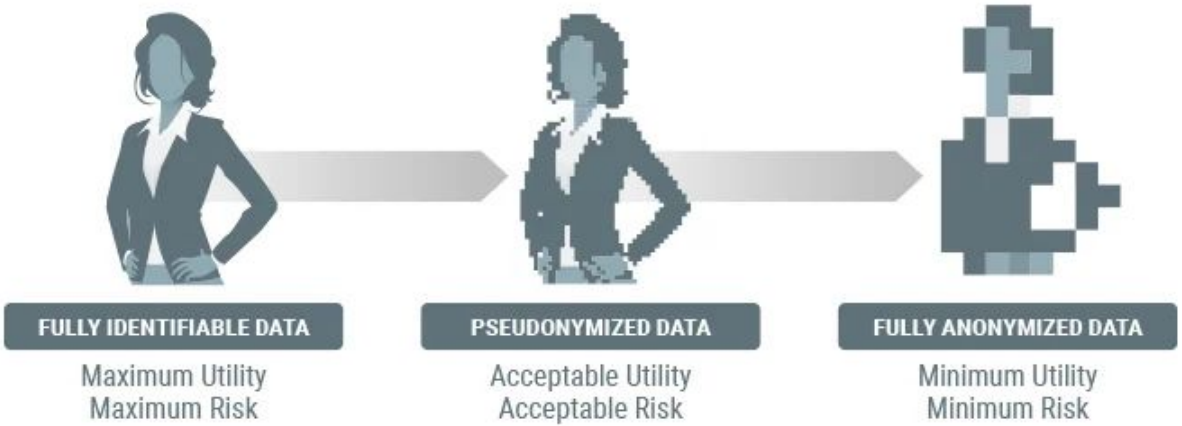
- Expecting high data utility
  - Higher risk of personal disclosure

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 measurements factors in choosing a right policy
- Five different cases
  - What are the implications?

Personal Data

Personal data collection:

- In various forms such as microdata (our focus), 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 (*QIDs*): subjective
- Sensitive Attributes (*SATs*): contextual
- Non-Sensitive Attributes (NATs)

Attribute Mapping

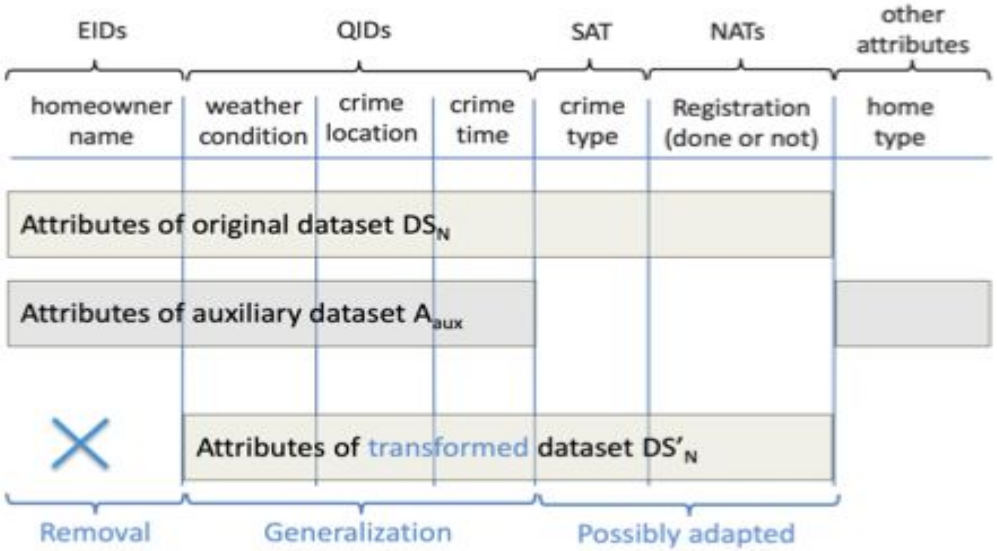


Figure 1. An illustration of attribute mapping for a data environment without the original microdata set.

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 more Formal Approaches:

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

ε-differential

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

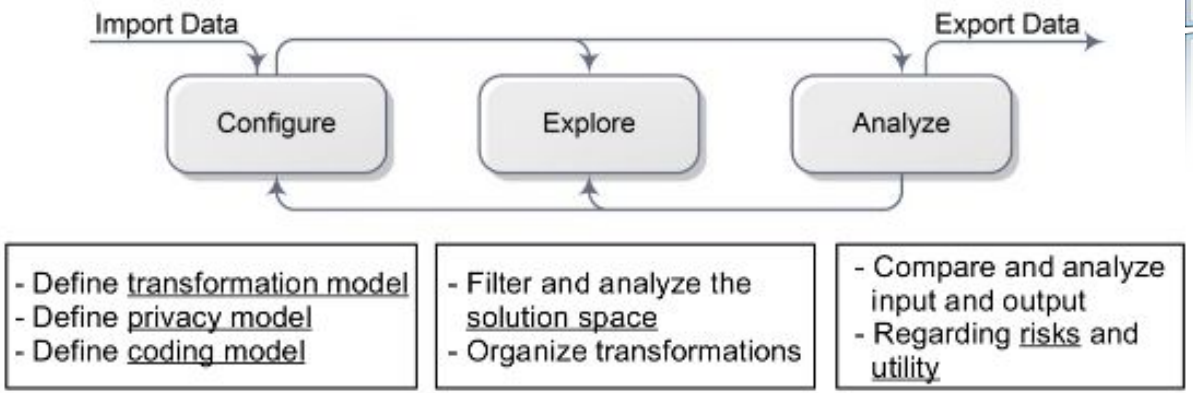
ARX:

- 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**.



Motivation

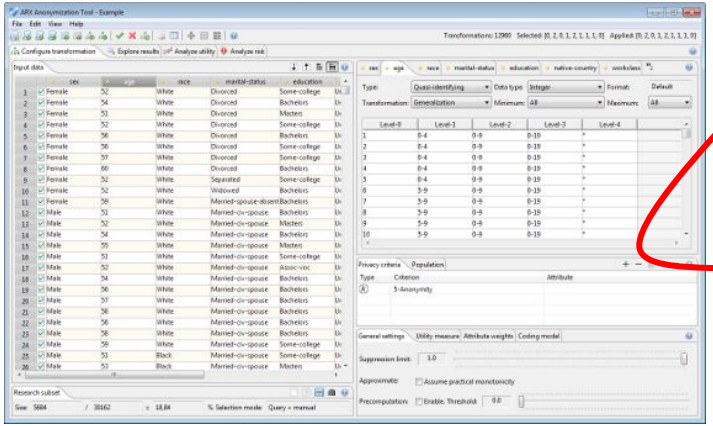
You would like to make a serie of **experiments (our case studies)**:

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, …

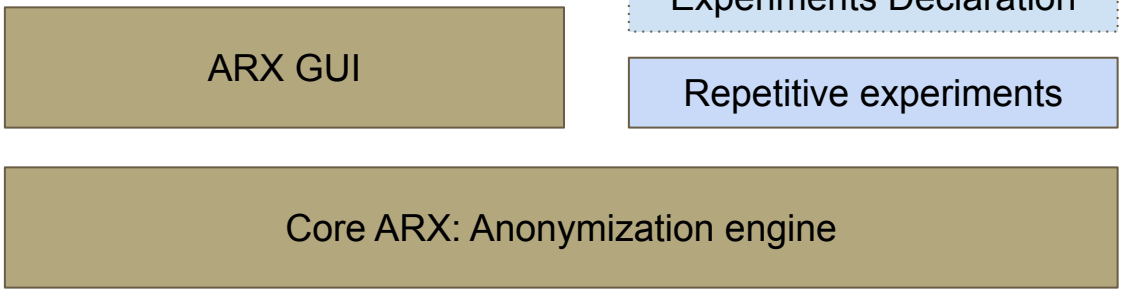
ARX is an open source software library:

- **You can add your features and use the library code.**

Using ARX



- ARX as a powerful GUI-based tool:
  - Multiple experiments: will be tedious
- ARX as a Software Library:
  - Possibility for extensions ...



Our Results

Our Method

All the cases: declared and executed

All the measurements are logged and visualized

Our main goal: behaviours of **risk and utility**

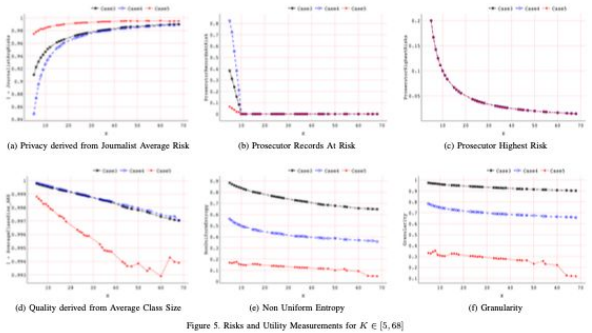


Figure 5. Risks and Utility Measurements for $K \in [5, 68]$

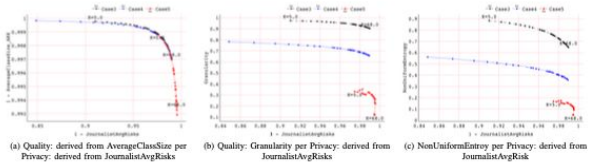


Figure 6. Quality per Privacy Measurements

Experiments Declaration

ARX GUI

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

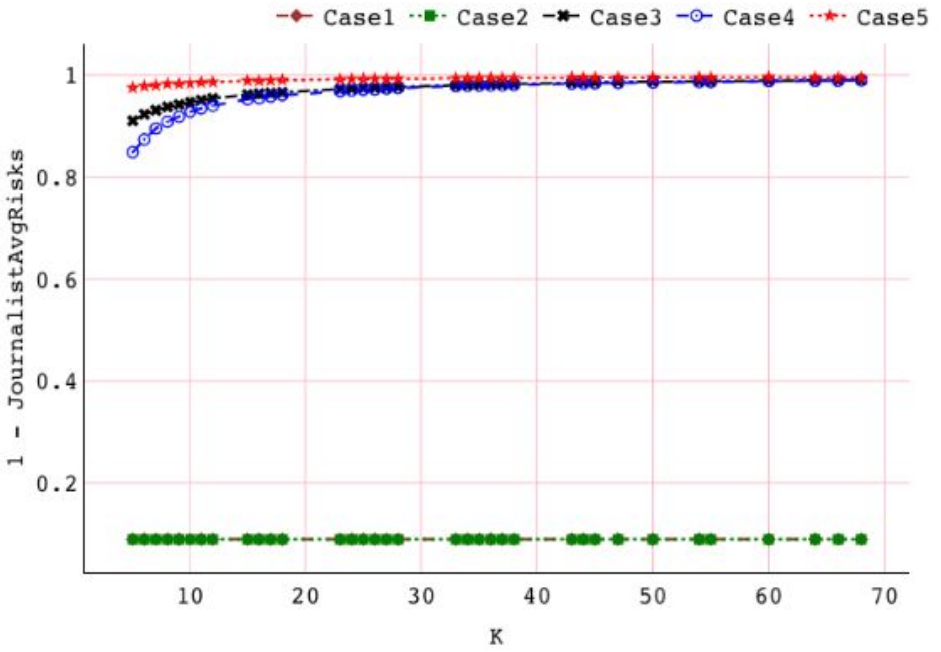


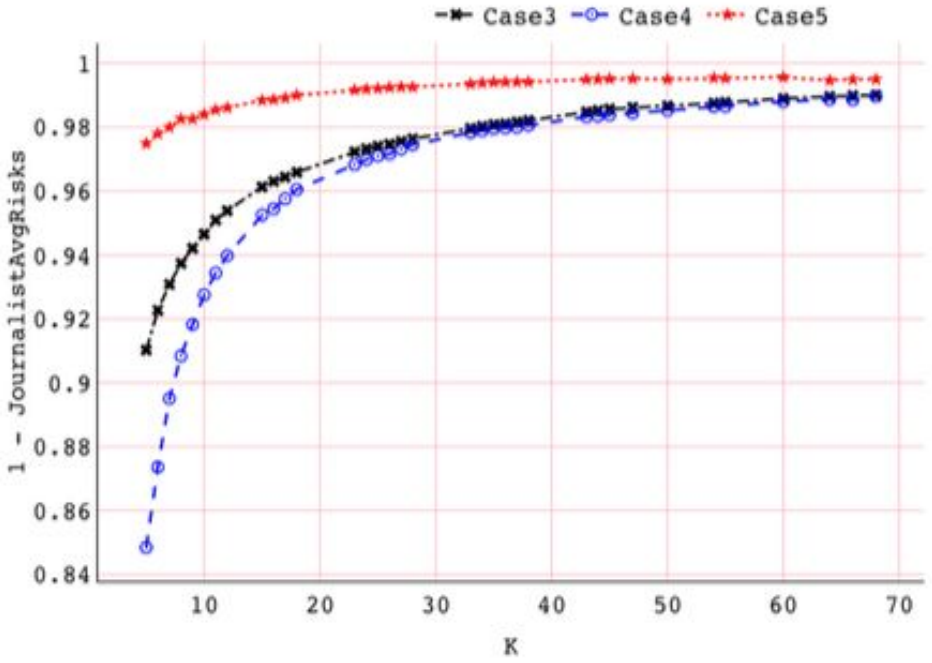
Figure 4. All cases Privacy derived from Journalist Average Risks

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)

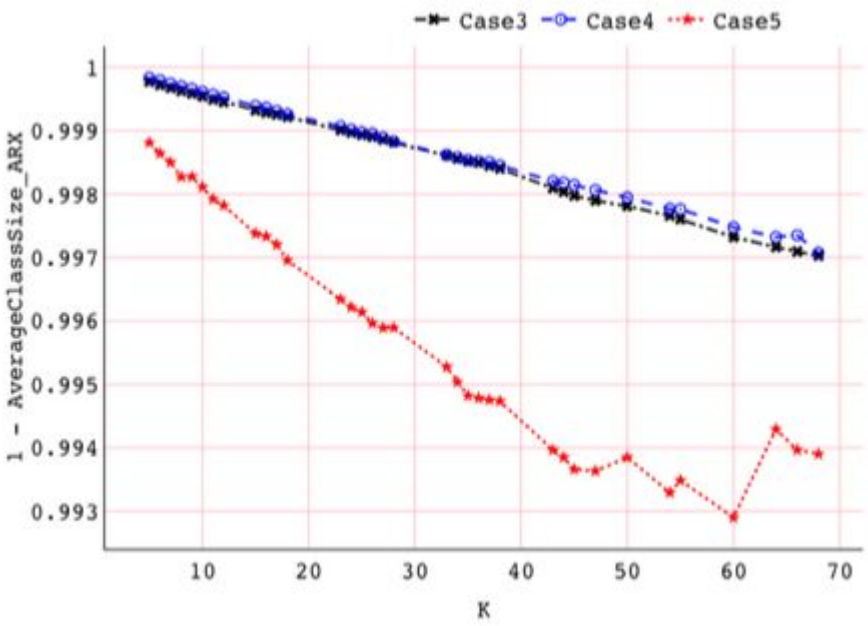


(a) Privacy derived from Journalist Average Risk

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

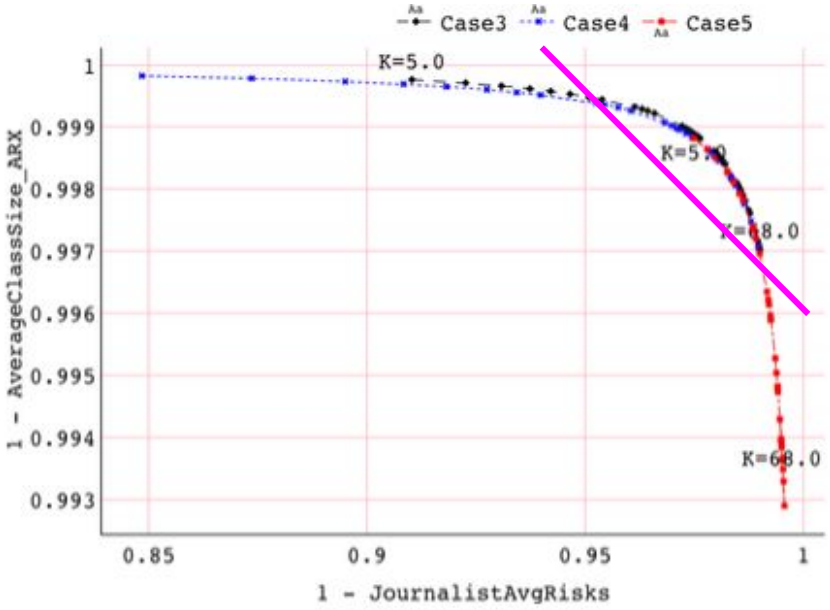


(d) Quality derived from Average Class Size

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

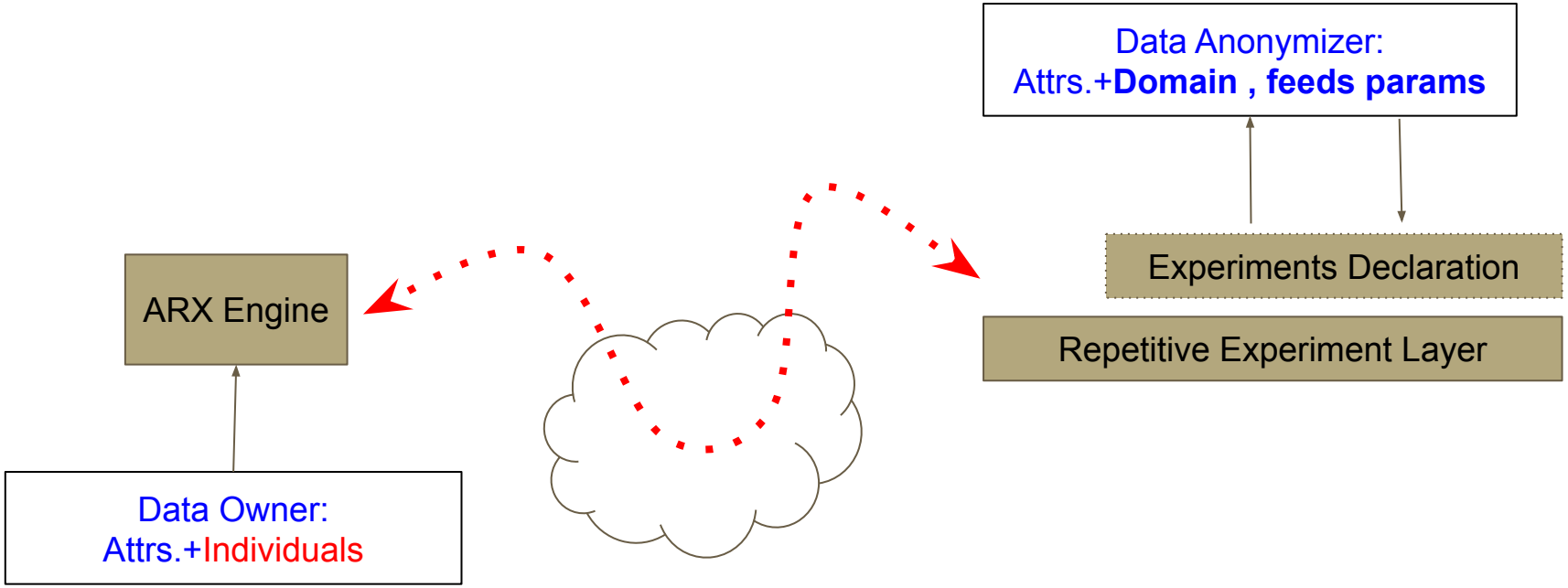
Conclusion

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
-