$\mathbb{X}^2\mathbb{R}^2$: a Tool for Explainable and Explorative Reidentification Risk Analysis

Tom Rolandus Hagedoorn
Eurecat, Spain
tom.rolandus@eurecat.org

Rohit Kumar
Eurecat
Barcelona, Spain
rohit.kumar@eurecat.org

Francesco Bonchi
ISI Foundation, Italy
Eurecat, Spain
francesco.bonchi@isi.it

ABSTRACT

Reidentification-risk analysis and anonymity have received a great deal of attention in the last two decades. While the research community has been developing several privacy notions and the algorithms to achieve them, these tools have faced difficulties in being transferred to the wider audience of practitioners, for they require a considerable amount of data privacy technical knowledge.

We demonstrate $\mathbb{X}^2\mathbb{R}^2$ (Explainable Explorative Reidentification Risk), a *data anonymization tool for the laymen.* $\mathbb{X}^2\mathbb{R}^2$ guides the user through a transparent explorative process, during which the existing reidentification risks are explained and quantified, possible data transformation options are recommended, and the consequences of these operations, in terms of privacy risk and data utility, are clearly shown.

PVLDB Reference Format:
Tom Rolandus Hagedoorn, Rohit Kumar, Francesco Bonchi. $\mathbb{X}^2\mathbb{R}^2$: a Tool for Explainable and Explorative Reidentification Risk Analysis. *PVLDB*, 13(12): xxxx-yyyy, 2020.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx

1. INTRODUCTION

In the last two decades, the landscape of data privacy has evolved drastically. The rise of public awareness concerning the risks related to the ubiquity of data, in combination with the pressure of privacy advocates, has led to stronger data privacy regulations. In the EU, the General Data Protection Regulation (GDPR) [11], enforced since May 2018, marks the latest governmental attempt at protecting and regulating user data. These regulations clearly define concepts such as personally identifiable information, sensitive attributes, user consent, right to explanation, etc., as well as defining the hefty fines incurred for breaching these regulations [10]. However, they provide minimal information on how to de-identify a dataset, although they advice the data curators to do so. Therefore, an increasing amount of individuals with a lack of extensive data privacy knowledge find themselves in a position of the data controller, without fully understanding how to comply with the lengthy new regulations. This knowledge gap is particularly exacerbated in small and medium enterprises [8].

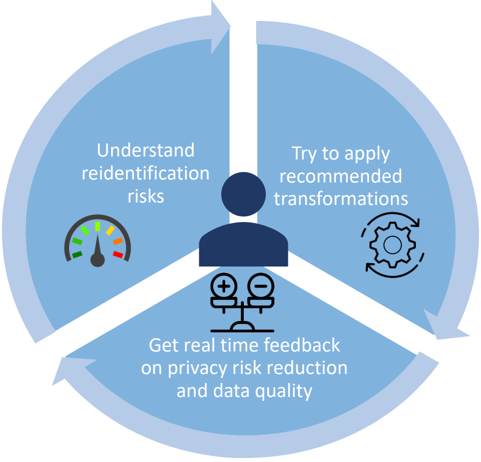

Understand reidentification risks
Try to apply recommended transformations
Get real time feedback on privacy risk reduction and data quality

Figure 1: Depiction of the $\mathbb{X}^2\mathbb{R}^2$ data anonymization process with the *human-in-the-loop.* The tool explains the reidentification risks currently existing in the data and recommends a set of the best data transformation operations. The user can try out the different recommended data transformations and monitor their effects on the trade-off between reidentification risk and data quality.

Parallelly, the research community has been developing privacy notions such as k-anonymity [7, 9], l-diversity [4], t-closeness [3], differential privacy [1] (just to mention a few). Efficient algorithms to achieve these privacy notions have been developed and they have been made available in several libraries and tools. However, these tools require proficiency with advanced data analysis and privacy concepts, have many parameters, and their execution remains opaque. Due to these limitations, these privacy notions and tools have faced difficulties in being transferred to the wider audience of practitioners.

Due to this gap between strict regulations applying to an increasingly wide audience and a lack of means to take actions in order to comply with these regulations, it is important to develop tools that open the black-box of anonymization, allowing laymen to identify and understand the reidentification risks present in their data, and anonymize their data without the need to set complex parameters or acquire sophisticated privacy notions.

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/xxxxxxx.xxxxxxx

$\mathbb{X}^2\mathbb{R}^2$(Explainable Explorative Reidentification Risk) is a tool aimed at filling this gap. The main idea behind $\mathbb{X}^2\mathbb{R}^2$ is to guide the user through a transparent explorative process of data anonymization, explaining the various data transformation options and their consequences in terms of privacy and data utility.

During this explorative process the user comes to understand the reidentification risks of their data, how different data transformation operations affect their data, and thus become able to drive the anonymization process, keeping in consideration the business needs and constraints, while keeping an eye on the intrinsic trade-off between reidentification risks and data utility [5, 2].

Figure 1 depicts the vision of the iterative, explorative $\mathbb{X}^2\mathbb{R}^2$ data anonymization process:

- The user uploads the database under analysis: i.e., a relational table where each row corresponds to an individual whose identity and privacy we need to protect.
- The tool explains the reidentification risks currently existing in the data, by explicitly showing to the user single tuples (or small groups of tuples), that can be easily reidentified due to the uniqueness of an attribute or a combination of attributes.
- The tool then recommends to the user a set of different data transformation operations: i.e., generalizing certain attributes, or suppressing rows.
- The user can apply some of these actions in the same way as applying a filter to a picture, seeing immediately the resulting dataset, an estimation of the corresponding reduction in privacy risks, while getting an estimate of the inherent reduction in data utility.

By iterating through these steps the users can explore their data, deciding on the best data transformation strategy and producing, in the end, an anonymized table. This process can be seen as a data anonymization algorithm (i.e., k-anonymity by attributes generalization and rows suppression) but with the *human-in-the-loop*.

$\mathbb{X}^2\mathbb{R}^2$ is just a component of a larger system developed within SMOOTH, a European Union's Horizon 2020 research and innovation project[1]. The goal of SMOOTH is to assist micro enterprises in adopting and complying with the GDPR by designing and implementing easy-to-use and affordable tools and thus helping strengthening the awareness on their GDPR obligations and analyzing their level of compliance with the new data protection regulation.

The next section provides an overview of the tool, while Section 4 discusses the organization of the demo.

2. SYSTEM OVERVIEW

$\mathbb{X}^2\mathbb{R}^2$ is designed following closely the core guiding principles that we defined, namely the platform ought to be intuitive with as few parameters as possible, provide explainable recommendations, and actively involve the user in the process, that is, have a human-in-the-loop approach. We adopted an iterative development process: early releases were distributed to end users within the SMOOTH project consortium, and their feedback was implemented in order to ensure that the tool is simple and intuitive to use.

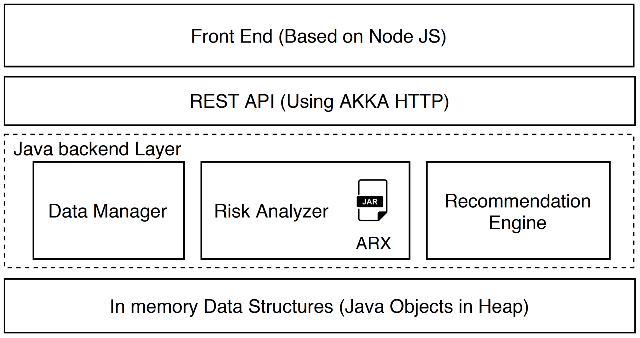

Front End (Based on Node JS)
REST API (Using AKKA HTTP)
Java backend Layer
Data Manager
Risk Analyzer
JAR
ARX
Recommendation Engine
In memory Data Structures (Java Objects in Heap)

Figure 2: $\mathbb{X}^2\mathbb{R}^2$ system architecture.

The tool is developed using Java for the backend logic and Node.js[2] – an open-source, cross-platform, JavaScript runtime environment – for the frontend dashboard interface. The dashboard interacts with the backend using REST API endpoints exposed by the Java back-end. The REST API endpoints are developed using AKKA HTTP interfaces[3]. Figure 2 shows the block diagram of the system architecture. This architecture design is very modular, with the REST API separating the frontend from the backend, which makes it extremely easy to extend the system by, e.g., adding new input formats, adding new risk analysis algorithms while keeping the frontend unchanged. The code of $\mathbb{X}^2\mathbb{R}^2$ is available at `https://github.com/rohit-nlp/x2-r2.git`.

The backend layer is divided into three main components, described next.

Data Manager. This component is responsible for handling different input data formats. The uploaded data is converted into an in-memory data structure.

Risk Analyzer. This component is responsible of (*i*) running reidentification risk analysis, and (*ii*) measuring utility loss in the data after each data transformation is applied. For reidentification risk analysis we take benefits of open source library ARX [6]. In particular, we use the size of each equivalence class of tuples w.r.t. each possible combination of attribute-values as a measure of reidentification risk: we define *Highest Risk* as the reciprocal of the size of the smallest equivalence class, and *Average Risk* as the reciprocal of the average size of equivalence classes (both then reported on a 0-100 scale). For utility loss we adopt a simple measure of the distance between the original table and the table after data transformation.

Recommendation Engine. This engine provides the data transformation recommendations. In particular, similarly to the algorithms for k-anonymity [7, 9], as data transformation operations $\mathbb{X}^2\mathbb{R}^2$ provides attribute generalization and the record suppression. The system uses an in-memory graph based data structure to support efficient traversal over applied recommendations history. After every recommendation is applied, this engine updates the data state in-memory and updates the risk and utility loss metrics accordingly.

The platform frontend, presented in detail in the next section, is designed as a dashboard, where most of the information is directly available to the user.

[1] `https://smoothplatform.eu/`

[2] `https://nodejs.org/`

[3] `https://akka.io/`

3. DASHBOARD

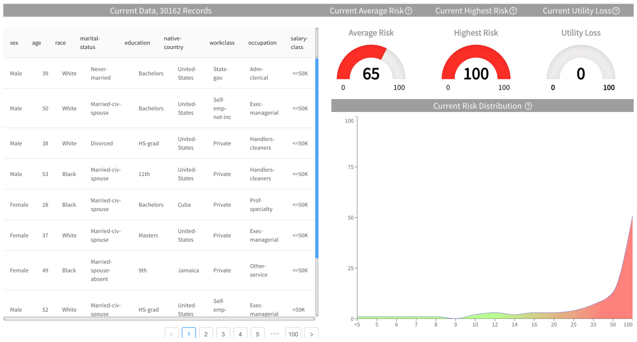
The dashboard interface is organized in three main views: *table view* (Figure 3) on the top-half of the screen, *attributes view* (Figure 4) and *rows view* (Figure 5) which share the bottom half of the screen in two different tabs.

3.1 Table view

The table view provides a global overview of the current state of the data. The left side shows a tabular view of the dataset, while the right side of the panel displays the reidentification risks and the utility loss with respect to the original data. The reidentification risk is presented in terms of *Highest Risk* and *Average Risk* as explained before, together with a risk distribution plot, showing the fraction of population which has a certain level of reidentification risk, where the risk of a tuple is defined as the reciprocal of the size of the equivalence class to which the tuple belongs (multiplied by 100). Throughout the illustrations, an intuitive green to red colour scheme is used to guide the user towards good values. For instance, low values for the gauges are represented by green and high values by red. The same reasoning applies for the risk distribution with low and high risk respectively, however here the area of the colour also guides the user. The reidentification risk of the dataset decreases as the red area decreases and the green area increases. Decreasing the reidentification risk entails a loss in data utility as is displayed in the third gauge. Initially, the utility loss will be zero as no recommendation is applied and the data is thus in its original state. Once a recommendation is applied, the gauge is updated to reflect the resulting loss in utility with respect to the original dataset.

Each measure is explained by a short text visible by hovering over the ? symbol. Furthermore, the system explains the value of *Highest Risk*, by showing the tuples which have the highest risk of reidentification in the rows view, and the attributes which are responsible for the largest fraction of the risk in the attributes view.

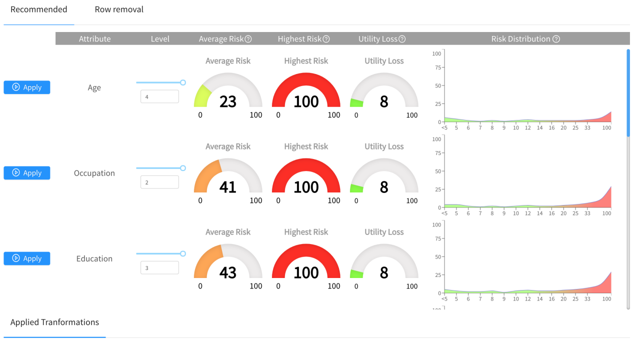
To guide the user through analyzing the reidentification risk of the dataset and decreasing it, recommendations are displayed that represent the optimal transformations to apply with respect to the risk and utility measures. The recommendations, which are the most innovative feature of $\mathbb{X}^2\mathbb{R}^2$, have the added benefit of simultaneously explaining where the risk comes from. Figures 4 and 5 display the left and right tab of the bottom half of the dashboard, which are the attribute generalization and row suppression recommendations respectively.



Current Data, 30162 Records

sex	age	race	marital-status	education	native-country	workclass	occupation	salary-class
Male	39	White	Never-married	Bachelors	United-States	State-gov	Adm-clerical	<=50K
Male	50	White	Married-civ-spouse	Bachelors	United-States	Self-emp-not-inc	Exec-managerial	<=50K
Male	38	White	Divorced	HS-grad	United-States	Private	Handlers-cleaners	<=50K
Male	53	Black	Married-civ-spouse	11th	United-States	Private	Handlers-cleaners	<=50K
Female	28	Black	Married-civ-spouse	Bachelors	Cuba	Private	Prof-specialty	<=50K
Female	37	White	Married-civ-spouse	Masters	United-States	Private	Exec-managerial	<=50K
Female	49	Black	Married-spouse-absent	9th	Jamaica	Private	Other-service	<=50K
Male	52	White	Married-civ-spouse	HS-grad	United-States	Self-emp-	Exec-managerial	>50K

Figure 3: Table view (top-half of the dashboard). On the left-hand side a tabular overview of the data and on the right-hand side the reidentification risk and utility loss panel.



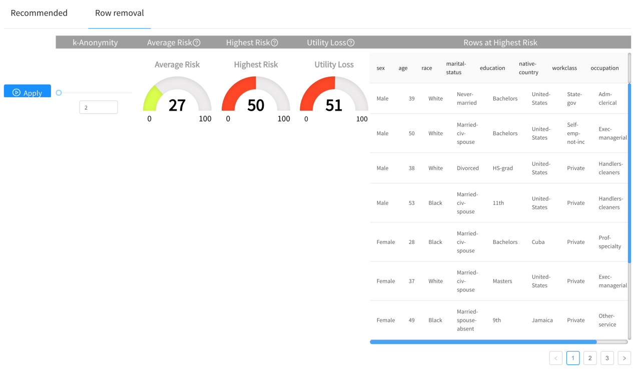
	Attribute	Level	Average Risk ?	Highest Risk ?	Utility Loss ?	Risk Distribution ?
Apply	Age	4	23	100	8	
Apply	Occupation	2	41	100	8	
Apply	Education	3	43	100	8	

Figure 4: Attributes view (bottom-half of the dashboard, first tab). Ranked list of the attributes which contribute more to the overall reidentification risk. The system offer the possibility of generalizing them to different levels and seeing in real-time the effects that such data transformation has on the risk and on the utility of the data.

3.2 Attributes view

The attributes view (left tab of the bottom-half of the interface) shows a ranked list of the attributes which are most responsible for reidentification risk and recommends to the user to generalize them.

Generalizing attributes is one of the main data-transformation operations towards achieving many different privacy policies. For this purpose, attribute generalization hierarchies are needed, and usually are requested to be provided by the user. In the spirit of limiting user-defined parameters and interventions, $\mathbb{X}^2\mathbb{R}^2$ automatizes the generalization hierarchies creation. For numerical attributes this is straightforwardly achieved by creating a hierarchy of nested intervals, which are defined by taking in consideration the distribution of values and a predefined depth of the hierarchy. This is less trivial for categorical values, as in this case an attribute generalization hierarchy is normally based on the semantic meaning of the levels of the attributes. For example, countries could be generalized to geographical regions such as Europe. However, numerous ways of grouping the same elements often exist, and choosing the most appropriate one is not always straightforward and is often subjective. Given the complexity inherent to automatically creating hierarchies based on the semantic meaning, we derived a different approach. The idea used is akin to that of k-anonymity, namely to make the values indistinguishable from each other by grouping them in sets of values. The sets of the first level of the generalization hierarchy are created by grouping together the least and most frequent values of the attributes, then the second least and second most frequent values, and so forth. This logic is repeated for the following levels, with the values being replaced by the sets of the previous level, until the set containing all values is obtained. When there is an odd number of values, or sets, the least and second least frequent values are grouped with the most frequent one. This insures that the distribution of the sets in the data are fairly even.



sex	age	race	marital-status	education	native-country	workclass	occupation
Male	39	White	Never-married	Bachelors	United-States	State-gov	Adm-clerical
Male	50	White	Married-civ-spouse	Bachelors	United-States	Self-emp-not-inc	Exec-managerial
Male	38	White	Divorced	HS-grad	United-States	Private	Handlers-cleaners
Male	53	Black	Married-civ-spouse	11th	United-States	Private	Handlers-cleaners
Female	28	Black	Married-civ-spouse	Bachelors	Cuba	Private	Prof-specialty
Female	37	White	Married-civ-spouse	Masters	United-States	Private	Exec-managerial
Female	49	Black	Married-spouse-absent	9th	Jamaica	Private	Other-service

Figure 5: Rows view (bottom-half of the dashboard, second tab): shows the tuples that are at highest risk of reidentification and provides the possibility to suppress them.

In the attributes view, the user can experiment with the different recommended attribute generalizations, as well as different levels of generalization for each of them, while seeing in real-time the effects that such data transformations have on the reidentification risk and on the utility loss of the data.

3.3 Rows view

The second tab of the bottom half of the dashboard contains the rows view. It shows the tuples that are at highest risk of reidentification and offers the possibility to suppress them. To decide on the number of records to suppress, the user sets the slider to the desired level of k-anonymity. The user does not need to understand the k-anonymity factor, as the gauges help identifying an appropriate level.

Both types of recommendations (attributes to be generalized and tuples to be suppressed) are instantly updated as soon as either has been applied, and in both cases the user can easily undo an applied data transformation. The table view on the top of the dashboard keeps updating instantly while offering a tabular view of the current state of the dataset. This allows the user to explore the possible data-transformation operations while immediately seeing the effects on the data. This process continues until the desired levels of reidentification risk and data utility loss have been achieved. Simultaneously, the user develops an understanding of the reidentification risks existing in the dataset and the inherent trade-off between mitigating these risks and maintaining data quality.

4. DEMONSTRATION OUTLINE

During the demonstration the audience will interact with all the functionalities of $\mathbb{X}^2\mathbb{R}^2$ on a standard census database. Only minimal prior explanations will be provided given that the intuitiveness of the tool, which is one of its core features, makes it easy even for non-expert users to browse through its functionalities.

The users will be first requested to upload the database in $\mathbb{X}^2\mathbb{R}^2$. A simple glance at the table view will provide a high level summary of the state of the data and the reidentification risk. Once the users have an overview of the data, they can move to the lower section of the dashboard, which is divided into the two tabs depicted in the Figures 4 and 5. Here, the users will find the recommendations regarding attribute generalization and record suppression. The users are encouraged to experiment with the recommendations of both tabs in any order that they see fit. To do so, they can apply and undo recommendations while simultaneously watching the impact that it has on the risk and utility measures. The idea of this experimentation is for the users to develop an understanding of what causes the reidentification risk in the dataset and how it can be decreased while controlling for the utility loss.

5. ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 786741 (SMOOTH).

6. REFERENCES

[1] C. Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC'08, page 1–19, 2008.

[2] F. Kohlmayer, F. Prasser, and K. A. Kuhn. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of biomedical informatics*, 58:37–48, 2015.

[3] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–es, Mar. 2007.

[5] B. Malin, D. Karp, and R. H. Scheuermann. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *Journal of Investigative Medicine*, 58(1):11–18, 2010.

[6] F. Prasser, F. Kohlmayer, R. Lautenschlaeger, and K. A. Kuhn. Arx-a comprehensive tool for anonymizing biomedical data. In *AMIA Annual Symposium Proceedings*, volume 2014, page 984. American Medical Informatics Association, 2014.

[7] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, 1998.

[8] S. Sirur, J. R. Nurse, and H. Webb. Are we there yet?: Understanding the challenges faced in complying with the general data protection regulation (gdpr). In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, pages 88–95. ACM, 2018.

[9] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[10] C. Tankard. What the gdpr means for businesses. *Network Security*, 2016(6):5–8, 2016.

[11] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.