International Conference on the Digital Society: Protecting Privacy in Open (& Big) Data Settings

A Framework for a User-friendly Statistical Disclosure Control Tool

Anshika Rawat
Delft University of Technology, a.rawat-1@student.tudelft.nl
Mortaza S. Bargh
Research and Documentation Centre, Ministry of Justice and Security, m.shoae.bargh@wodc.nl
Afshin Amighi
Rotterdam University of Applied Sciences, a.amighi@hr.nl
Marijn Janssen
Delft University of Technology, m.f.w.h.a.janssen@tudelft.nl

Wetenschappelijk Onderzoek- en
Documentatiecentrum
Ministerie van Veiligheid en Justitie

IARIA

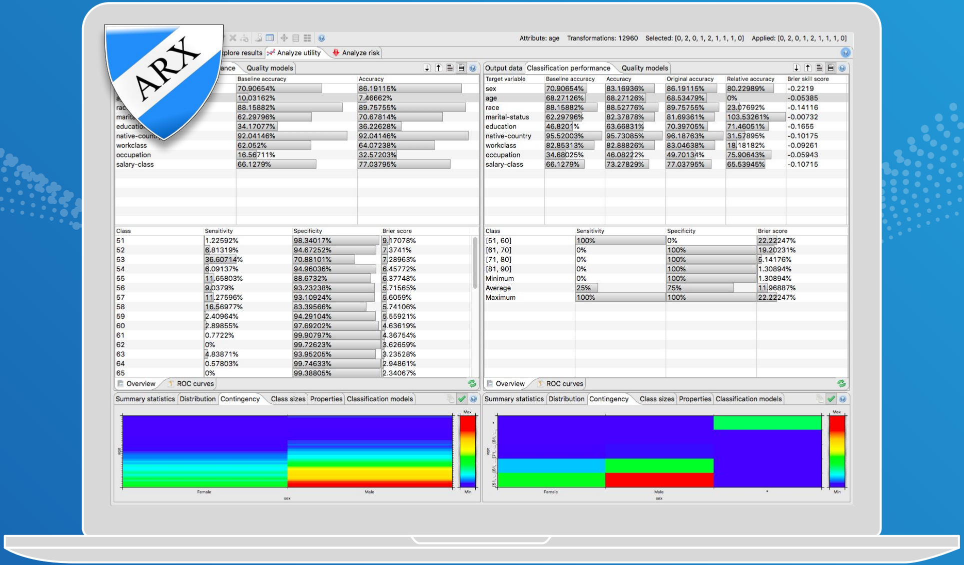
TU Delft Delft University of Technology

About me

- Background in computer science
- Masters in Management of Technology
- Interests lie in privacy, cyber security, data analytics, digital transformations

ARX

- Open-source application
- Supports several SDC techniques
- Undergoes regular updates



Attribute: age Transformations: 12960 Selected: [0, 2, 0, 1, 2, 1, 1, 1, 0] Applied: [0, 2, 0, 1, 2, 1, 1, 1, 0]

	Baseline accuracy	Accuracy
	70.90654%	86.19115%
a	10.03162%	7.46662%
rac	88.15882%	89.75755%
marita	62.29796%	70.67814%
educatio	34.17077%	36.22628%
native-count	92.04146%	92.04146%
workclass	62.052%	64.07238%
occupation	16.56711%	32.57203%
salary-class	66.1279%	77.03795%

Class	Sensitivity	Specificity	Brier score
51	1.22592%	98.34017%	9.17078%
52	6.81319%	94.67252%	7.3741%
53	36.60714%	70.88101%	7.28963%
54	6.09137%	94.96036%	6.45772%
55	11.65803%	88.6732%	6.37748%
56	9.0379%	93.23238%	5.71565%
57	11.27596%	93.10924%	5.6059%
58	16.56977%	83.39566%	5.74106%
59	2.40964%	94.29104%	5.55921%
60	2.89855%	97.69202%	4.63619%
61	0.7722%	99.90797%	4.36754%
62	0%	99.72623%	3.62659%
63	4.83871%	93.95205%	3.23528%
64	0.57803%	99.74633%	2.94861%
65	0%	99.38805%	2.34067%

Target variable	Baseline accuracy	Accuracy	Original accuracy	Relative accuracy	Brier skill score
sex	70.90654%	83.16936%	86.19115%	80.22989%	-0.2219
age	68.27126%	68.27126%	68.53479%	0%	-0.05385
race	88.15882%	88.52776%	89.75755%	23.07692%	-0.14116
marital-status	62.29796%	82.37878%	81.69361%	103.53261%	-0.00732
education	46.8201%	63.66831%	70.39705%	71.46051%	-0.1655
native-country	95.52003%	95.73085%	96.18763%	31.57895%	-0.10175
workclass	82.85313%	82.88826%	83.04638%	18.18182%	-0.09261
occupation	34.68025%	46.08222%	49.70134%	75.90643%	-0.05943
salary-class	66.1279%	73.27829%	77.03795%	65.53945%	-0.10715

Class	Sensitivity	Specificity	Brier score
[51, 60]	100%	0%	22.22247%
[61, 70]	0%	100%	19.20231%
[71, 80]	0%	100%	5.14176%
[81, 90]	0%	100%	1.30894%
Minimum	0%	100%	1.30894%
Average	25%	75%	11.96887%
Maximum	100%	100%	22.22247%

The Problem

- The tools are designed from the perspective of experts

- Slow progress in their development has resulted in limited support material and even smaller user base.

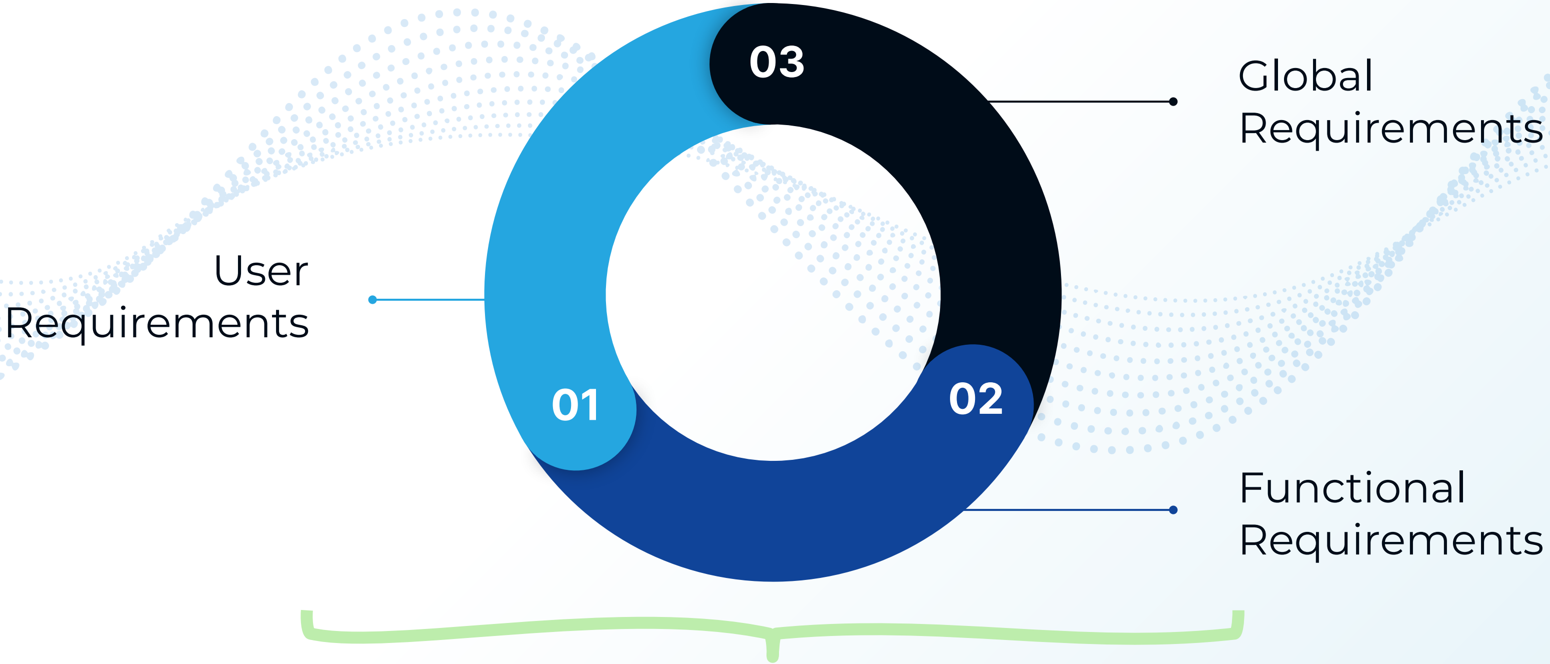
Potential Solution (Research Objective)

1. **Addressing the complexity of ARX** which is the knowledge needed to understand the concept of SDC

2. **Increasing the software usability**, to make it easier for entry-level users to adopt it without depending on external support material

Through the design of a new SDC tool

Prototype Development



Simulating the task of a complete data anonymization process

Usability Problems with ARX

1. Minimal Memory Load

ARX requires its users to recall from memory a great deal of information to complete a task

2. Self-descriptiveness

ARX systems lacks self-explanatory features. This is compounded by a lack of external supporting documentation

3. User Guidance

The UI does not provide clues to guide users on how to use its features collectively

4. Navigability

The design elements of ARX impede a smooth navigation experience for the user

5. Minimal Action

Lack of information and guidance leads to users finishing a task in more number of steps than actually intended

6. Familiarity

Given the extensiveness of ARX's features, it's design such as content display does not invoke feelings of familiarity in the user

User Requirements (I)

Problem Area	Solution (User requirement)
Minimal Memory Load	- Minimalistic design to avoid visual clutter
	- Consistent interface elements based on existing mental models
	- Offloading tasks by using default values or visual clues for decision making
Self-descriptivene ss	- Intrinsic methods to relay information
	- Use of simple, unassuming language
	- Providing contextual functions and information
	- Instinctive placing of visual metaphors
User Guidance	- Principle of tunnelling and selective attention through multi-step pathway forms with inline validation for task completion

User Requirements (II)

Problem Area	Solution (User requirement)
Navigability	- Defining a clear primary navigation area
	- Minimal hierarchical structures that embrace predictability such as a left-hand side navigation menu
Minimal Action	- Streamlining and grouping similar task actions on one page/tab of the screen
Familiarity	- Incorporating predictable design elements in pace with current trends

Simplifying Functions

- ARX has a range of features that can overwhelm new users

- **Paradox of Choice:** An overload of options does not necessarily lead to better results

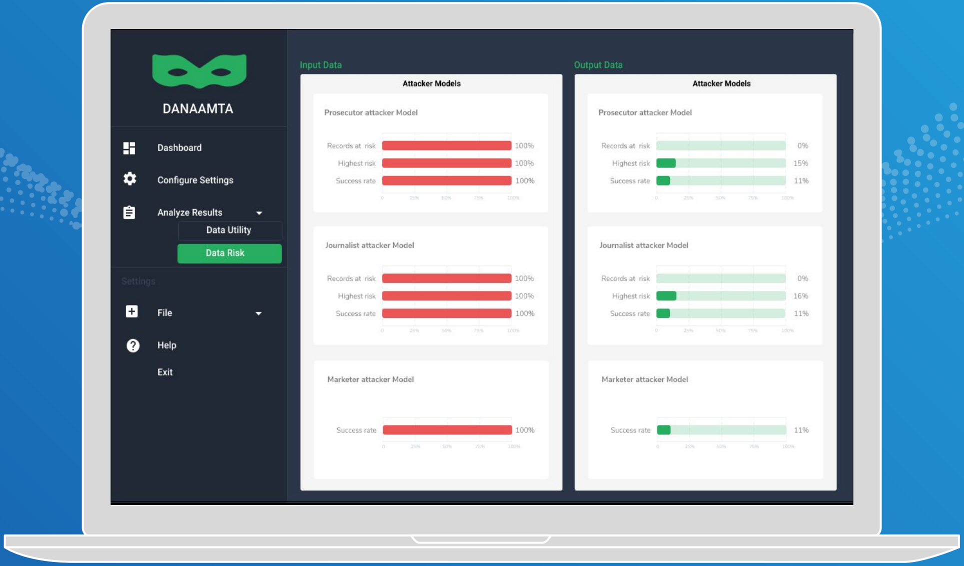
- Providing users with fewer options can result in them making decisions without facing decision fatigue

Functional Requirements

Function	Description
Anonymization Approach	- Privacy model approach
Data Utility Measures	- General-purpose metrics like Average Equivalence Class Size, Non-Uniform Entropy and Granularity
Risk Measures	- Risk evaluation metrics based on the Prosecutor, Journalist, and Marketer Attacker Models
General Configurations	- Suppression Limit

DANAAMTA

- Simplified the task of data anonymization process
- Guiding users from point A to point B
- No overload of expert-level concepts



Prototype

In Practice

- The prototype can be used as a stepping stone to expose entry-level users of an organization to the field of SDC without overwhelming them with its complexities

- Through micro-learning employees can be managed to move on to much more advanced tools (ARX) which might be a more practical approach given the complexity of actual data sets

Future Work

1. Integrating the prototype with the **APIs of ARX** to provide a fully functional tool. Such a prototype can be **better evaluated** by comparing the results of anonymizing the same data set with ARX and the prototype
2. To evaluate the prototype with **larger sample size or within the context of the organisation** such as participants who could be the potential data processors
3. A **similar study can be conducted with experts** to see the difference between the different user levels and their preferences
4. Incorporating other approaches to data anonymization such as **differential privacy**