Demo : Data Minimization and Informed Consent in Administrative Forms

Nicolas Anciaux
Inria, Saclay, France
Université Paris Saclay - Versailles, Saclay, France,
nicolas.anciaux@inria.fr

Sabine Frittella
LIFO, INSA Centre Val de Loire, Bourges France
Université d'Orléans, Orléans, France
sabine.frittella@insa-cvl.fr

Baptiste Joffroy
LIFO, INSA Centre Val de Loire, Bourges France
Université d'Orléans, Orléans, France
baptiste.joffroy@insa-cvl.fr

Benjamin Nguyen
LIFO, INSA Centre Val de Loire, Bourges France
Université d'Orléans, Orléans, France
benjamin.nguyen@insa-cvl.fr

ABSTRACT

This article proposes a demonstration implementing the data minimization privacy principle, focusing on reducing data collected by government administrations through forms. Data minimization is defined in many privacy regulations worldwide, but has not seen extensive real-world application. We propose a model based on logic and game theory and show that it is possible to create a practical and efficient solution for a real French welfare benefit case.

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

Privacy, Data minimization, GDPR, Informed consent.

ACM Reference Format:
Nicolas Anciaux, Sabine Frittella, Baptiste Joffroy, and Benjamin Nguyen. 2023. Demo : Data Minimization and Informed Consent in Administrative Forms. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23), November 26–30, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3576915.3624363

1 INTRODUCTION

Overview. In this demonstration, we propose a tool to assist users with the problem of the *minimization of personal data* collected in the context of administrative forms. *Data minimization* is a fundamental privacy concept defined in many privacy regulations (such as the European General Data Protection Regulation or GDPR [7]) that states that only the data necessary to take a decision must be collected and stored. Data minimization has yet to see any kind of large scale (non trivial) implementation. We have recently proposed in [2] a model, based on propositional logic and game theory, to capture the general problem of data minimization. We propose to demonstrate the applicability of this approach on real forms.

CCS '23, November 26–30, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0050-7/23/11.
https://doi.org/10.1145/3576915.3624363

General Context. Forms are filled by individuals who wish to apply for welfare benefits (e.g. in France). The volume of personal data thus collected is massive and concerns millions of individuals. For example, hundreds of different types of aids are offered in France at local or national level, to meet different social needs or provide incentives, with a very large number of beneficiaries. This includes aid for family, education, savings, access to housing, health, transport, mobility, energy saving, legal aid, etc. Governmental portals offer simulators so that citizens can identify the aids they can apply for. For example, aide-sociale.fr offers a rather holistic simulator (more than 1000 aids); the governmental portal mesdroitssociaux.gouv.fr is targeted on the most solicited aids (58 welfare aids) and also offers a simulator for young adults (887 aids). In order to be obtained, each proposed aid requires the beneficiary to fill out a form to collect their personal data showing that they meet the eligibility criteria and calibrating the aid. The allocation decision is made based on the collected data, which leads to the processing of millions of forms per year, with several weeks of instruction by the administration. For example, the public data set provided by the French government indicates that 6 million forms per year were processed on average over the period 2016-2020 for only 10 family welfare benefits. As another example, the social aid benefit related to "complementary health coverage" (which is one of those we propose to demonstrate) concerns 7.19 million beneficiaries in 2022 (see annual report of the complementary health insurance on p.10), all of whom had to send in their corresponding form to activate and subsequently renew the aid each year (see same report at bottom of p.8).

Objective. Given these staggering numbers, an effective solution to minimize data collected via forms, while properly informing the individual and obtaining their informed consent to the collection process, may lead to significant resource gains and more rigorous compliance with legal principles [6, 7]. We will show in this demonstration a PET (Privacy Enhancing Technology) which analyzes a given user's data, proposes all possible data minimizations, and quantifies the information that will subsequently be sent by the user, in order to let them make an informed decision on which minimization to choose.

2 DATA MINIMIZATION RELATED WORK

Data minimization prior to processing has long been considered an unsolvable problem [13], with an overly negative influence on the quality of the processing [5], sometimes in contradiction with other

legal obligations such as discrimination prevention [9], fairness [10] or fraud detection [8]. Implementation is deemed too complicated for developers [1, 11], which has led to an overall limited adoption and compliance with this principle. Some barriers were removed by the introduction of a formalism clarifying the general objectives of a data minimizer [3]. Indeed, recent contributions show that such a formalism can be applied in the context of using machine learning to predict recommendations/personalization through performance metrics [4, 12], while minimizing the data taken as input and still producing a useful model as output, even though its quality is slightly reduced. In [2], we study the case of administrative forms, where such a trade-off between data minimization (or privacy) and performance (utility of the processing) does not apply. Indeed the full range of benefits due to individuals *must* be offered by virtue of law: any metric that reduces the legitimate benefits due to a user in exchange for better data minimization would not be acceptable. Moreover, using logic and game theory, we show how to bring meaningful information to the individual in order to obtain their informed consent on the collection process, which is also required by regulations. This demonstration is based on this last result. We explain our approach on an example in the following section, which will be used in the demonstration.

3 A DATA MINIMIZATION EXAMPLE

The RSA benefit. In this section, we walk the reader through our approach, using the "active solidarity income" (in French: "revenu de solidarité active", abbreviated as RSA) welfare benefit form. The volume of RSA requests is around 2 million forms processed per year[1] for a total spending of around 12 billion euros, which is the highest welfare benefit (in term of euros spent) in France. Note that our demonstration will feature the RSA, but also the complementary health coverage aids which is the welfare benefit with the most forms processed annually in France (over 7 million forms) for a total spending of around 2.8 billion euros, as detailed in [2]. We believe these examples show that our approach can be used to manage real large scale use cases, and that there is a real impact on 1) the minimization of data sent and 2) the quantification of the data sent, thus improving the users information with a view to consent.

General approach. We consider an automatic decision making system, which processes user data, collected as a form of attribute/value pairs (e.g. $age = 25$ or $pregnant = false$). We assume that the automatic decision making system uses logical rules, which we write in disjunctive normal form (DNF) (e.g. a disjunction of conjunctions, each conjunction representing one possibility of obtaining a benefit). This is the case of many classical AI approaches such as decision trees, which benefit from explainability. By considering the values of all a given user's attributes, we deduce the predicates triggered by the user.

Building the RSA form. The RSA example was built using data from its official description[2]. 17 predicates are used, which are a direct translation of the official texts, negations included :

p_1 : nationality = FRENCH
p_2 : nationality = EUROPEAN COUNTRY
p_3 : nationality = SWISS
p_4 : valid residence permit for at least 5 years = TRUE
p_5 : age ≥ 25
p_6 : pregnant = TRUE
p_7 : children > 0
p_8 : $18 \leq$ age < 25
p_9 : working = TRUE
p_{10} : living in France = TRUE
p_{11} : single parent = TRUE
p_{12} : High school student = FALSE
p_{13} : Higher education student = FALSE
p_{14} : Intern = FALSE
p_{15} : on leave = FALSE
p_{16} : sabbatical = FALSE
p_{17} : non active status = FALSE

As some predicates are linked to each other, we add some consistency rules, noted R_{ADD}, as follow :

$p_1 \rightarrow p_2$ (if you are French you are from Europe)
$p_1 \rightarrow \neg p_4$ (if you are French you do not have a green card)
$p_4 \rightarrow \neg p_1$ (if you have a green card you are not french)
$p_5 \rightarrow \neg p_8$ (if you are over 25, then you are not 18-25)
$p_8 \rightarrow \neg p_5$ (if you are 18-25, then you are not over 25)

Eligibility criteria and benefits for RSA are outlined on the French "Family Allowance Fund". We directly translated the rules to a propositional formula, and then converted the formula to DNF using the Logical Expression Converter program (Haskell code available here). We then use a parser to transform this expression, and combine it with R_{ADD} to produce the DNF format used by our application (which is a DNF version of the DIMACS CNF format[3]). The `RSA.dnf` file is available here. It is composed of the 17 predicates and a disjunction of 31 conjunctions of up to 12 different predicates. An example rule to obtain the benefit (e.g. the first of the 31 conjunctions) states that one needs to be French (p_1), be over 25 years old (p_5), live in France (p_{10}), be a single parent (p_{11}), and be neither conducting an internship (p_{14}), a sabbatical (p_{16}) or non-active (p_{17}). The final rule is obtained by adding the predicates triggered in R_{ADD}, thus the final rule is : $p_1 \wedge p_2 \wedge \neg p_4 \wedge p_5 \wedge \neg p_8 \wedge p_{10} \wedge p_{11} \wedge p_{14} \wedge p_{16} \wedge p_{17}$.

Computing the Minimal Accurate Subvaluations. We precompute (offline) the Minimal Accurate Subvaluations (MAS) for all possible forms that can obtain the benefit. MAS are a subset of the 17 predicates that correspond to enough information to grant the benefit to a user, even if the user had multiple ways of obtaining the benefit. These subvaluations are minimal in the sense that no predicate can be removed while still granting the benefit to the user. In this example, there are 1296 cases to consider, and 24 MAS. One example of a MAS would be noted m_1 = _1001__0_11___111 where a 0 or 1 in a given position means the predicate must be revealed, along with its value true (1) or false (0) and _ (blank) means the predicate is not sent.

Linking MAS to forms. We next build a bipartite graph linking each MAS to all the forms that it is a subvaluation of. It is possible that a MAS is linked to several forms, and it is possible that a form is linked to several MAS. If a form is only linked to one MAS, then

[1] See the Court of Auditors website, and for more details its recent report: "The Active Solidarity Income", Evaluation of a public policy, Jan. 2022 (link to the report).
[2] See here for the official description (in French).
[3] See this link for details on the DIMACS CNF format.

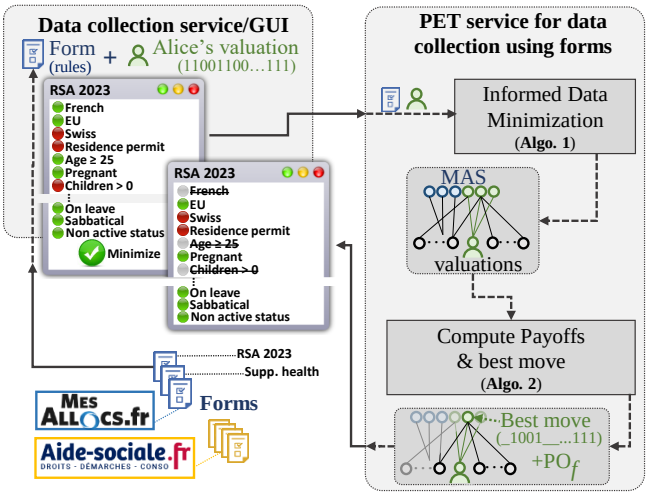

Figure 1: Informed Consent Data Minimization Architecture

a user with this form has no choice when minimizing their form, they will need to send the MAS. If a form is linked to several MAS, then the user may choose which MAS to send to the administration.

Choosing the MAS using game theory. In order to choose which MAS to send, we need to define a cost function, for instance the number of forms sending the MAS. We note this payoff function $PO_{SM}(m)$ (Same MAS) where m is a MAS. We use game theory to compute the correct value of PO_{SM}. The game is to decide which MAS each player (a form) will send to the administration, and the objective of the game is to send the MAS with the highest PO_{SM} value. This is a one round game, all players announcing simultaneously their move. We show in [2], Theorem 4.6 that this game admits a Nash equilibrium (thus the moves of rational players can be accurately predicted and computed), under a hypothesis on lexicographical ordering of the MAS : all players with only one choice play this choice, all players with equivalent choices will play the first MAS in lexicographical order. Note that each player reduces the value of $PO_{SM}(m)$ for all the m that the player could have played, but in fact did not.

Example. Consider a user with the form (as in the demo video) : f = 1100110111000111 (a 0 or 1 in a position means a false or true value for the ith predicate, bear in mind some are negative such as $p_{15} = true$, which means the person is *not* on leave). This user has the choice between 2 different MAS : m_1 = _1001__0_11___111 which corresponds to 32 other different forms or m_2 = _100_1___11___111 which corresponds to 128 other different forms. Note that the number of different forms is less or equal to 2^b where b is the number of blank predicates. Our informed consent minimizer thus informs the user of this possible choice, and lets the user decide, suggesting the better choice of m_2

4 DEMONSTRATION OUTLINE

Our demonstration is divided into 3 parts. The overall architecture is presented in Fig. 1.

(1) Presentation/creation of the decision process. We have already analyzed 2 French administrative welfare benefit procedures (RSA and Welfare benefits). It is also possible for participants to create their own decision process rules.
(2) Processing of the decision process rules (PET service for data collection). This demonstrates the feasibility of our approach on real world examples. Processing of the examples takes up to a few minutes on a regular laptop. This process constructs the MAS then generates the bipartite graph (Algo. 1) and computes the payoff function using game theory (Algo. 2).
(3) Informed consent for users (Data collection service/GUI). Participants are invited to fill in a form. The form is then processed using the output of the previous service. Users are suggested minimized forms, and are informed of the corresponding payoff functions (i.e. number of other players who share the same "blanked" form).

5 CONCLUSION

We introduced a novel Privacy Enhancing Technology for data minimization in administrative forms, informing eligible users about data removal and its impact. This confirms data minimization's practicality in decision rule contexts.

Acknowledgements. *This work was supported by grants ANR JCJC 2019 "PRELAP" (ANR-19-CE48-0006), PEPR "iPOP" (ANR-22-PECY-0002) and ANR France 2030 "CyberINSA" (ANR-23-CMAS-0019).*

REFERENCES

[1] A. Alhazmi and N. A. G. Arachchilage. I'm all ears! listening to software developers on putting gdpr principles into software development practice. *Personal and Ubiquitous Computing*, 25(5):879–892, 2021.
[2] N. Anciaux, S. Frittella, B. Joffroy, B. Nguyen, and G. Scerri. A new PET for data collection via forms with data minimization, full accuracy and informed consent. In *Proc. of the 27th Extending Database Technology Conference*, 2024, to appear.
[3] T. Antignac, D. Sands, and G. Schneider. Data minimisation: a language-based approach. In *IFIP International Conference on ICT Systems Security and Privacy Protection*, pages 442–456. Springer, 2017.
[4] A. J. Biega, P. Potash, H. Daumé, F. Diaz, and M. Finck. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 399–408, 2020.
[5] B. Custers and H. Uršič. Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. *International data privacy law*, 6(1):4–15, 2016.
[6] L. Determann and J. Tam. The california privacy rights act of 2020: A broad and complex data processing regulation that applies to businesses worldwide. *Journal of Data Protection & Privacy*, 4(1):7–21, 2020.
[7] European Council. Regulation EU 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.
[8] L. Găbudeanu, I. Brici, C. Mare, I. C. Mihai, and M. C. Șcheau. Privacy intrusiveness in financial-banking fraud detection. *Risks*, 9(6):104, 2021.
[9] G. Galdon Clavell, M. Martín Zamorano, C. Castillo, O. Smith, and A. Matic. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 265–271, 2020.
[10] Q. Ramadan, D. Strüber, M. Salnitri, J. Jürjens, V. Riediger, and S. Staab. A semi-automated bpmn-based framework for detecting conflicts between security, data-minimization, and fairness requirements. *Software and Systems Modeling*, 19(5):1191–1227, 2020.
[11] A. R. Senarath and N. A. G. Arachchilage. Understanding user privacy expectations: A software developer's perspective. *Telematics Informatics*, 35(7):1845–1862, 2018.
[12] S. Shabanian, D. Shanmugam, F. Diaz, M. Finck, and A. Biega. Learning to limit data collection via scaling laws: Data minimization compliance in practice. arXiv, July 2021.
[13] O. Tene and J. Polonetsky. Privacy in the age of big data: a time for big decisions. *Stan. L. Rev. Online*, 64:63, 2011.