Empowering Innovation: Unlocking the Potential of Privacy-Enhancing Technologies

Univ.-Prof. Dr. Dominique Schröder October 15, 2024







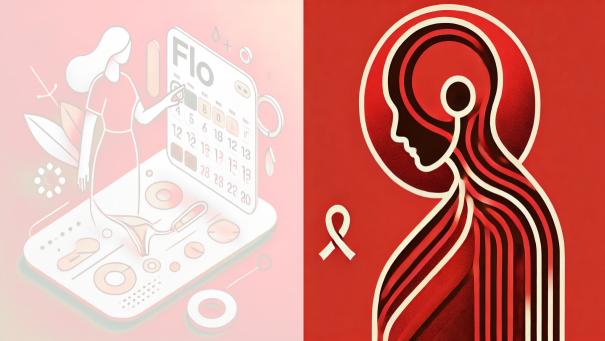




.









The Flo App





The Flo App

- 380 million downloads
- · 68 million monthly active users
- ISO 270001 certification and refers to this certification as "the internationally recognized standard for information security"
- · Collects information such as (in privacy mode!):
 - year of birth
 - place of residence
 - (... gender...)







Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	:	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice		19	
	1234	27	M
Charly	4854	45	F
	1277	28	M
Eve	9584	68	F



Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	:	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice	1161	19	F
Bob	1234	27	М
Charly	4854	45	F
Dave	1277	28	М
Eve	9584	68	F
:	:	: I	:



Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	:	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice	1161	19	F
Bob	1234	27	М
Charly	4854	45	F
Dave	1277	28	М
Eve	9584	68	F
:	:	i I	:



Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	:	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice	1161	19	F
Bob	1234	27	М
Charly	4854	45	F
Dave	1277	28	М
Eve	9584	68	F
:	:	i I	:



Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	÷	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice	1161	19	F
Bob	1234	27	М
Charly	4854	45	F
Dave	1277	28	М
Eve	9584	68	F
:	:	: 1	:



Anonymized dataset containing confidenital information

Zip	Age	Sex	Confidential
15XX	70-75	F	
12XX	25-30	М	
95XX	65-70	F	
11XX	15-20	М	
12XX	45-50	F	
:	:	:	:

Unanonimized dataset containing no confidential information

Identity	Zip	Age	Sex
Alice	1161	19	F
Bob	1234	27	М
Charly	4854	45	F
Dave	1277	28	М
Eve	9584	68	F
:	:	: 1	:





Q The Dataset



Case Studies



Open Question

How can we access the unanonymized dataset?



- · We gather publicly available statistical data.

$$A_{\mathcal{P}}(\vec{a} \mid \vec{b}) = \psi(\vec{a}) \cdot \Pr\left[\vec{b} \mid \vec{a}\right].$$



- · We gather publicly available statistical data.
- Using population statistics, we estimate the anonymity set size $\psi(\vec{a})$.
- · We refine the set size by each auxiliary information we have
- · We define the conditional anonymity set for attributes \vec{a} and \vec{b} via

$$A_{\mathcal{P}}(\vec{a} \mid \vec{b}) = \psi(\vec{a}) \cdot \Pr[\vec{b} \mid \vec{a}].$$



- · We gather publicly available statistical data.
- Using population statistics, we estimate the anonymity set size $\psi(\vec{a})$.
- · We refine the set size by each auxiliary information we have.
- · We define the conditional anonymity set for attributes \vec{a} and \vec{b} via

$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b}) = \psi(\vec{a}) \cdot \Pr\left[\vec{b} \mid \vec{a}\right].$$

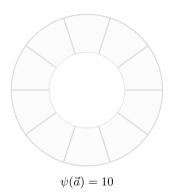


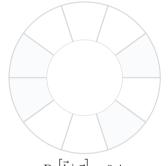
- · We gather publicly available statistical data.
- Using population statistics, we estimate the anonymity set size $\psi(\vec{a})$.
- · We refine the set size by each auxiliary information we have.
- \cdot We define the conditional anonymity set for attributes $ec{a}$ and $ec{b}$ via

$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b}) = \psi(\vec{a}) \cdot \Pr\Big[\vec{b} \mid \vec{a}\Big].$$



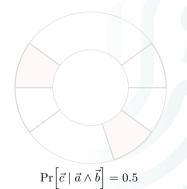
$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr\left[\vec{b} \mid \vec{a}\right] \cdot \Pr\left[\vec{c} \mid \vec{a} \wedge \vec{b}\right]$$



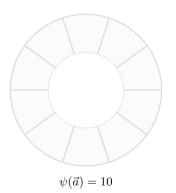


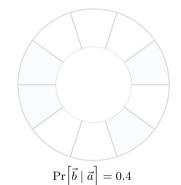


$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$

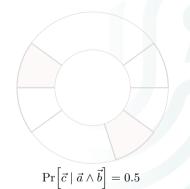


$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$

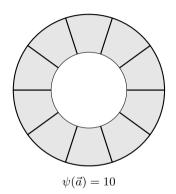




$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$



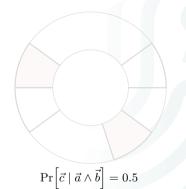
$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$



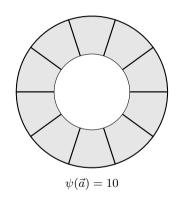


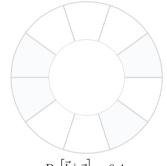


$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$



$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$







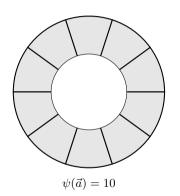
$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$

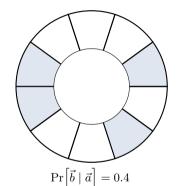


 $\Pr\!\left[\vec{c}\mid\vec{a}\wedge\vec{b}\right]=0.5$



$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$





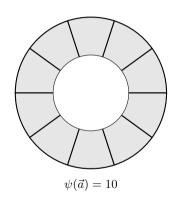


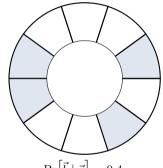
$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$



 $\Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big] = 0.5$

$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \left[\vec{b} \mid \vec{a} \right] \cdot \Pr \left[\vec{c} \mid \vec{a} \wedge \vec{b} \right]$$







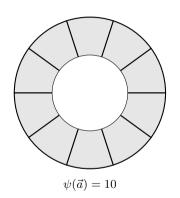
$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$

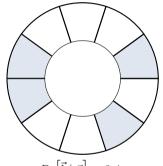


 $\Pr\Big[\vec{c} \mid \vec{a} \wedge \vec{b}\Big] = 0.5$



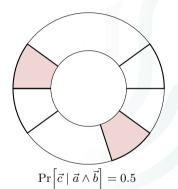
$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$



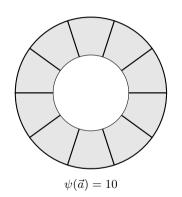


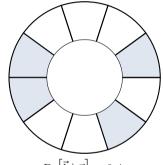


$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$



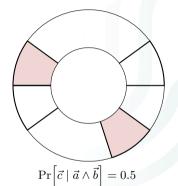
$$\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = \psi(\vec{a}) \cdot \Pr \Big[\vec{b} \mid \vec{a} \Big] \cdot \Pr \Big[\vec{c} \mid \vec{a} \wedge \vec{b} \Big]$$







$$A_{\mathcal{P}}(\vec{a} \mid \vec{b} \mid \vec{c}) = 10 \cdot 0.4 \cdot 0.5 = 2$$





Q The Dataset



Case Studies



Data Request



Paul Gerhart

22 November 2021 at 11:F

Consensus-data request for researching purposes

Bcc: census.customerservices@ons.gov.uk, Eurostat Helpdesk_EN, User Information Services Stats SA, leosanni@nigerianstat.gov.ng,

de

STATCAN infostats-infostats.STATCAN@canada.ca, Atencion a Usuarios, lbge@lbge.gov.br, info, Stat, info@stats.gov.cn, ddu.rgi@nic.in, pbs@pbs.gov.bk, client.services@abs.gov.au, info@stats.govt.nr, nstac-info@nstac.go.jp, statistics@un.org, statistics@un.org, statistics@un.org, statistics@un.org

To whom it may concern,

My name is Paul Gerhart, and I am part of a privacy research team at the chair of applied cryptography of the Friedrich Alexander-University Erlangen-Nurnberg in cooperation with the University of Birmingham.

My team and I are working on a web app to inform people about the anonymity set they are currently living in. That is the number of people who fit in the same data bucket created by several data points one may provide voluntarily without worries. With our app, we want to create awareness of how sensitive personal data is to help people protect their privacy.

Our work is based on the paper Pandemic Privacy by Berrang and Schröder, but we want to stress the insights to a worldwide dataset.

Therefore, we are interested in census data that gives insights into the population count by postcode separated by age groups and sex. Moreover, we are interested in the distribution of height and weight by age group and sex.

Based on this data alone, we cannot deanonymize people, but we can show how anonymity decreases by the publication of personal data that might seem irrelevant.

Hence we were hoping you could provide the desired data for us.

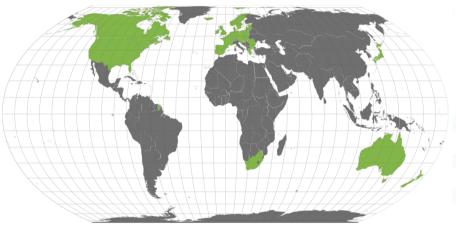
Best regards

Paul Gerhart



Our Dataset

Currently, we can calculate anonymity sets for 1 084 230 346 people.



Data Response I



☐ VisualAnon 8. December 2021 at 13:33













Data Response II



минэкономразвития россии ФЕДЕРАЛЬНАЯ СЛУЖБА

ФЕДЕРАЛЬНАЯ СЛУЖБА ГОСУДАРСТВЕННОЙ СТАТИСТИКИ (РОССТАТ)

Герхарт П.
paul.gerhart@fau.de

Мженишан ум., в. 39, стр. 1, г. Москва, 107450 », Тел.: (495) 607-49-02, факс: (495) 607-22-06 http://www.gks.ru; e-mail: statigigks.ru OG. IA. MAN ye. CAICOT

Уважаемый госполин Герхарт!

В связи с Вашим обращением направляем имеющуюся официальную статистическую информацию о распространенности роста и всеа в разбивке по возрастным группам (в возрасте 15 лет и более) и полу. Даниые предоставлены по итотам Выборочного наблюдения состояния здоровыя населения 2020 года, материалы и база микроданных которого размещены на официальном сайте Росстата (https://rosstat.gov.ru/): Статистика/ Переписы и обследования/ Федеральные статистические изблюдения по социально-демографическим проблемам/ Итоти выборочного наблюдения состояния здоровыя населения.



Q The Dataset



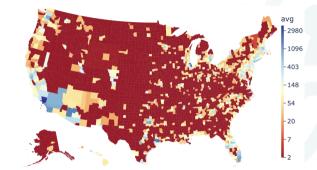
Case Studies

Q Visual Anon



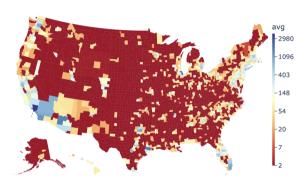
Case Study: USA

- · Avg. CAS: 77
- · Avg. CAS in red area: 2
- · Avg. CAS below 5 in 97% of the counties





Roe v. Wade: Flo App



Status of Abortion Bans in the United States as of October 7, 2024

Hover over state for more details

Abortion Banned (12 states)

■ Gestational limit between 6 and 12 weeks LMP (6 states)

Gestational limit between 15 and 22 weeks LMP (5 states)

Gestational limit at or near viability (17 states)

No gestational limits (9 states & DC)



Note. Use refers to Last Mentatual Period. Violably is the point when a fetus can survive outdisk the worth and in generally presumed to occur at around 24 weeks gestation. However, violability. It has never been properly defined by courts and depends on the individual pregnancy and on various factors, including gestational age, fetal weight and see, and medical interventions a variable. For more details please see our trackers on exceptions to state abortion have and confirmation of the proposed present and the individual pregnancy and our XFT state.

For more details please see our trackers on exceptions to state abortion bans and early gestational limits, abortion-related ballot initiatives, state and federal Itigation, and our RFF Stat Health Facts page on abortion policies.

Source: RFF analysis of state policies and court decisions, as of October 7, 2024. • Get the data - Embed • Download PNG

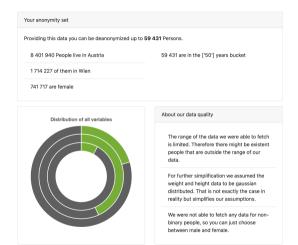




KFF

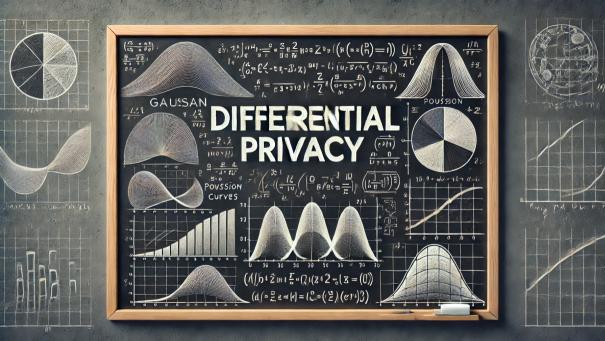
Visual Anon (Age)











Example: Grade Release in Schools

Grade	Count	
1	2	
2	4	
3	6	
4	3	
5	1	
Mean	2.8125	



Example: Grade Release in Schools

Grade	Count	Revealed
1	2	2
2	4	4
3	6	6
4	3	3
5	1	0
Mean	2.8125	2.5

- There are 16 students in class.
- · Teacher publishes mean grade: 2.8125
- Students learn the grade of each student except for one
- The mean of the publishing students is 2.5 (assigning a 0 to the unpublishing student)
- · They compute

$$(2.8125 - 2.5) * 16 = 5$$

and leak the unpublished grade

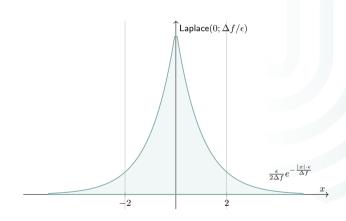




Example: Histogram Queries

$$\mathsf{cnt}'(x) = \mathsf{cnt}(x) + \mathsf{Laplace}(0, 1/\epsilon)$$

x	cnt(x)	$\epsilon = 2$
1	2	1.96
2	4	3.46
3	6	6.08
4	3	3.16
5	1	1.62
$\mathbb{E}(X)$	2.8125	2.9398







Example: Grade Release in Schools

Grade	Count	Revealed
1	1.96	2
2	3.46	4
3	6.08	6
4	3.16	3
5	1.62	0
Mean	2.9398	2.5

Computing the missing grade:

$$(2.9398 - 2.5) \cdot 16 = 7.03$$

Impact of Histogram Queries

Industry	Medicine	
Customer Behavior	Patient Health Data	
(Amazon, Walmart)	(Mayo Clinic, Cleveland Clinic)	
- Analyzes purchases	- Summarizes patient data	
- E.g., purchases per month	- E.g., age distribution of patients	
Log Analysis	Drug Effectiveness	
(AWS, Azure)	(Pfizer, Novartis)	
- Monitors system logs	- Analyzes treatment responses	
- E.g., server response times	- E.g., drug dosage effectiveness	
Financial Risk	Epidemiology	
(JP Morgan, Goldman Sachs)	(CDC, WHO)	
- Categorizes risk levels	- Tracks infection rates	
- E.g., asset risk distribution	- E.g., COVID-19 spread	
Supply Chain	Medical Imaging	
(FedEx, Toyota)	(Radiology, MRI)	
- Tracks delivery times	- Analyzes image intensity	
- E.g., shipment times	- E.g., MRI scan analysis	







- The dataset of a single hospital is too sparse
- Idea: Combine the datasets of multiple hospitals
- Problem: The data cannot leave the hospital
- **Solution:** We design an MPC protocol and apply differential privacy











- The dataset of a single hospital is too sparse
- Idea: Combine the datasets of multiple hospitals
- Problem: The data cannot leave the hospital
- Solution: We design an MPC protocol and apply differential privacy

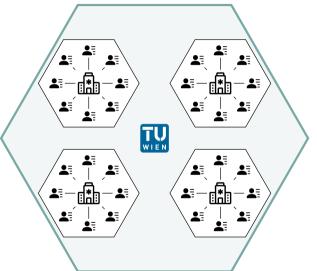




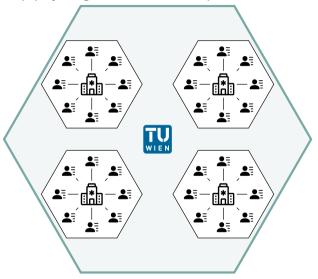




- The dataset of a single hospital is too sparse
- Idea: Combine the datasets of multiple hospitals
- Problem: The data cannot leave the hospital
- **Solution:** We design an MPC protocol and apply differential privacy



- The dataset of a single hospital is too sparse
- Idea: Combine the datasets of multiple hospitals
- Problem: The data cannot leave the hospital
- Solution: We design an MPC protocol and apply differential privacy



- The dataset of a single hospital is too sparse
- Idea: Combine the datasets of multiple hospitals
- Problem: The data cannot leave the hospital
- Solution: We design an MPC protocol and apply differential privacy

PRIVACY GUARANTEE

No patient data ever leaves any hospital





How PETS Work Under the Hood

PreRound(pk)	Lagrange(S,i)	$SignAgg(pk, \rho, \{\sigma_i\}_{i \in S}, m)$
1: $X \leftarrow pk$	1: $\Lambda_i \leftarrow \prod_{j \in S \setminus \{i\}} j/(j-i)$	1: $X \leftarrow pk$
$2: d_i, \leftarrow \mathbb{Z}_p; e_i, \leftarrow \mathbb{Z}_p$	$_2$: return Λ_i	$2: (D, E) \leftarrow \rho$
$3: D_i \leftarrow g^{d_i}; E_i \leftarrow g^{e_i}$		$3: b \leftarrow H_{\mathrm{non}}(X, S, \rho, m)$
$4: state_i \leftarrow (d_i, e_i)$	$SignRound(sk_i, pk, S, state_i, \rho, m)$	$4: R \leftarrow DE^b$
$5: \rho_i \leftarrow (D_i, E_i)$	1 : $/\!\!/$ called at most once per secret state $state_i$	$5: s' \leftarrow \sum_{i \in S} \sigma_i$
6: return $(state_i, \rho_i)$	$2: \ \overline{x}_i \leftarrow sk_i; \ X \leftarrow pk$	$6: \sigma \leftarrow (R,s)$
	$3: (D, E) \leftarrow \rho$	7: return σ
$PreAgg(pk, \{\rho_i\}_{i \in S})$	$4: (d_i, e_i) \leftarrow state_i$	
$1: X \leftarrow pk$	5: $b \leftarrow H_{\mathrm{non}}(X, S, \rho, m)$	$Verify(pk, m, \sigma)$
$2: \{(D_i, E_i)\}_{i \in S} \leftarrow \{\rho_i\}_{i \in S}$	$6: R \leftarrow DE^b$	1: $X \leftarrow pk$
$3: D \leftarrow \prod_{i \in S} D_i$	7: $c \leftarrow H_{\mathrm{sig}}(X, R, m)$	$2: (R,s) \leftarrow \sigma$
$4: E \leftarrow \prod_{i \in S} E_i$	$8: \Lambda_i \leftarrow Lagrange(S,i)$	$3: c \leftarrow H_{\mathrm{sig}}(X, R, m)$
$5: \rho \leftarrow (D, E)$	9: $\sigma_i \leftarrow d_i + be_i + c\Lambda_i \overline{x}_i$	4: return $(g^s = RX^c)$
6: return $ ho$	10: $\mathbf{return} \ \sigma_i$	



Practical Schnorr Threshold Signatures without the Algebraic Group Model Hien Chu, Paul Gerhart, Tim Ruffing & Dominique Schröder CRYPTO'23





Current Research

Work Published by E192-08









PRIVACY ENHANCING TECHNOLOGIES