The Observatory of Anonymity: An Interactive Tool to Understand Re-Identification Risks in 89 countries

Luc Rocher
Data Science Institute
Imperial College London
UK
lrocher@imperial.ac.uk

Meenatchi Sundaram Muthu
Imperial College London
UK
msm518@ic.ac.uk

Yves-Alexandre de Montjoye
Data Science Institute
Imperial College London
UK
demontjoye@imperial.ac.uk

ABSTRACT

Anonymity offers strong guarantees for people to use technology without fear of mass surveillance, identity theft, and data misuse. To anonymize and share data widely, de-identification is the main tool used in academia and industry. Yet mounting evidence suggest that de-identification may not protect people's privacy in practice. We present *The Observatory of Anonymity*, an interactive website demonstrating how few pieces of personal data can easily re-identify us. Taking advantage of modern web technologies, it allows participants to explore their correctness score—the likelihood to be correctly and uniquely identified from their demographics only. Trained on census data from 89 countries, it demonstrates the effectiveness of re-identification attacks on deemed-anonymous data. The website further allows analysts to upload their own data samples to train our machine learning models in real time. The Observatory provides a unique tool for individuals, researchers, and practitioners to assess whether current de-identification practices satisfy the anonymization standards of modern data protection laws such as GDPR and CCPA.

CCS CONCEPTS

• **Security and privacy** → **Pseudonymity, anonymity and untraceability**; *Social aspects of security and privacy*; • **Human-centered computing** → *Information visualization.*

KEYWORDS

Data privacy, Visualization, Public Demonstration

ACM Reference Format:
Luc Rocher, Meenatchi Sundaram Muthu, and Yves-Alexandre de Montjoye. 2021. The Observatory of Anonymity: An Interactive Tool to Understand Re-Identification Risks in 89 countries. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3442442.3458606

1 INTRODUCTION

The collection of our personal data online raises legitimate privacy concerns. From financial and medical services to filling in online forms and surveys, our digital traces are increasingly monitored and tracked. While these online traces are useful for modern data-driven research, driving scientific progress in healthcare, social science, and AI [6, 8, 13, 20], more than 72% of U.S. citizens report being worried about sharing their information online [12].

In response to these privacy concerns, anonymity provides strong guarantees for people to use technology without fear of mass surveillance, identity theft, and data misuse. To allow data to be used and shared widely, de-identification is widespread in academia and industry, processing personal data so that it can no longer be attributed to specific individuals before sharing it [3, 10, 15, 17]. The effectiveness of de-identification is still debated [1, 4, 11, 14, 16, 19]; mounting evidence now suggest that few pieces of collected information can often identify us in deemed anonymous data [2, 7, 9, 18].

Building upon our recent method to estimate the likelihood of any re-identification to be successful [18], we propose an interactive website demonstrating how few pieces of demographic information can easily breach our anonymity. Using census data from 89 countries, the Observatory of Anonymity, available at https://cpg.doc.ic.ac.uk/observatory, allows users to explore their correctness—the likelihood to be correctly re-identified when found in any dataset.

The Observatory is engineered with privacy in mind, training our statistical model in the client-side browser. The browser trains a Gaussian Copula model, fitted on individual-level sample data, to predict the correctness of a re-identification in the complete population of a country. Taking advantage of modern web technologies such as WebAssembly [5] and emscripten [21], the Observatory demonstrates re-identification risks using pre-trained demographic data. It further allows researchers and practitioners to estimate if data they collected would be anonymous or not, by training our model on uploaded sample records.

2 SYSTEM ARCHITECTURE

The Observatory of Anonymity consists of four main components: a frontend application, a computational backend, a data processing pipeline and, lastly, a collection of pre-trained country-level data (see Fig. 2). This tool is engineered to run entirely in the client-side browser, guaranteeing that no personal data is sent back to the server. The entire demo is written in ReactJS and TypeScript with computationally heavy routines compiled to WebAssembly using emscripten.

Frontend component. The users interact with the Observatory using the frontend web application, which allows them to switch between multiple views and explore two main facets of re-identification. An interactive quiz (see Fig. 1) first produces a personalized estimation of the risk of re-identification based on the collection of

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3458606

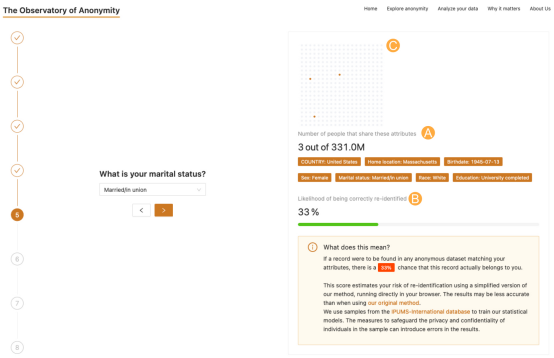

Figure 1: Qualitative and quantitative elements to display the individual correctness

country-level data. This individual correctness is presented to the user through both qualitative and quantitative elements. Users are presented with the number of people (Fig. 1a) who match their demographics (estimated using our Gaussian Copula model) and the resulting correctness score (Fig. 1b), from 0% to 100%. Additionally, they are also presented with a dot map uniqueness visualization (Fig. 1c), which forms a visual representation of how many individuals match the user's demographics. As more information is selected, dots disappear until the user becomes unique. To represent from 1 to 1.4 Billion individuals [largest country, China], we use a custom scale designed to be exact at small counts (≤ 50) and logarithmic at larger counts. Usability testing on a panel of participants showed that this custom linear-logarithmic scale conveys the re-identification risks far better than a linear or pure logarithmic scale.

Lastly, the analyst module allows users to test the anonymity of their own collected data. Analysts first upload a CSV file containing discrete multivariate individual-level data. In response, the module generates an interactive form to visualize the risk of re-identification of individuals in the complete population. The underlying model learns how the collected data are distributed and predicts, from the uploaded sample, who can be anonymous in a much larger population. For computational reasons, we currently limit the sampling factor to 100x: from a sample of 10k records, the model will predict the correctness amongst at most 1M people.

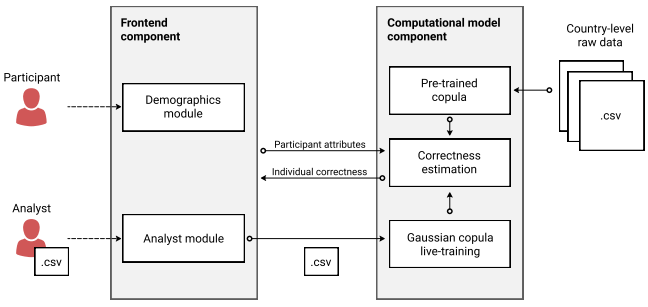

Figure 2: Interactions and data flow in the Observatory.

Data processing. The data processing pipeline takes discrete and categorical individual-level CSV files, with one row per individual. The processing includes data normalization to remove outliers (e.g., missing values), discretization of string values, running the computational model, and exporting the model parameters (marginals, correlation matrix) in a JSON format. This component is run before release on 89 country-level files (pre-training), but also runs in the browser when an analyst uploads a CSV file (live-training). The in-browser live training is fast even on large data files. On a modern, mid-tier laptop, live-training the model on the ADULT dataset (from the UCI Machine Learning Repository) takes 2.58s for 32, 562 records and 11 attributes. Even on older devices and mid-tier mobile devices, the training takes less than 10s.

Computational model. This backend model has two main purposes. It first extracts empirical marginal distributions (Ψ) and a correlation matrix (Σ) from a given dataset and uses them to fit a statistical model for multivariate distributions, a Gaussian Copula. For any individual not necessarily in the training sample, the model then computes the correctness of re-identification given the parameters (Σ, Ψ) and the individual's attributes (x) using multivariate integration:

$$\kappa_x = \frac{1-(1-q(x))^n}{nq(x)}$$

with n the population size and $q(x)$ the expected probability to draw a record matching the individual's attributes in the complete population:

$$q(x) = \int_{F_1^{-1}(x_1-1|\Psi)}^{F_1^{-1}(x_1|\Psi)} \dots \int_{F_d^{-1}(x_d-1|\Psi)}^{F_d^{-1}(x_d|\Psi)} c_\Sigma(u)\mathrm{d}u$$

We refer the reader to our recent article [18] for an in-depth description of the statistical method. To estimate the correctness, the algorithm ran on the browser executes 50 iterations of the above multivariate integral, each with a random marginal configuration. Pseudo-random seeds are utilized to ensure that the computations are reproducible across browsers and input data.

Originally designed for a server with 48 Intel Xeon cores, the statistical model was implemented using a combination of Fortran and Julia code. The Observatory includes a version of our source code ported to TypeScript, overcoming two main challenges: computing speed and numerical stability. We compile the routines computing multivariate normal integrals from Fortran, using the DragonEgg compiler[1] (Fortran to WebAssembly). We then use the emscripten toolchain [21] to interface the resulting WebAssembly code with the backend. Performance-wise, each integration takes 0.36 ms/it using the original Julia code and 1.28 ms/it using WebAssembly on a mid-tier laptop. User testing shows that the interactivity is not affected by this small performance loss.

3 DEMONSTRATION

The Observatory of Anonymity is instantiated with demographic data from 89 countries, obtained from the Integrated Public Use Microdata International (IPUMS-International) database. IPUMS compiles data from national statistical services worldwide. The Belgium data files originate from the 2011 Belgium census, collected by the statistical office Statbel. For this demonstration, we selected

[1] DragonEgg – Using LLVM as a GCC backend, https://dragonegg.llvm.org/.

ten attributes to model the correctness: country and region, age, sex, marital status, number of children, religion, race, education, and employment status. These attributes were chosen due to the availability of attribute data across various countries and the ability to easily infer this data from auxiliary sources.

The Observatory is a free and open-source project licensed under GPLv3. Its source code[2] is documented for easy deployment and to add new data sources. Designed for a mobile- and computer-friendly responsive user experience, it runs from tablets in museum installations to large interactive displays such as Imperial's Data Observatory[3]. Taking advantage of recent performance optimizations in client-side JS engines, the Observatory of Anonymity demonstrates how collected information can easily identify us. It questions whether current de-identification practices satisfy the anonymization standards of modern data protection laws such as GDPR and CCPA. We believe such tools are essential to understand the risk of re-identification when sharing personal data.

ACKNOWLEDGEMENTS

We acknowledge support from the Information Commissioner Office for the development of an initial version of this online demonstration tool.

REFERENCES

[1] Daniel Barth-Jones. 2012. The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. https://ssrn.com/abstract=2076397

[2] Chris Culnane, Benjamin I P Rubinstein, and Vanessa Teague. 2017. Health Data in an Open World. arXiv:cs.CY/1712.05627 http://arxiv.org/abs/1712.05627

[3] Yves-Alexandre de Montjoye, Ali Farzanehfar, Julien M Hendrickx, and Luc Rocher. 2017. Solving Artificial Intelligence's Privacy Problem. *Field Actions Science Reports* Special Issue 17 (Dec. 2017), 80–83. http://journals.openedition.org/factsreports/4494

[4] K El Emam and L Arbuckle. 2014. De-identification: A critical debate. https://fpf.org/2014/07/24/de-identification-a-critical-debate/. https://fpf.org/2014/07/24/de-identification-a-critical-debate/

[5] Andreas Haas, Andreas Rossberg, Derek L Schuff, Ben L Titzer, Michael Holman, Dan Gohman, Luke Wagner, Alon Zakai, and J F Bastien. 2017. Bringing the web up to speed with WebAssembly. In *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI 2017)*. Association for Computing Machinery, New York, NY, USA, 185–200. https://doi.org/10.1145/3062341.3062363

[6] A Halevy, P Norvig, and F Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* 24, 2 (2009), 8–12. https://doi.org/10.1109/MIS.2009.36

[7] Alex Hern. 2017. 'Anonymous' browsing data can be easily exposed, researchers reveal. http://www.theguardian.com/technology/2017/aug/01/data-browsing-habits-brokers

[8] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and Others. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (2009), 721. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2745217/

[9] Grigorios Loukides, Joshua C Denny, and Bradley Malin. 2010. The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc.* 17, 3 (May 2010), 322–327. http://dx.doi.org/10.1136/jamia.2009.002725

[10] Bradley Malin, Kathleen Benitez, and Daniel Masys. 2011. Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. *J. Am. Med. Inform. Assoc.* 18, 1 (Jan. 2011), 3–10. http://dx.doi.org/10.1136/jamia.2010.004622

[11] Gregory J Matthews and Ofer Harel. 2011. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Stat. Surv.* 5, 0 (2011), 1–29. http://projecteuclid.org/euclid.ssu/1296828958

[12] Timothy Morey, Theodore Forbath, and Allison Schoop. 2015. Customer data: Designing for transparency and trust. *Harv. Bus. Rev.* 93, 5 (May 2015), 96–105. https://hbr.org/2015/05/customer-data-designing-for-transparency-and-trust

[13] Travis B Murdoch and Allan S Detsky. 2013. The Inevitable Application of Big Data to Health Care. *JAMA* 309, 13 (April 2013), 1351–1352. http://dx.doi.org/10.1001/jama.2013.393

[14] Arvind Narayanan and Edward W Felten. 2014. No silver bullet: De-identification still doesn't work. http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf.

[15] Office for Civil Rights, HHS. 2002. *Standards for privacy of individually identifiable health information.* Federal Register. https://www.ncbi.nlm.nih.gov/pubmed/12180470

[16] P Ohm. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57 (2010), 1701. http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1450006

[17] Jules Polonetsky, Omer Tene, and Kelsey Finch. 2016. Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification. *Santa Clara Law Rev.* 56, 3 (June 2016), 593–629. http://digitalcommons.law.scu.edu/lawreview/vol56/iss3/3

[18] Luc Rocher, Julien M Hendrickx, and Yves-Alexandre de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* 10, 3069 (July 2019). http://dx.doi.org/10.1038/s41467-019-10933-3

[19] David Sánchez, Sergio Martinez, and Josep Domingo-Ferrer. 2016. Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata". *Science* 351, 6279 (March 2016), 1274. http://dx.doi.org/10.1126/science.aad9295

[20] Rosemary Wyber, Samuel Vaillancourt, William Perry, Priya Mannava, Temitope Folaranmi, and Leo Anthony Celi. 2015. Big data in global health: improving health in low- and middle-income countries. *Bull. World Health Organ.* 93, 3 (March 2015), 203–208. http://dx.doi.org/10.2471/BLT.14.139022

[21] Alon Zakai. 2011. Emscripten: an LLVM-to-JavaScript compiler. In *Proceedings of the ACM international conference companion on Object oriented programming systems languages and applications companion (OOPSLA '11)*. Association for Computing Machinery, New York, NY, USA, 301–312. https://doi.org/10.1145/2048147.2048224

[2]The source code of the Observatory of Anonymity is available at: https://github.com/computationalprivacy/observatory.

[3]Data Observatory, https://www.imperial.ac.uk/data-science/data-observatory.