Measuring Conditional Anonymity—A Global Study

Pascal Berrang
University of Birmingham
United Kingdom

Paul Gerhart
Friedrich-Alexander Universität
Erlangen-Nürnberg
Germany

Dominique Schröder*
TU Wien, Austria
Friedrich-Alexander Universität
Erlangen-Nürnberg, Germany

ABSTRACT

The realm of digital health is experiencing a global surge, with mobile applications extending their reach into various facets of daily life. From tracking daily eating habits and vital functions to monitoring sleep patterns and even the menstrual cycle, these apps have become ubiquitous in their pursuit of comprehensive health insights. Many of these apps collect sensitive data and promise users to protect their privacy – often through pseudonymization. We analyze the real anonymity that users can expect by this approach and report on our findings. More concretely:

(1) We introduce the notion of *conditional* anonymity sets derived from statistical properties of the population.
(2) We measure anonymity sets for two real-world applications and present overarching findings from 39 countries.
(3) We develop a graphical tool for people to explore their own anonymity set.

One of our case studies is a popular app for tracking the menstruation cycle. Our findings for this app show that, despite their promise to protect privacy, the collected data can be used to identify users up to groups of 5 people in 97% of all the US counties, allowing the de-anonymization of the individuals. Given that the US Supreme Court recently overturned abortion rights, the possibility of determining individuals is a calamity.

1 INTRODUCTION

The global landscape of digital health is undergoing an unprecedented surge, fueled by legislative changes facilitating broader access to health data for research and the exponential growth of healthcare applications. According to a study by Fortune Business Inside, the digital health application market is projected to skyrocket from $38.89 billion USD in 2021 to an astonishing $314.60 billion USD by 2028[1]. This surge, however, raises serious concerns about the protection of sensitive user information, as these applications promise privacy safeguards while simultaneously accumulating vast datasets.

Despite these assurances, the frequent occurrence of data breaches affecting massive user bases cannot be ignored [20, 30]. The aftermath of such breaches, estimated by IBM Security to cost an average of $10.10 million in the healthcare sector alone[2], underscores the urgent need for innovative privacy solutions.

It is noteworthy that many of these applications employ pseudonymization techniques in an attempt to protect user privacy. Pseudonymization involves replacing personally identifiable information with pseudonyms, rendering the data more challenging but not impossible to directly attribute to individual users. An illustrative example of such a recent data leakage is the case of the GetHealth platform, where 60 million personal data records were disclosed [29].

Our research ventures into the critical realm of privacy protection for applications handling vast sociodemographic and medical datasets. In light of the escalating data breaches and the prevalence of pseudonymization, we seek to address fundamental questions:

- *How can we determine the individual anonymity level for users of such apps?*
- *What is the typical anonymity level faced by users?*

We address both questions. We introduce a novel and simple mechanism to compute anonymity sets without having access to the apps' original data sets; we refer to this technique as *conditional anonymity sets*. These sets are derived from publicly available statistical sources. We provide the tool *VisualAnon*[3] to explore the individual anonymity set for a large amount of countries.

We measure the effectiveness of our approach through the examination of two representative examples of such apps: the Flo app and a (medical) data donation app. These case studies serve as illustrative instances to evaluate the practical implications of our methodology. Additionally, we present overarching findings derived from a broader analysis, encompassing several countries and providing a comprehensive understanding of the generalizability and impact of our approach in the realm of digital health data privacy.

1.1 Roe v. Wade—Privacy Matters

In the US, there are 11 states in which abortion has already been made illegal, almost immediately followed by law enforcement trying to access relevant sensitive information from various apps. Facebook, for example, had to hand over private chat messages of a 17-year-old girl living in Nebraska to the police. This allowed law enforcement to charge the girl and her mother in an abortion case, as the Guardian reported [11].

The Flo app is one of the most popular apps for tracking the period. The app was downloaded over 300 million times and used by roughly 50 million people every month. Flo advertises that the app has ISO 270001 certification and refers to this certification as

*The research in this work was partially conducted at Friedrich-Alexander Universtiät Erlangen-Nürnberg, Germany.

[1] https://www.medicaldevice-network.com/news/digital-health-apps/

[2] https://www.ibm.com/reports/data-breach

[3] https://visualanon.org. For the reviewing process, we made VisualAnon anonymous. The username to access VisualAnon is "PETS", and the password is "VisualAnon".

Proceedings on Privacy Enhancing Technologies 2024(4), 947–966

https://doi.org/10.56553/popets-2024-0150

"the internationally recognized standard for information security[4]". This certificate suggests to the user that the data is safe and their privacy is protected. A closer look at Flo's privacy statement shows that the app collects name, email address, year of birth, place of residence, and associated location information, including time zone and language. Flo states that they can infer the gender. Furthermore, the user may choose to share information like weight, body temperature, menstrual cycle dates, and further information via, e.g., Apple Health.

In view of the recent decision of the Supreme Court, collecting this information is alarming as it allows the tracking of abortion. Flo tries to solve this problem by introducing an anonymous mode [3] in which the name, email, and technical identifiers get removed.

However, our conditional anonymity sets reveal that this is insufficient. Due to the unequal distribution of the population density, 97% of the counties in the US have an average anonymity set size of less than 5 for women of age 20-60. We shared our insights with the Flo app developers.

1.2 Data Donation App

Another example is the data donation app of the Robert Koch Institute (RKI). The RKI is the leading institute in Germany that focuses on the investigation and prevention of infectious diseases. Moreover, it is also responsible for nationwide health monitoring [8, 9]. Its data donation app collects health information and computes a fever curve to predict further outbreaks and identify COVID hotspots. More than 1 million people currently use the app, and it collected over 400 million data records [2, 38]. The RKI advertises the app as being pseudonymous – it uses generalization techniques to protect the privacy of its users and assigns each user a unique ID to associate new data with that user.

The collected attributes yield anonymity set sizes of less than 20 for only a bit more than 5% of the general population. This number deteriorates quickly when considering simple additional knowledge. For example, if we know that a target is a smartwatch user, then this information yields anonymity sets smaller than 20 for more than 35% of the cases. Furthermore, the knowledge that a target participated in the study of the RKI uniquely identifies 87% of the participants. We shared our results with the RKI in responsible disclosure.

1.3 Our Contribution

Our main contributions are as follows:

- We introduce the notion of conditional anonymity sets to estimate the level of anonymity of individuals. These sets are computed from publicly available statistics and thus do not require access to the original data set.
- We present two real-world case studies, analyzing the conditional anonymity sets on the Flo app in the USA and the RKI's data donation app in Germany.
- We have assembled statistical data from 39 countries all over the world, covering over one billion people, allowing us to globally measure anonymity on the internet using conditional anonymity sets.
- We evaluate the accuracy of our conditional anonymity sets based on a fictional dataset of 102.5 million users.
- We develop an online tool called *VisualAnon* that builds upon the collected data and allows users to estimate their own anonymity level, and we actively develop VisualAnon. VisualAnon is designed to serve a dual purpose: as well as raising awareness, it is also a research tool, helping to reveal how anonymity varies across different demographic landscapes. It supports researchers and privacy advocates by facilitating the assessment of de-identification risks in different datasets. In addition, policymakers and application developers can use VisualAnon to test and improve their privacy protocols.

1.4 Ethics Discussion

All experiments presented in this work solely utilize publicly available data, as disseminated by the census bureaus of respective countries. This data is anonymized and aggregated, and we only perform secondary data analyses on it. Our methodology estimates anonymity set sizes but cannot identify actual individuals. We do not engage with nor process any form of personal or individual-specific data[5]. The project gained ethics approval from the lead institution's review board.

2 RELATED WORK

With a vast amount of data being created and collected, our society has long identified the need for privacy as an existential right. Our research community has since explored and developed ways to protect our privacy in the digital age.

General Identification Attacks. Pioneered with the early work of Latanya Sweeney, such approaches have been used in the past to deanonymize people in the US [33] and in the context of videos [26]. The work of Sweeney showed how to deanonymize concrete persons by combining a medical database and an election registry, and the work of Narayanan matched videos from an anonymized Netflix database to the publicly available IMDB database. In contrast, *conditional anonymity sets* provide a methodology for the computation of meaningful bounds for anonymity sets without access to concrete datasets and solely based on publicly available data.

Data Reconstruction Attacks. Data reconstruction attacks (where attackers aim to complete an incomplete dataset) differ from our method for two main reasons. Firstly, our approach is based on statistical data, unlike these attacks, which require specific partial data sets. Second, while data reconstruction attacks focus on identifying individuals with a certain probability, such as "John Doe can be re-identified with probability x", our research estimates anonymity within a group, such as "a person with these attributes has an anonymity set of 5". Data reconstruction attacks have been studied extensively [25, 28, 32]. Recently, the work by Rocher et al. [31] demonstrates the application of machine learning to these attacks. In addition, [31] presented an online tool to illustrate data reconstruction attacks using machine learning with data from the

[4] https://flo.health/press-center/flo-achieves-isfo-27001-certification

[5] We use the term personal data in accordance with the ethics and data protection guidelines of the European Commission (https://commission.europa.eu/system/files/2020-06/5._h2020_ethics_and_data_protection_0.pdf)

US, England, and Wales. The tool aims to answer a question different from VisualAnon. Furthermore, the tool of Rocher et al. [31] needs to run ML training with specific data for extension to other countries. In contrast, Visual Anon covers 39 countries and operates without the need to train machine learning models.

In a recent work [1], Michael Hawes explores a matching attack focusing on the reconstruction of microdata from the US census. The authors systematically reconstruct individual-level records from published census tables by solving a system of equations using mixed-integer linear programming. In a second step, the authors match a personal data source file with the reconstructed tables. We want to use this work to examine the differences between conditional anonymity sets and reconstruction attacks.

Unlike the specific attack demonstrated in [1], conditional anonymity sets estimate the likelihood of success for a *wide range* of possible deanonymisation attacks *without* requiring actual reconstruction of the dataset. This stems from the fact that conditional anonymity sets estimate the biggest anonymity a user can hope for by only revealing minimal data, regardless of how the actual revealed data is used to break anonymity. In addition, conditional anonymity sets can be computed only using statistical data without relying on heavy computations. This contrasts reconstruction attacks that rely on actual data sets and heavy computation.

In conclusion, the goal of conditional anonymity sets is to facilitate providing bounds and guidelines without directly compromising anonymity, while the goal of reconstruction attacks (thus also the goal of [1]) is to show that people can actually be deanonymized by running an attack on the anonymity.

Health Privacy. The privacy of health data is of paramount concern due to the sensitivity of such data. Due to the large body of work in this area, we can only touch on some of the more recent work.

Furthermore, many applications of differential privacy to health data exist. For example, in the area of pharmacogenetics, Fredrikson et al. show that differential privacy can induce inadequate warfarin dosing and expose patients to increased risk of mortality [19]. For genetic data, genome-wide association studies have been a primary concern, and many papers are studying the application of differential privacy for this use case [23, 36, 37, 39]. In epigenetics, Backes et al. [14] and Berrang et al. [16] study linkage attacks like ours (but on concrete datasets). Backes et al. [14] also provide suitable trade-offs between utility and privacy for a local, differentially private model.

3 MEASURING PRIVACY

The main goal of our work is to understand the impact of data breaches for allegedly anonymized (medical) data. In this section, we explore the impact of a data breach, beginning with the definition and formalization of a threat model. In this model, we assume the attacker possesses two databases $\mathcal{D}_1$ and $\mathcal{D}_2$, out of which only one is anonymized. The adversary's goal is the computation of a matching between both databases.

In many real-world applications and settings, the full amount of data that is collected and anonymized is unknown and not made public to the users. For example, the Flo app supports an anonymous mode where "Flo user [have] the option to access the app without name, email address, and technical identifiers[6]". But how anonymous are the users in this setting? Since precise information is missing, it is impossible to compute exact anonymity sets. Nevertheless, the users should have an estimate of their degree of anonymity by the (additional) information that they provide, such as age in a certain range. We introduce *conditional anonymity sets* (CAS) as meaningful bounds of the anonymity set size independent of a concrete application. We derive conditional anonymity sets from publicly available statistical information only.

3.1 Threat Model

Our approach estimates the potential success of a concrete attacker. While the attacker is assumed to have access to two concrete databases with user data, our estimates do not require actual access to these databases. Instead, we demonstrate how to provide meaningful bounds of the anonymity set sizes based on population statistics only. We also show how these estimates relate to anonymity sets in concrete databases based on the addition of background knowledge.

Databases: We assume that the adversary has access to two concrete databases $\mathcal{D}_1$ and $\mathcal{D}_2$. One database $\mathcal{D}_1$ is *not* pseudonymous and contains socio-demographic information as well as the real identity of potential victims. Such as database can either be obtained (maliciously) by data leakages, or it is a state actor who is trying to de-anonymize some of its citizens. The other database $\mathcal{D}_2$ is anonymized and contains socio-demographic data of individuals and additional attributes such as longitudinal health data. Due to the anonymization process, the information in the second database might be imprecise. As an example, consider the information that the Flo app collects in its anonymous mode, where the age is stored in buckets only.

Adversary: We focus on a PPT adversary $\mathcal{A}$ that may have some *auxiliary information* or background knowledge and which tries to link both the databases to de-anonymize an individual in $\mathcal{D}_2$ and gain additional knowledge about individuals in $\mathcal{D}_1$. As a concrete example, think about the prosecution trying to de-anonymize a woman who has had an abortion. The prosecution has its own database of citizens ($\mathcal{D}_1$) and forces the Flo application to reveal its "anonymized" database ($\mathcal{D}_2$).

We measure the adversary's success of de-anonymizing an individual in $\mathcal{D}_2$ as the probability of them producing a correct link with its entry in $\mathcal{D}_1$. For this purpose, we assume that every individual in $\mathcal{D}_2$ is present in $\mathcal{D}_1$ – slightly abusing notation that is $\mathcal{D}_2 \subseteq \mathcal{D}_1$. Since the adversary's success would be 0 for any individual not being part of $\mathcal{D}_1$, this assumption implicates a strictly stronger adversary. We capture additional relationships between the two databases in the form of *auxiliary information* about individuals in $\mathcal{D}_1$.

Definition 3.1 (Adversary's Success). Let $\mathcal{D}_1, \mathcal{D}_2$ be databases as defined in our adversarial model. Let i be an individual in $\mathcal{D}_2$ and $\mathcal{D}_2(i)$ be the data of this individual within the database. $\sigma(i)$ is the corresponding entry of i in $\mathcal{D}_1$, and by κ we denote auxiliary information the adversary has. We define the adversary's success

[6] https://flo.health/privacy-portal/anonymous-mode-fa

in de-anonymizing an individual i as

$$success(i) = \Pr[\mathcal{A}(\mathcal{D}_2(i), \mathcal{D}_1, \kappa) = \sigma(i)],$$

where $\mathcal{A}$ is defined by the adversarial strategy.

To protect the privacy of a user i, we want to minimize $success(i)$. Note that this adversary has access to concrete databases and does not rely on population statistics. Instead, we will use population statistics to provide reliable estimates on $success(i)$.

The adversarial model we propose covers our case studies but is also applicable more generally. As already mentioned, the databases might originate from different data leakages. Alternatively, one of the databases could have been collected by the adversary themselves (a curious app provider, signup data from an insurance company, or obtained from prosecution, etc.). $\mathcal{D}_2$ might have been published for scientific purposes and could be part of a study.

Relevance of our adversarial model. Cyber attacks in the real world clearly demonstrate that such databases are frequently obtained by hacking [20, 30] or be made available by human mistake [29]. Data miners and brokers are trying to link such databases and sell the resulting information in a "business worth billions" [35]. In one very recent instance, 60 million data records were leaked from the GetHealth platform. Describing itself as a "unified solution to access health and wellness data from hundreds of wearables, medical devices, and apps" [29], their platform achieves the same goal as the framework used by most of the health applications. The leaked data is similar in nature and included names, dates of birth, weight, height, gender, and GPS logs, among others.

Adversarial strategy. When regarding databases containing unperturbed information, we assume an adversary who links an entry from $\mathcal{D}_2$ to $\mathcal{D}_1$ by an exact matching of the overlapping attributes. The adversary limits the choices by applying their auxiliary information and, if multiple matches exist, chooses one of the remaining matches at random. This corresponds to the adversarial strategy maximizing the adversary's success given that both databases contain precise information (that may still be generalized into bins). Using the random choice, whenever the adversary is uncertain about the matching, allows us to assume adversaries that have no background knowledge of the dataset. We want to emphasize that our approach also works with more sophisticated metrics, such as entropy (and min-entropy) (c.f. [18]). Yet, we will see in Section 4 that the random selection is predominantly performed on small sets (e.g., this set size is below 5 for citizens of 97% of all US counties). Therefore, we use random selection, which serves as a trade-off between utility and precision.

3.2 Conditional Anonymity Sets

Given the adversarial model and the exact matching strategy described in Section 3.1, it is easy to see that $success(i)$ is inversely proportional to the number of individuals in $\mathcal{D}_1$ exposing the same attributes as $\mathcal{D}_2(i)$ and satisfying the auxiliary knowledge.

If there is only a single matching entry in $\mathcal{D}_1$ that is in line with the auxiliary information, this has to be the correct link (since $\mathcal{D}_2 \subseteq \mathcal{D}_1$). Hence, the adversary's success is 1. If there are k matching entries in line with the auxiliary information, the adversary randomly chooses one of them. The probability of choosing the correct entry – and thus the adversary's success – is $\frac{1}{k}$. We say k is the size of the anonymity set for the attributes $\mathcal{D}(i)$.

Definition 3.2 (Anonymity Set). Given a vector of attributes $\vec{x}$ and a database $\mathcal{D}$, we define the anonymity set for $\vec{x}$ as:

$$\mathsf{A}_{\mathcal{D}}(\vec{x}) = \{j \mid \mathcal{D}(j) \overset{\cap}{=} \vec{x}\},$$

where $\mathcal{D}(j) \overset{\cap}{=} \vec{x}$ is defined as an exact match of all overlapping attributes between $\mathcal{D}(j)$ and $\vec{x}$.

We did not incorporate the auxiliary knowledge directly into the notion of anonymity sets and will instead use it to filter the database first. Given auxiliary information κ, we usually talk about the anonymity sets $\mathsf{A}_{\mathcal{D}'}(\vec{x})$, where $\mathcal{D}' = \mathcal{D}|_{\kappa,\vec{x}}$ is the subset of the original database, which is in line with the background knowledge.

For example, if the adversary knows $\vec{x}$ is a smartwatch user and the adversary knows which individuals in $\mathcal{D}$ are smartwatch users, they can exclude all others to form $\mathcal{D}'$. Using our previous argumentation, we can reformulate the adversary's success.

Proposition 3.3. *Given an adversary $\mathcal{A}$ as defined in Section 3.1 with auxiliary knowledge κ, an exact matching strategy yields an adversary's success of*

$$success(i) = |\mathsf{A}_{\mathcal{D}'}(\mathcal{D}_2(i))|^{-1},$$

where $\mathcal{D}' = \mathcal{D}_1|_{\kappa,\mathcal{D}_2(i)}$.

To protect the privacy of an individual i, we want to minimize the adversary's success $success(i)$. Thus, the size of the anonymity set $k = \left|\mathsf{A}_{\mathcal{D}_1|_{\kappa,\mathcal{D}_2(i)}}(\mathcal{D}_2(i))\right|$ is a good metric for the privacy of the individual. This metric implies a form of k-anonymity for the set of i's attributes [34]. We can calculate k from population statistics only and do not require actual instantiations of $\mathcal{D}_1$ and $\mathcal{D}_2$. For any possible combination of attributes in $\mathcal{D}_2$, we can estimate the number of individuals in a given population who exhibit these attributes.

This brings us to the definition of *conditional anonymity sets*. Conditional anonymity sets estimate anonymity sets from public population statistics. Given a particular instantiation $\vec{a}$ over attributes α (e.g., *gender = female*), a population statistic $\psi(\vec{a})$ returns the number of people in this population $\mathcal{P}$ exhibiting these attributes. It is also the size of the anonymity set of any such individual with respect to the population: $|\mathsf{A}_{\mathcal{P}}(\vec{a})|$.

Definition 3.4 (Conditional Anonymity Set). Let $\vec{b}$ denote an instantiation over a non-overlapping set of attributes β, such that $\alpha \cap \beta = \emptyset$. We define the *conditional anonymity set* $\mathsf{A}_{\mathcal{P}}(\vec{a} \mid \vec{b}) = \mathsf{A}_{\mathcal{P}|_{\vec{b}}}(\vec{a})$ as the anonymity set capturing the part of the population with both these attributes. Given a conditional probability distribution $\Pr[\beta \mid \alpha]$, we can calculate its size as:

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

Cumulative Distribution Function (CDF). A *Cumulative Distribution Function (CDF)* characterizes the probability that a random variable X assumes a value less than or equal to a given point x, expressed as

$$\mathsf{CDF}(\mathsf{x}) = \Pr[X \leq x].$$

In the subsequent sections, we will assess the CDF function for the distribution of Conditional Anonymity Sets (CAS) sizes. The CDF facilitates the following insights:

- It enables the estimation of how many CAS are below a specific threshold.
- The steepness of the CDF indicates a higher prevalence of CAS for a given size, while a flatter CDF suggests a lower count of CAS for the same size.
- When comparing CDFs for two CAS distributions, the disparities between the CDFs highlight the differences between the distributions. If a CDF is shifted toward the origin, it indicates a relatively larger number of small CAS, while a shift toward the positive axis indicates a prevalence of larger CAS.

3.3 Comparison to Sweeney [33]

In pioneering work, Latanya Sweeney showed that matching two databases, one anonymized and one with personally identifiable information, is relatively easy. She also built an online platform[7] to measure anonymity when *precise data* about a subject, such as their date of birth plus zip code, is leaked.

We generalize Sweeney's basic idea from constructing intersections of *precise* data sets to computing intersections of *distributions* over data sets. In contrast, we take the statistical distribution of people living in an area in combination with the statistical information about the age distribution in that area.

Because of Sweeney's work, only bucketed information is published these days, e.g., year of birth (or even decade of birth) rather than the exact date of birth. However, our work shows this is often insufficient because we can build intersections with the statistical information provided by the governments. In many cases, these resulting anonymity sets are very small.

3.4 Accuracy of Conditional Anonymity Sets

We evaluate the accuracy of conditional anonymity sets through a comprehensive ground truth survey to ensure their effectiveness. We first create a synthetic dataset for AnonLand, a fictional country with a population of 102.5 million. The demographic attributes of the dataset are generated based on specific distributions:

- The age distribution is modeled using a trapezoidal shape: ages from 0 to 40 are uniformly distributed, while ages from 41 to 90 follow a triangular distribution, reflecting a more varied age distribution in this range.
- Gender is assigned by sampling a uniformly random bit with an equal chance of being male or female.
- Height follows a normal distribution, with an average (mean) height of 180 cm for men and 175 cm for women, and a standard deviation of 10 cm for both genders.
- Weight is normally distributed, with a mean of 80 kg for males and 70 kg for females and a standard deviation of 10 kg for both genders.
- In order to accurately represent both densely populated and sparsely populated areas, we divide the districts into five classes. Accordingly, we create five metropolises of five million inhabitants each, 25 cities of one million inhabitants each, 250 counties of 100,000 inhabitants each, 2,500 areas of 10,000 inhabitants each, and 2,500 villages of one thousand inhabitants each.

In the second step, we conducted a census on the people of AnonLand to derive statistics on the district population by gender and age, as well as height and weight statistics per age and gender. We do this census twice: once without applying differential privacy and once applying laplacian noise with an epsilon of $\epsilon = 2$ and a sensitivity of 1 (for histogram queries).

Finally, we randomly selected five thousand random citizens (one thousand per district class) and compared the real anonymity sets of these citizens to the conditional anonymity sets estimated using our methodology. We select a thousand inhabitants per district class to evaluate the accuracy of CAS for both small and big RAS sizes.

3.4.1 The Impact of Differential Privacy. Our first observation is that the size of a conditional anonymity set, which is computed with noised data, differs only marginally from the CAS computed on the plain data. For each of our 5000 test citizens, the noised CAS differed no more than 0.2 from the unnoised CAS.

3.4.2 Accuracy of the CAS. To analyze the accuracy of the CAS, we compare the CAS to the real anonymity set (RAS) by computing the normalized error (CAS − RAS)/RAS, which we henceforth call divergence. We depict the results for the divergence in Fig. 1.

Our first observation is that the CAS-RAS divergence narrows down with increasing RAS size. The second observation is that the absolute divergence exceeds the value 0.25 mainly for small RAS sizes. We observe that the normalized error of the CAS is minimal for RAS sizes above 25.

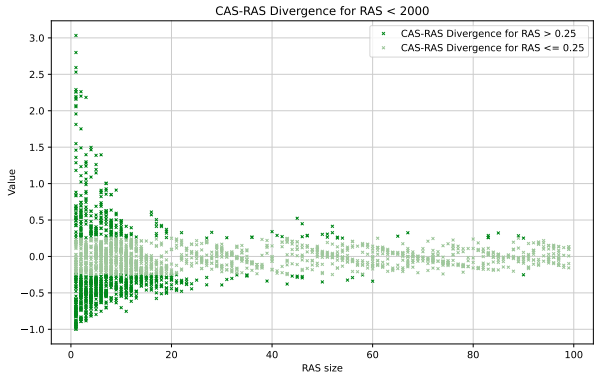


Figure 1: CAS-RAS **divergence below and above** 0.25 **for all participants. For** RAS **sizes below** 100**, primarily isolated outliers above an absolute divergence of** 0.25 **exist.**

Given that the normalized error for RAS values above 25 is minimal, we focus our analysis on the CAS when the RAS size is less than 25. Figure Fig. 2 presents a boxplot illustrating the CAS scatter for each RAS value. Additionally, we depict the ideal relation where CAS = RAS. We observe that the median of all CAS values is

[7] https://aboutmyinfo.org

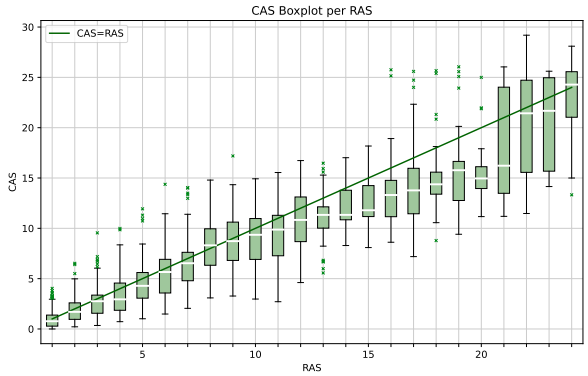


Figure 2: Boxplot of the CAS **per** RAS **for** RAS **sizes below** 25**. The** CAS **is a narrow collar of the** RAS**.**

mostly below the actual RAS size. Furthermore, the majority of the boxes have narrow interquartile ranges around or below the ideal line, indicating that CAS sizes are close to their corresponding RAS. For a more detailed analysis of the accuracy of CAS, we provide additional plots in Appendix B.

4 CASE STUDIES

There is a plethora of health applications on the market that all collect similar data. This section shows that storing simple and only slightly pseudonymized data in practical applications does not provide sufficient privacy protection. The novelty here is that we use our notion of conditional anonymity sets to estimate the level of anonymity one can hope for, given the information provided to the apps.

We exemplarily analyze two such apps in a case study: the Flo and the RKI's data donation app. The Flo app tries to predict the user's menstruation cycle by analyzing previous cycles, body temperature, and other symptoms. The primary purpose of the RKI app is to collect health data to predict and estimate the spread of the coronavirus and to improve the early detection of hotspots by calculating a fever curve.

4.1 Ovulation Apps in the US

Ovulation apps are widely recognized for their usefulness in helping individuals understand and track their menstrual cycles. However, the recent decision to outlaw abortion adds a new dimension to the use of these apps. While they continue to serve their primary purpose, the evolving legal landscape adds an unexpected layer of significance, transforming them into potential security threats. In this section, we assess the emerging threat to ovulation app users in the US This assessment aims to shed light on the implications and risks associated with using such apps, considering the broader context of legal changes and their impact on reproductive health decisions. We showcase this threat with the example of the Flo app, one of the most widely used ovulation apps.

4.1.1 Flo App. The Flo app was downloaded over 300 million times and used by roughly 50 million people monthly[8]. It provides multiple statistics over the user's ovulation circle, like an ovulation calculator, a period calculator, a pregnancy calculator, and a pregnancy due date calculator. As the Flo app is proprietary, no detailed information about the data usage is publicly known. However, according to the privacy policy [4], weight, body temperature, and menstrual cycle dates are amongst the stored data. Furthermore, the app allows the integration of data provided by external services like Apple HealthKit or GoogleFit.

4.1.2 Dataset. Our primary data source is the United States Census Bureau [13], which is responsible for conducting the official census and providing comprehensive demographic statistics. Specifically, we rely on the dataset derived from the American Community Survey, referred to as table *S0101*. This dataset contains tuples representing the attributes of the US population, including county, sex, age, and count. While the American Community Survey dataset provides the granularity necessary for our analysis, it includes information on minors and lacks age group distinctions beyond 75. As a result, our analysis focuses on age groups from 18 to 75. The population counts in this dataset reflect the year 2021.

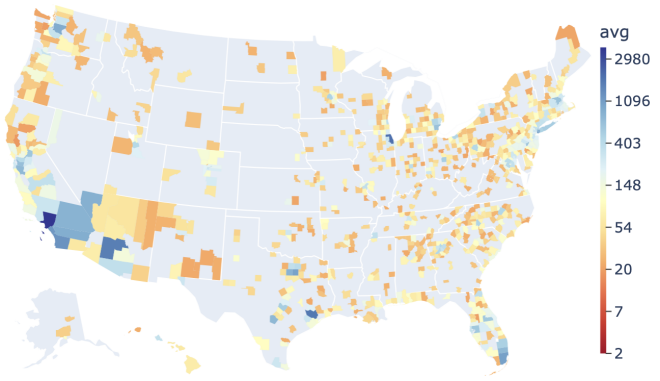
To enrich our analysis, we include data on the height and weight of the US population from a separate source [27]. The Centers for Disease Control and Prevention (CDC) dataset from a 2002 survey provides means and standard errors for weight and height across gender and age groups. In particular, we assume that height and weight are independent of the county of residence (but not of the country). Our assumption is based on the normal distribution of both height and weight around their respective means. While height generally follows a normal distribution, weight exhibits a slight right skew [22]. Nevertheless, for practical purposes, a normal distribution serves as a reasonable approximation of body weight [21]. To maintain physiological coherence, we restrict the body mass index to the range of 17 to 30, effectively filtering out implausible combinations of height and weight.

The American Community Survey provides data for only 840 of the 3221 official counties. For the remaining counties, we calculate averages by dividing the remaining state population by the number of uncounted counties in that state, multiplying the result by the number of age groups, and finally multiplying this by the mean values of the height and weight distributions. As a result of the calculation of these averages, the American Community Survey already has coverage of the vast majority of the population in the 840 counties. Consequently, for the counties missing specific data, we fill each conditional anonymity set with these calculated averages. The difference between the unfilled and filled datasets is visualized in Figure 3.

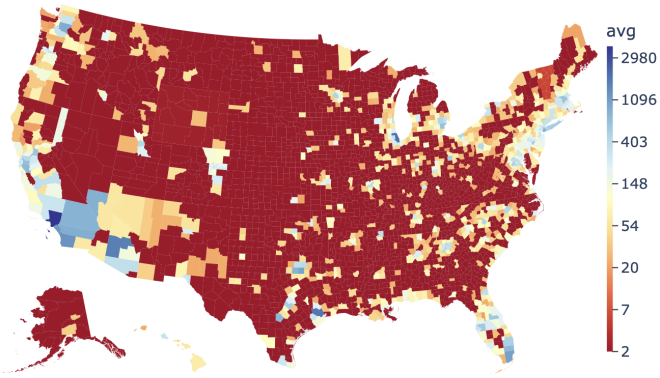
4.1.3 Findings. We evaluate the conditional anonymity sets of U.S. citizens in three different scenarios. In the first case, we examine the entire U.S. population without additional information. Then, in the second case, we narrowed our focus to fertile women, estimated by restricting the age range from 20 to 60. This particular case includes individuals directly affected by Roe v. Wade, as described in Section 1.1. Finally, our third case explores the effects of additional

[8] https://flo.health/about-flo

knowledge, specifically the knowledge that a woman is using the Flo app. This third scenario simulates the potential risks associated with a data breach within the Flo app. To upper-bound the probability that a woman in the US is using the Flo app, we use the factor of 50/169.03. We establish this factor since there are about 169.03 million women in the US [12], and the Flo app has about 50 million active users. While this only constitutes an upper bound (since not every Flo user is a woman in the US), it still provides interesting implications to the conditional anonymity set, as we show in the following paragraphs.



(a) Average set sizes, without filling uncovered counties.



(b) Average set sizes, uncovered counties are filled with the average value.

Figure 3: Average conditional anonymity set sizes for the female population between 20 and 60 years in the USA.

No Auxiliary Information. The maximum anonymity set size is offered by Los Angeles County with a size of 11,026 for males between 25 and 30 years of age, between 175cm and 180cm in height (5.7–5.9ft), and a weight of 80–85kg (176–187.3lbs). The other four districts that are not in Los Angeles County but provide the biggest anonymity set size are Cook County, Illinois (5,535); Harris County, Texas (4,914); Maricopa County, Arizona (4,475); and San Diego County, California (3,815), all for 25-year-old males, with a height between 175cm and 180cm (5.7–5.9ft) and a weight between 80kg and 85kg (176–187.3lbs). The smallest anonymity set size can be found in Ada County, Idaho, for males of 20 years with a height of 150cm (4.92ft) and a weight of 40kg (88.18lbs). A person with these attributes has a BMI of 17.8 and hence is covered by our evaluation. The overall average anonymity set size for the 840 counted counties (out of 3,221 official counties) is 77, but as we show in Figure 3, the average anonymity set size in each uncounted county is around 2. Looking at the CDF for the US data (c.f. Figure 13), we observe that 20% of the US citizens have a CAS size of less than 70, and 80% of the US citizens have a CAS size of below 900. The biggest difference between two anonymity sets in a single district can be found in Dallas County, where the anonymity set for 25–30-year-old males is 3.3 times larger than the one for 70–75-year-olds. The average difference between the smallest and largest set per district is 2.34 times.

Another surprising observation is that San Francisco County offers only relatively low anonymity. The largest anonymity set in this county is of size 54,292 (30–35-year-old male inhabitants), which is a mere 35% of the *smallest* anonymity set in Los Angeles County. San Francisco County also exhibits surprisingly small anonymity set sizes for 20–25-year-old inhabitants. For young people above the age of 25, the anonymity set sizes increase significantly (c.f. Figure 16 in the Appendix).

Auxiliary Information: Females between 20 and 60 years. The maximum anonymity set size for females between 20 and 60 years is in Los Angeles County, California, with a size of 9,089 for females aged between 25 and 30 years, a height of 160–165cm (5.24–5.41ft), and a weight of 70–75kg (154,32–165,34lbs). The smallest CAS size contains a single person in Ada County, Idaho, for 20 to 25-year-old females with a height of 185–190cm (6.06–6.23lbs) and a weight of 100–105kg (220.46–231.48lbs). The average anonymity set size is 70, and looking at the CDF function (c.f. Figure 13), the CAS sizes of females between 20 and 60 years are nearly identically distributed as the CAS sizes of the overall US population without background assumptions.

Considering the ten most populous counties in the US, young women (25–30) usually constitute the biggest anonymity sets, which is positive in light of our case study. In Miami Dade, however, this is not the case; instead, women between 55 and 60 make up the largest CAS, exposing young women to a greater risk of deanonymization. In fact, in Miami Dade and King County, the smallest anonymity set consists of 20–25-year-old women, and the largest CAS in King County is 57% larger than the set of this vulnerable subgroup. Considering a woman in Miami Dade aged 20–25, 180–185cm of height and 80–85kg of weight, her anonymity set is only of size 96.

Flo app users (upper bound). To find an upper bound for Flo app users, we assume that at most 29.5% of all American females use the Flo app. This upper bound is computed by dividing the number of 50 million Flo app users by the number of 169.03 million American females as already discussed in Section 4.1.3. As we apply this filter to each female in the US, the minimal and maximal conditional anonymity sets remain the same but with a size of one-third of the original CAS size. This shift is also visible in the CDF function (c.f. Figure 13), which is equally shifted to the left, indicating the same distribution but for smaller anonymity set sizes.

4.1.4 Impact. Coming back to the Flo app, our analysis yields that, even in "anonymous mode", the average CAS has a mere size of 20 (and this assumes an upper bound on Flo app users). Even in the scenario where an adversary cannot distinguish between Flo app users and non-users, 20% of potential users fall into a CAS of size 70 or lower. Moreover, there exist especially vulnerable groups in 97% of all US counties, for which the CAS size is smaller or equal to 5. We determine sparsely populated counties as the main factor impacting anonymity set sizes but note that this is not the only risk factor. Given our surprising results for Miami Dade County, we caution that anonymity sets can be very location dependent. We call for better privacy protections for users of such apps. We list potential mitigation techniques in Section 5.

4.2 Data Donation Apps in Germany

In the midst of the COVID-19 pandemic, several apps emerged to monitor the prevalence and transmission of the virus. Given the noble cause these apps championed and the need for collective efforts to contain the pandemic, individuals were more generous in donating their personal data than they might have been in other circumstances. In addition, a significant number of these apps actively advocated for and implemented strong privacy standards. In our examination of the privacy threats posed by these tracking apps, we focus on the example of the data donation app developed by the Robert Koch Institute (RKI). The RKI, a respected German government agency, and research institute dedicated to disease control and prevention, serves as a notable case study for understanding the implications and challenges associated with the use of such tracking applications in general.

4.2.1 RKI Data Donation App. The RKI Data Donation App has gained significant traction, with over one million users in Germany and a colossal collection of over 400 million data records [2, 38]. Positioned as a pseudonymous platform, the app uses generalization techniques to protect user privacy, assigning users a unique ID for associating new data.

During registration, the RKI app collects socio-demographic data. It then uses wearables to collect longitudinal health metrics, including activity and pulse, steps, calories burned, distance traveled, stairs climbed, sleep patterns, and body temperature over time. To strengthen privacy measures, the RKI is adopting a strategy of data generalization prior to transmission. Instead of collecting precise values, the data is stored with a certain level of granularity. Sociodemographic categories include location (district), gender (male/female), age (in 5-year increments), height (in 5cm increments), and weight (in 5kg increments). For example, a person who is 23 years old would be categorized as 20-25 years old. Notably, all participant information is linked to a pseudonymous user ID, providing an additional layer of privacy protection.

4.2.2 Dataset. Our data originates from the German Federal Statistical Office [5], the agency responsible for conducting the official census and providing comprehensive demographic statistics. The census dataset, identified as table *12411-0018*, consists of tuples denoting the attributes of the German population, including district, gender, age, and number. Similar to the approach taken in the US case study, our analysis focuses on age groups between 18 and 75, with population counts from December 2020.

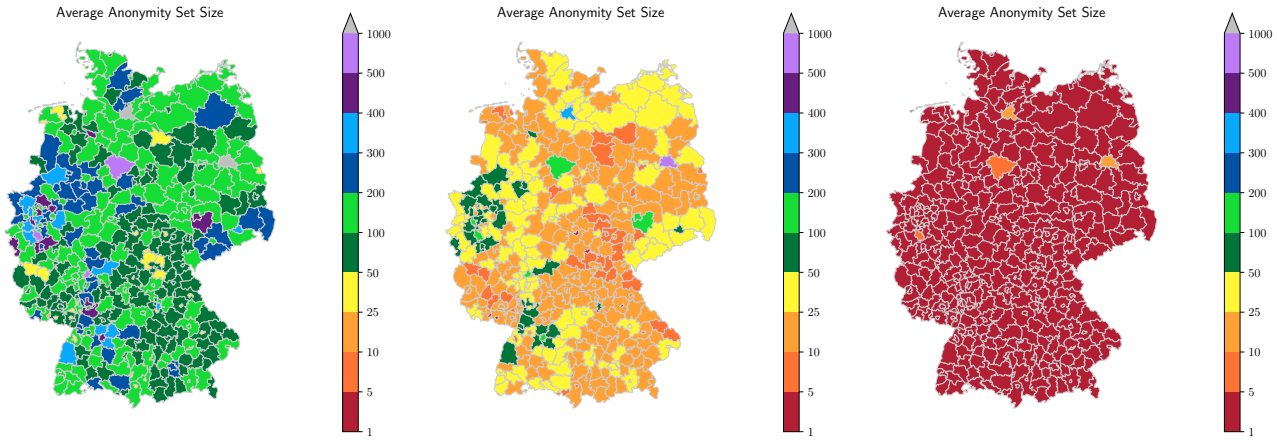
To enhance our insights, we augment this dataset with information on the height and weight of the German population from table *12211-9018*. The German dataset provides tuples indicating mean weight, mean height, age, and sex from a 2017 health survey. Supplementary statistics on the distribution of body weight and height from this survey, provided by the German Federal Statistical Office [5] and detailed in Appendix C, contribute to our analysis. We assume that height and weight are independent of district of residence (but not of country), and we note that the differences between West and East Germany in 1999, as analyzed by Bergmann and Mensink [15], were generally small.

Like in the case study for the US, we assume both height and weight are normally distributed around the mean and restrict the body mass index to a range between 17 to 30 to filter for impossible combinations of height and weight.

The German Federal Statistical Office's publication on using smartwatches by age groups [10] becomes another dimension of our analysis. We assume independence from district and gender, although deviations may exist in practice. Nevertheless, this assumption allows us to make approximate estimates and to illustrate the diminishing anonymity in the presence of additional adversary background knowledge. In addition, we assume that the demographic distribution of app participants is consistent with census data. In the context of the Covid data donation app, the Robert Koch Institute (RKI) conducted a detailed analysis and concluded that the age distribution of their data matches the total population, with minor deviations. They also observe that female donors are slightly younger than the national average, while male donors are slightly older, resulting in a slightly higher proportion of women in their dataset compared to national statistics [6].

The RKI Data Donation App is specifically designed for Germany, so our analysis focuses solely on this country. We perform a detailed examination of app-specific types of adversarial background knowledge in three progressive steps. In the first step, we evaluate the scenario without any additional background information. This forms the baseline for our analysis. In the second step, we consider individuals using a smartwatch. Since the RKI app incorporates longitudinal health data from wearables, assuming smartwatch users provide a first and approximate estimate of potential RKI app users. In the final step, we examine the scenario where a user has not only adopted a smartwatch but has also installed the RKI app. This represents a more specific and refined level of background knowledge that accounts for the user's active engagement with the RKI app. We give an overview of the average anonymity set sizes with the respective additional background information in Figure 4 in the Appendix.

No Auxiliary Information. Berlin and Hamburg offer the highest protection on average, followed by cities such as Hannover, Leipzig, and Stuttgart (c.f. Figure 4a and Figure 20). Interestingly, some large cities, such as Munich, only offer lower anonymity set sizes despite having a higher population density than Berlin. Regarding the maximum anonymity set sizes, all districts have at least one combination of attributes possessed by at least 100 individuals.



(a) Average set sizes, no background knowledge. **(b) Smartwatch users.** **(c) RKI participants.**

Figure 4: Average anonymity set sizes by district given different auxiliary information in Germany.

The largest anonymity set overall is of size 7,903 and contains females between 65 and 70 years old and 160—165cm tall, weighing 70—75kg, living in Berlin.

Figure 14 shows the cumulative distribution function (CDF) for the anonymity set sizes. The blue line shows the CDF, considering no auxiliary information. We can see that more than 10% of the population has an anonymity set of size less than 40. About 30% of the population have an anonymity set greater or equal to 300.

Auxiliary Information: Smartwatch Users. Adding seemingly trivial background knowledge – such as which individuals are users of a smartwatch – can already result in a detrimental loss of privacy. We compare the distribution of average anonymity set sizes without auxiliary knowledge with the one assuming knowledge of smartwatch users. With the exception of Berlin, none of the areas can offer an average anonymity set size of more than 500 anymore. In fact, most of the districts now average below 50 (354 of 401 districts).

We observe a similar decay of privacy for the maximum set sizes (c.f. Figure 21). Only Berlin, Hamburg, Hannover, Frankfurt, and Cologne still have anonymity sets beyond the size of 400.

Due to the low number of smartwatch users in the age group 65—70, the previously largest anonymity set of female pensioners has shrunk from 7,903 to a count of 570. The new top spot is now allocated to the group of female citizens of Berlin aged 30—35 and being of height 165—170cm and weight 65—70kg. This group has an anonymity set size of 1,889.

Figure 14 indicates that this auxiliary information corresponds to a shift of the CDF of about one order of magnitude. More than 30% of smartwatch users have an anonymity set of size less than 20. Only about 15% belong to a set with more than 100 members.

Auxiliary Information: App Participants. The RKI's blog provides detailed information on the regional distribution of its donors [7], providing us with the necessary probability distribution to calculate the corresponding CAS.

Figure 4c strikingly demonstrates the disastrous consequences of an adversary having detailed membership information about the app's participants. All districts except Berlin, Hamburg, Hannover, and Cologne offer average anonymity sets of fewer than 5 members. And even those exceptions do not go beyond a size of 25 on average.

The maximum anonymity set size overall can be found in Berlin and has the same combination of attributes as in the first scenario. This time, the set has a size of a mere 57 members. Most other districts can only provide a maximum anonymity set size of less than 5 (c.f. Figure 22 in the Appendix).

The CDF (Figure 14) shows that more than 87% of the app's participants are uniquely identifiable and can be de-anonymized by an adversary with $success(i) = 1$.

4.2.3 Impact. Our findings for the data donation app are even more concerning than those for the Flo app. An adversary with knowledge about the app participants can uniquely identify 87% of the users. Even if we relax the assumptions on the attacker and only assume knowledge of seemingly harmless information such as the *smartwatch user* background information, anonymity set sizes tend to be small. We can again identify vulnerable subgroups that are especially prone to de-anonymization. One of these groups, for example, is the set of female pensioners with smartwatches. Moreover, location also plays a crucial role. Inhabitants of sparsely populated districts tend to be more vulnerable than inhabitants of big cities. Overall, our findings affirm the need for better privacy protections, and we refer to Section 5 for potential mitigations.

4.3 Visual Anon

The case studies we have provided so far focus primarily on Germany and the United States. However, recognizing the diversity of conditional anonymity sets across multiple countries, we aim to broaden awareness and make the impact of these sets tangible to a wider audience. To achieve this goal, we present *VisualAnon* (https://visualanon.org), an online tool designed to assess conditional anonymity sets worldwide. For the reviewing process, we

made VisualAnon anonymous. The username to access VisualAnon is "PETS", and the password is "VisualAnon". Currently, we have collected statistical data for 39 countries, encompassing a population of over one billion users who can estimate their highest achievable level of privacy. In this section, we illustrate the process of building VisualAnon and demonstrate how this online tool can effectively estimate conditional anonymity on a global scale. Through VisualAnon, we aim to provide a comprehensive understanding of the nuanced variations in conditional anonymity across countries, thus contributing to a more comprehensive and globally relevant perspective on privacy implications.

4.3.1 Goals. The primary goal of VisualAnon is to expand the awareness of conditional anonymity beyond our initial case studies in Germany and the US. We aim to provide an experiential understanding of how the knowledge of only a few attributes can already lead to the deanonymization of individuals into small and identifiable groups. Through VisualAnon, users can interactively explore the dynamic variations in conditional anonymity set sizes considering attributes such as age, gender, height, and weight across different districts. These attributes, which are common in health applications, serve as a basic starting point for our exploration. As we move forward, we intend to expand the scope by incorporating additional attributes to enrich the analysis. By continually improving VisualAnon, we aim to provide a more comprehensive and nuanced perspective on the challenges and implications of conditional anonymity and understanding in this critical area.

4.3.2 Dataset. We built the dataset for VisualAnon based on our existing data from the US and Germany. To expand its scope, we collected additional data from various sources, including Eurostat for the European Union and Switzerland, Statistics South Africa via email, NZ.Stat for New Zealand, the Statistics Bureau of Japan, the Australian Bureau of Statistics for Australia, Statistics Canada, and the Census and Statistics Department for Hong Kong.

The collection of height and weight data varies from country to country, and currently, we have extracted this information only for the US, Germany, and Japan. In cases where height and weight data is unavailable for a specific country, VisualAnon defaults to the German dataset. We are actively seeking and incorporating height and weight data for additional countries to improve the tool's global applicability and accuracy. We plan to make our preprocessed data publicly available on the VisualAnon website to facilitate further research on measuring anonymity. At the time of writing, VisualAnon covers 1,084,230,346 people from 39 countries, with at least one country per continent, which is roughly 13.98% of the world population. We defer an example of the use of VisualAnon to Appendix A.

4.4 Global Measurements

Using our rich dataset, we aim to explore the global variation in CAS across countries. This investigation focuses exclusively on CAS determined by the attributes of district, gender, and age, as these three factors alone significantly narrow the scope of CAS. In this section, we present noteworthy findings from our extensive dataset and offer perspectives that we find particularly interesting. Specifically, we address the following questions:

(1) *What are the countries with the smallest* CAS *size?*
(2) *Is there an observable difference in* CAS *size between males and females?*
(3) *To what extent is it possible for a citizen to enhance their conditional anonymity?*

We address each question in its own section, followed by a more detailed examination of countries with notable findings. We expect, for this section, that the average CAS size for Germany and the US are higher than in Section 4.1 and Section 4.2, since we now don't consider the height and weight attributes anymore.

4.4.1 The Smallest CAS *Size.* We begin our analysis by asking: which of our covered countries have the smallest non-zero CAS, given the attributes of county, age, and gender? This question is answered in Table 1, which shows all countries with a minimum CAS per gender between 1 and 100. Australia emerges as the country with the smallest CAS size. In Acton, there is only one male (and one female) in a given age group. It is also noteworthy that the US ranks 5th with a CAS size of 82 males in Bastrop County, Texas, within a specific age group. This observation is particularly remarkable considering that the American Community Survey provides data for only 840 of the 3,221 official counties (c.f. Section 4.1).

Table 1: Minimal CAS **size** ($\neq 0$) **for countries with a minimal** CAS **size of at most** 100**.**

Country	District	Minimal Set
Australia	Acton	1 male
	Alps - East	1 female
NZ	Chatham Islands Territory	3 males
	Chatham Islands Territory	3 females
Canada	Churchill–Keewatinook Aski, Man	5 males
	Northwest Territories, NWT	10 females
Japan	Nakagusuku-son	19 males
	Nakagusuku-son	34 females
US	Bastrop County, Texas	82 males
	Riley County, Kansas	165 females

4.4.2 Gender-Specific CAS *Differences.* In the next question, we address gender-specific differences in the CAS. Specifically, we want to explore possible differences between the average CAS size for males and females per country. To answer this question, we calculate the average CAS size per country and gender, along with the quotient between the average male and female CAS sizes. This quotient allows us to assess gender-specific differences.

The most interesting results of our evaluation are depicted in Table 2. The top half of the Table 2 shows countries where the male/female ratio is less than 0.9, indicating that the average CAS size for males is more than 10% smaller than for females. Conversely, the bottom half of Table 2 contains countries where the male/female ratio is above 1, indicating that the average CAS size favors females.

Notably, Hong Kong stands out as the only country where the average CAS for males exceeds that of females by more than 10%,

with a remarkable quotient of 3.95. Since Hong Kong is so outstanding, we further explore the anonymity implications of people living in Hong Kong in Section 4.4.4.

Table 2: Average anonymity set size by district, gender, and age. We depict countries from our dataset where the quota of male/female is below 0.9 and above 1.

Country	CAS (Male)	CAS (Female)	Male/Female
Latvia	9,398.46	11,349.89	0.83
Lithuania	8,276.40	9,859.30	0.84
Estonia	7,045.06	8,264.81	0.85
Hungary	14,041.77	15,690.20	0.89
Sweden	13,020.80	12,983.48	1.00
Iceland	4,381.61	4,357.64	1.01
Norway	7,465.53	7,350.08	1.02
Hong Kong	10,551.02	2,672.81	3.95

4.4.3 CAS *Variance.* To this point, we have estimated the minimal and average CAS sizes of each country in our dataset. However, our analysis of the US in Section 4.1 has already shown that there are countries where the actual CAS size deviates heavily from the average CAS size for a country (e.g., Figure 3, where we have isolated bright zones with a high average CAS, and most red zones with a low average CAS). Therefore, we ask how starkly the CAS sizes deviate from the average CAS size in our collected countries.

To answer this question, we computed for each country and gender how many conditional anonymity sets deviate more than 30% from the average CAS size. We then calculated the percentage of these deviating sets over all possible CAS. If the percentage is low, then most of the attribute combinations lead to a similar CAS size, and switching the attribute set by, e.g., moving from one district to another has no big effects on the own CAS. If the percentage is high, then with a high probability, changing the own attribute set can influence the CAS size. In Table 3, we show the countries with a deviation of less than 50% from their average CAS size (top half) and countries with a deviation of more than 80% (bottom half).

Using this first result, we now want to focus further on the six countries with more than 80% deviating attribute combinations. In particular, we want to examine how the average CAS size changes when we remove the outmost quartiles of the CAS sizes. For these countries, we presume that the majority of people live in isolated districts, and hence, the average CAS size of all other districts should be much lower. We examine this question by considering the average CAS size, the average CAS size without the upper quartile of all CAS sizes, and the average CAS size without the upper and lower quartile (of all CAS sizes). Table 4 shows the results of this examination, and we can see that all six countries have much lower CAS sizes when we remove the quartile of the upper CAS sizes. In addition, Table 4 also shows the percental decrease of the average CAS size for each country and gender. The lower this percentage is, the smaller the remaining average CAS after removing the quartiles, meaning that there are many small and less big CAS sizes. We observe a notable case of the percental decrease in Japan, where the remaining average CAS size for males is only 27% of the original average CAS size. In contrast, all other countries show a value of at least 39%. Because of this outlier, we take a closer look at Japan in Section 4.4.4 and Section 4.4.5.

Table 3: Countries, and the percentage of CAS that deviate more than 30% from the mean CAS size. We show countries with less than 50% (the CAS is similar in the whole country) and more than 80% (the CAS tends to be different).

Country	Male Dev.	Female Dev.
Cyprus	0.357143	0.428571
Slovakia	0.446429	0.348214
Canada	0.509026	0.499704
Latvia	0.476190	0.523810
Poland	0.544372	0.465368
New Zealand	0.797575	0.804104
Greece	0.798319	0.824930
Estonia	0.842857	0.742857
United States of America	0.858333	0.854836
Japan	0.873281	0.871206
Sweden	0.863946	0.880952

Table 4: Changes of the average CAS under different conditions for countries with a high deviation from the mean CAS (c.f. Table 3). The first column describes the mean CAS without conditions, the second column the mean CAS if the highest quartile is removed from the dataset, and the third column the mean CAS if both the lower and the upper quartile are removed. The upper half depicts the data for males and the lower half for females.

Country	Average M	Lower M	Mid M
New Zealand	1,867	728 (39%)	839 (44%)
Greece	6,183	2,895 (46%)	3,320 (53%)
Estonia	7,045	4,891 (69%)	5,705 (80%)
US	8,981	3,950 (43%)	4,546 (50%)
Japan	5,440	1,505 (27%)	1,781 (32%)
Sweden	13,020	6,726 (51%)	8,076 (62%)
Country	**Average F**	**Lower F**	**Mid F**
New Zealand	1,939	781 (40%)	886 (45%)
Greece	6,444	2,921 (45%)	3,429 (53%)
Estonia	8,264	6,440 (77%)	6,796 (82%)
US	9,416	4,340 (46%)	4,808 (51%)
Japan	5,770	1,640 (28%)	1,960 (33%)
Sweden	12,983	6,734 (51%)	7,906 (60%)

4.4.4 Comparing Outliers. Our previous examinations yielded the following outliers: Australia, since it has the smallest minimal CAS size (c.f. Table 1). Hong Kong since it has an outstanding ratio between female and male CAS sizes (c.f. Table 2). Japan since it has the biggest drop of an average CAS when removing the highest quartile (c.f. Table 4). The US, since it has the fifth smallest minimal

CAS size (c.f. Table 1), although only 840 of the 3,221 official counties are counted.

In this section, we inspect the cumulative distribution functions for each of these countries to elaborate on the differences in the CAS distribution of each country. In contrast to the evaluations in Section 4.1, and Section 4.2, we consider the CDF of the CAS sizes only for the attributes district, gender, and age, yielding bigger CAS sizes. We compute the graph of the CDF based on three assumptions and depict the graph of the CDF in Figure 28:

- The first CDF assumes no background knowledge (Figure 28a).
- The second CDF depicts only females in an age range between 20 and 60 (Figure 28b).
- The third CDF depicts males in an age range between 20 and 60 (Figure 28c).

The CDF, without further assumptions, provides a baseline. We can observe that a factor of at least 100 shifts the CDF of Australia compared to the other countries. This matches our observation that Australia has the smallest minimum CAS size and indicates that most CAS sizes in Australia are smaller than in the other countries. Furthermore, the CDF curve of Australia is much smoother than the other lines. This indicates that we have more data points for Australia compared to the other countries (most likely due to smaller district sizes).

A second observation is that Japan has the flattest CDF curve. This indicates that the distribution of the CAS sizes has a higher variance compared to the other countries, which is in accordance with our observations in Table 4. At the same time, the CDF curve of Hong Kong is the steepest one, indicating that the distribution of the CAS sizes has low variance.

Our third observation arises from the difference between the male CDFs (Figure 28c) and the female CDFs (Figure 28b). These CDFs look similar for each country except for Hong Kong. For Hong Kong, the female CDF is shifted by a factor of 10 to the origin. This complies with our computed male/female quotient (c.f. Table 2), which is 3.95 in Hong Kong and is the only quotient differing significantly from the value 1 which is roughly the quotient for all other countries we investigated.

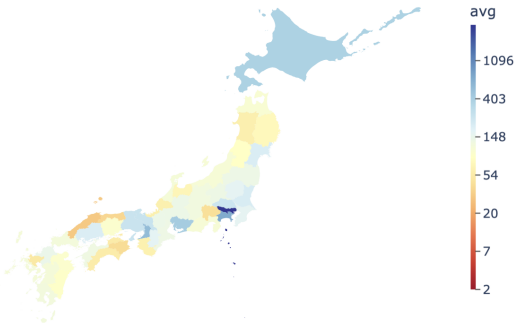
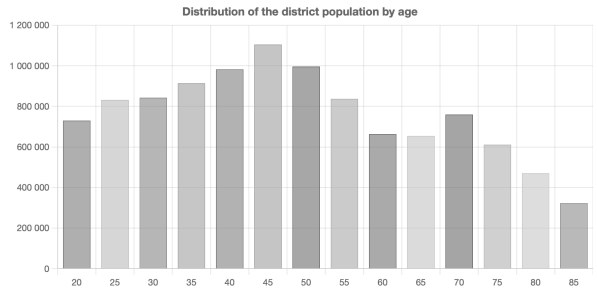
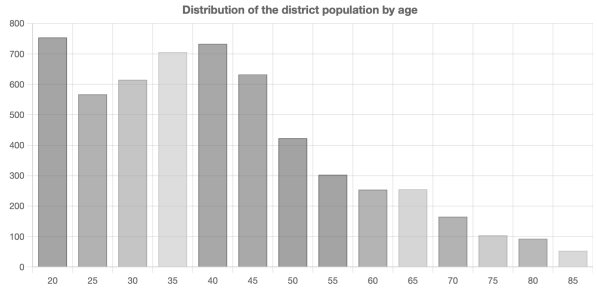


Figure 5: Average conditional anonymity set sizes for Japan for the attributes district, gender, age, height, weight.

4.4.5 Case Study: Japan. We also examine the case of Japan and show that even in a densely populated country, the conditional anonymity set (CAS) decreases to single digits for many attribute combinations. In contrast to Germany and the US, Japan has a higher average CAS size, as shown in Figure 5. To facilitate the comparison, we also incorporated the attribute's height and weight to compute the average CAS size for this figure. In Figure 5, we find that there are only a few districts with a high CAS size and multiple districts with a low CAS size. This aligns with our findings in Table 4. Furthermore, these differences become apparent if we compare one of the most populated districts (namely Tokyo-to) with a less populated district (namely Nakaguskuku-son). We do so in Figure 6 and see that the population of Tokyo-to is by a factor of 1000 higher than in Nakaguskuku-son. This confirms our previous findings that the density of the population within a district is a major influence on conditional anonymity.



(a) Population Distribution for Tokyo-to.



(b) Population Distribution for Nakaguskuku-son.

Figure 6: Comparing the population distribution in Japan of the districts Tokyo-to and Nakaguskuku-son.

To illustrate an example where the conditional anonymity set (CAS) is a single-digit number, we examine the ward Akaiwa-shi. This ward has a total population of 13,676, of which 5,225 are males. Notably, there are only 239 individuals in the specific demographic category of males aged 20, as shown in Figure 7.

When we narrow down our assumptions by assigning a height of 172kg, we have 57 people in the CAS, and when we set a weight of 62kg, we are left with only 6 people in this CAS.

5 CONCLUSION

In this work, we analyzed the degree of anonymity of several subgroups in 39 countries based on attributes that are frequently collected by health apps. We derived how anonymity set sizes can be

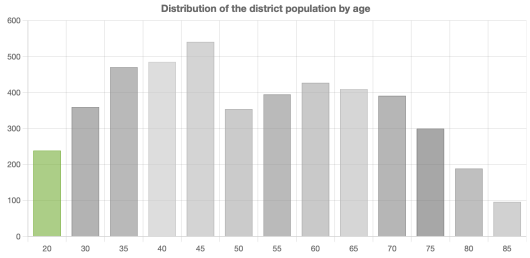


Figure 7: Male population in Akaiwa-shi, Japan.

used to measure the adversary's success in our model and proposed calculating those set sizes based on population statistics only. This way, we can quantify the privacy of individuals without having access to concrete instantiations of the datasets.

We evaluate our approach through two case studies on popular health apps in the US and Germany. We investigated the anonymity sets under different adversarial background knowledge. In the USA, 97% of the counties have average anonymity set sizes of less than 5 for women between ages 18 and 60, and the average anonymity set size for the whole US population is even smaller. For the RKI application, assuming knowledge of participation, the sets can be reduced to a single member for over 87% of the participants. We also demonstrated that a simple attribute such as "owning a smartwatch" reduces the anonymity set sizes to less than 20 for more than a third of the population. These small anonymity sets allow unique identification and de-anonymization of the individuals in case of data breaches.

We then present overarching findings from 39 different countries. These corroborate our previous results: being an inhabitant of a less densely populated district significantly impacts one's anonymity set. Thus, sharing one's district with a health app can often deteriorate the expected anonymity. We also evaluate gender imbalances and find that Eastern European countries provide larger anonymity sets to female than male inhabitants. In Nordic countries, we observe a reversal of this trend. We also investigate findings in places such as Hong Kong and Japan and identify vulnerable subgroups. We validate the accuracy of our approach using a fictional dataset of 102.5 million users.

We also developed an online tool called VisualAnon (https://visualanon.org), which allows exploring these sets. We aim to increase awareness of conditional anonymity and provide an easy-to-use tool for users to determine their vulnerability to deanonymization attacks.

5.1 Possible Mitigations

While evaluating mitigation techniques goes beyond the scope of this work, we want to highlight possible directions for future work. As outlined in Section 2, our community has already created a large body of literature on protecting health data. This includes early approaches such as k-anonymity [34] and l-diversity [24] as well as many variations of Differential Privacy [17], which provides provable privacy against strong adversaries. Unfortunately, the adoption of such techniques is scarce. A notable exception is the use of Differential Privacy by tech giants such as Google and Apple.

While employing privacy-enhancing methods would be the preferred solution, other simple steps could be taken to increase anonymity sets for existing applications. Such measures include increasing the bucket size of crucial attributes or simply collecting less data. Since sparsely populated districts are a primary factor in anonymity set sizes, this attribute could be a good target. Moreover, our analysis and tool VisualAnon enable app developers to tailor bucket sizes depending on vulnerable groups. If necessary, data could be collected in variable bucket sizes to guarantee minimum anonymity set sizes, an approach that resembles the guarantees of k-anonymity.

ACKNOWLEDGMENTS

This work was partially supported by Deutsche Forschungsgemeinschaft as part of the Research and Training Group 2475 "Cybercrime and Forensic Computing" (grant number 393541319/GRK2475/1-2019), grant 442893093, by the state of Bavaria at the Nuremberg Campus of Technology (NCT) which is a research cooperation between the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and the Technische Hochschule Nürnberg Georg Simon Ohm (THN), and by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation program in the scope of the CONFIDENTIAL6G project under Grant Agreement 101096435, The contents of this publication are the sole responsibility of the authors and do not in any way reflect the views of the EU. We thank the anonymous reviewers for their very helpful comments and suggestions.

REFERENCES

[1] The census bureau's simulated reconstruction-abetted re-identification attack on the 2010 census. https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attack-on-the-2010-census.html. Accessed: 2024-02-15.
[2] Corona data donation app. https://corona-datenspende.de/science/en/. Accessed: 2021-10-08.
[3] Flo Anonymous Mode. https://flo.health/press-center/flo-launches-anonymous-mode. Accessed: 2022-11-10.
[4] Flo Privacy Policy. https://flo.health/privacy-policy. Accessed: 2022-11-10.
[5] German Statistical Office Population. https://www-genesis.destatis.de/genesis/online. Accessed: 2021-10-08.
[6] RKI App Donor Demographics. https://corona-datenspende.de/science/en/reports/demographie/. Accessed: 2021-10-08.
[7] RKI App Donor Regional Distribution. https://corona-datenspende.de/science/en/reports/users/. Accessed: 2021-10-08.
[8] Robert Koch-Institut. https://www.rki.de. Accessed: 2021-10-08.
[9] Robert Koch-Institut. https://en.wikipedia.org/wiki/Robert_Koch_Institute. Accessed: 2021-10-08.
[10] Smartwatch users in germany. https://www.destatis.de/DE/Presse/Pressemitteilungen/Zahl-der-Woche/2020/PD20_39_p002.html. Accessed: 2021-10-08.
[11] The Guardian. https://www.theguardian.com/us-news/2022/aug/10/facebook-user-data-abortion-nebraska-police. Accessed: 2022-11-11.
[12] Total population in the united states by gender from 2010 to 2027. https://www.statista.com/statistics/737923/us-population-by-gender/. Accessed: 2024-02-15.
[13] United States Census Bureau. https://www.census.gov. Accessed: 2022-11-10.
[14] Michael Backes, Pascal Berrang, Anne Hecksteden, Mathias Humbert, Andreas Keller, and Tim Meyer. Privacy in Epigenetics: Temporal Linkability of MicroRNA Expression Profiles. page 18.
[15] Karl E Bergmann and Gert B Mensink. Körpermasse und Übergewicht. *Gesundheitswesen*, 61:S115–20, 1999.
[16] Pascal Berrang, Mathias Humbert, Yang Zhang, Irina Lehmann, Roland Eils, and Michael Backes. Dissecting Privacy Risks in Biomedical Data. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 62–76, London, April 2018. IEEE.
[17] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Proceedings of the 42nd ACM symposium*

on Theory of computing - STOC '10, page 715, Cambridge, Massachusetts, USA, 2010. ACM Press.
[18] Barbara Espinoza and Geoffrey Smith. Min-entropy as a resource. *Information and Computation*, 226:57–75, 2013. Special Issue: Information Security as a Resource.
[19] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. page 17.
[20] Mike Freeman. UC San Diego Health sued over data breach that may have exposed records of 500,000 patients. *The San Diego Union-Tribune*, September 2021. https://www.sandiegouniontribune.com/business/story/2021-09-23/sd-fi-ucsandiego-cyber-attack.
[21] Guido Heineck. Height and weight in germany, evidence from the german socio-economic panel, 2002. *Economics & Human Biology*, 4(3):359–382, 2006.
[22] M Hermanussen, H Danker-Hopfe, and GW Weber. Body weight and the shape of the natural distribution of weight, in very large samples of german, austrian and norwegian conscripts. *International journal of obesity*, 25(10):1550–1553, 2001.
[23] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '13, pages 1079–1087, Chicago, Illinois, USA, August 2013. Association for Computing Machinery.
[24] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.
[25] Arvind Narayanan and Edward W Felten. No silver bullet: De-identification still doesn't work. *White Paper*, 8, 2014.
[26] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.
[27] Cynthia L Ogden. *Mean body weight, height, and body mass index: United States 1960-2002*. Number 347. Department of Health and Human Services, Centers for Disease Control and …, 2004.
[28] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev.*, 57:1701, 2009.
[29] Charlie Osborne. Over 60 million wearable, fitness tracking records exposed via unsecured database. *ZDNet*, September 2021. https://www.zdnet.com/article/over-60-million-records-exposed-in-wearable-fitness-tracking-data-breach-via-unsecured-database/.
[30] Danny Palmer. Unsecured servers and cloud services: How remote work has increased the attack surface that hackers can target. *ZDNet*, June 2021. https://www.zdnet.com/article/unsecured-servers-and-cloud-services-how-remote-work-has-increased-the-attack-surface-that-hackers-can-target/.
[31] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1):3069, 2019.
[32] Mark A Rothstein. Is deidentification sufficient to protect health privacy in research? *Am J Bioeth*, 10(9):3–11, Sep 2010.
[33] Latanya Sweeney. Simple demographics often identify people uniquely. *Health (San Francisco)*, 671(2000):1–34, 2000.
[34] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
[35] Sam Thielman. Your private medical data is for sale – and it's driving a business worth billions. *The Guardian*, January 2017. https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns.
[36] Florian Tramèr, Zhicong Huang, Jean-Pierre Hubaux, and Erman Ayday. Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1286–1297, Denver Colorado USA, October 2015. ACM.
[37] Caroline Uhler, Aleksandra Slavkovic, and Stephen Fienberg. Privacy-Preserving Data Sharing for Genome-Wide Association Studies. *J Privacy Confidentiality*, 5, May 2012.
[38] Marc Wiedermann, Annika H Rose, Benjamin F Maier, Jakob J Kolb, David Hinrichs, and Dirk Brockmann. Evidence for positive long-and short-term effects of vaccinations against covid-19 in wearable sensor metrics–insights from the german corona data donation project. *arXiv preprint arXiv:2204.02846*, 2022.
[39] Fei Yu, Stephen E. Fienberg, Aleksandra Slavković, and Caroline Uhler. Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies. *Journal of biomedical informatics*, 50:133–141, August 2014.

A VISUALANON EXAMPLE: BRISTOL

In this section, we demonstrate how VisualAnon estimates the conditional anonymity set using an example from Bristol. Yet, we encourage readers to explore VisualAnon directly at https://visualanon.org to estimate their own conditional anonymity sets. In this sample case, we evaluate the CAS of a 25–29-year-old male from Bristol, UK, with a height between 180–184cm and a weight between 90–94kg. VisualAnon categorizes this request in the following result:

- 63,182,180 People live in the United Kingdom
- 428,235 of them in Bristol, City Of
- 172,750 are male
- 20,605 are in the ['25'] years bucket
- 5,248 of them are ['180'] cm high
- 573 of them weight ['90'] kg

With our example request, the candidate has a conditional anonymity set of 573 people. Besides showing the CAS, VisualAnon also allows a user to comprehend how the CAS can be influenced by changing each attribute or by expanding the buckets of the attributes. We visualize the evaluation of VisualAnon for our given attributes in Figure 8. The selected attributes are highlighted in green color.

B ACCURACY OF THE CAS

In this section, we provide additional measurements on the accuracy of the CAS.

C ADDITIONAL STATISTICS

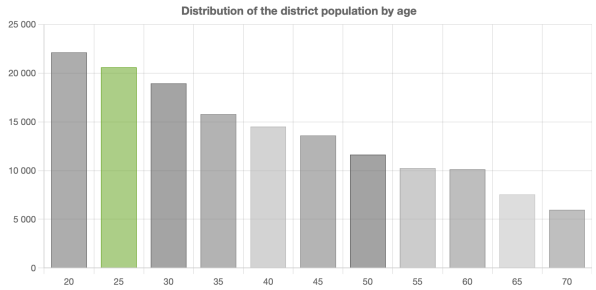
The Statistisches Bundesamt has provided us with additional statistics on the body height and body weight for table *12211-9018* in the Genesis database. The relevant statistics can be found in Table 5.

D ADDITIONAL FIGURES

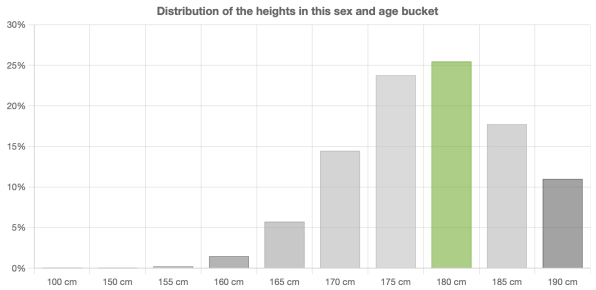
In this section, we provide additional figures from our evaluations.

Gender	Age group	Height (cm)			Weight (kg)		
		Interval	Mean	Standard Deviation	Interval	Mean	Standard Deviation
female	[18, 20)	[100, 199]	167.6	6.8	[33, 199]	61.4	10.9
	[20, 25)	[100, 197]	167.6	6.8	[34, 175]	63.0	11.5
	[25, 30)	[100, 205]	167.3	6.6	[30, 170]	65.4	12.9
	[30, 35)	[117, 195]	167.2	6.6	[34, 180]	67.0	13.9
	[35, 40)	[120, 201]	167.4	6.6	[36, 200]	67.9	14.0
	[40, 45)	[100, 200]	167.2	6.7	[35, 180]	68.8	13.9
	[45, 50)	[100, 201]	167.2	6.6	[34, 180]	69.4	13.7
	[50, 55)	[120, 197]	166.7	6.4	[30, 200]	70.1	14.1
	[55, 60)	[122, 200]	165.8	6.4	[30, 182]	70.4	13.9
	[60, 65)	[120, 225]	164.8	6.3	[33, 200]	70.9	14.0
	[65, 70)	[100, 225]	163.9	6.2	[30, 160]	71.3	13.7
	[70, 75)	[120, 192]	163.8	6.1	[32, 186]	70.5	13.3
male	[18, 20)	[150, 206]	181.2	7.8	[33, 178]	75.9	13.1
	[20, 25)	[117, 225]	181.2	7.6	[32, 200]	79.4	14.0
	[25, 30)	[140, 213]	180.8	7.5	[40, 200]	82.8	14.6
	[30, 35)	[128, 210]	180.5	7.3	[39, 200]	84.6	14.9
	[35, 40)	[120, 212]	180.4	7.2	[45, 200]	86.0	15.1
	[40, 45)	[120, 217]	180.2	7.3	[38, 200]	87.4	15.1
	[45, 50)	[100, 210]	179.9	7.3	[40, 200]	87.7	15.0
	[50, 55)	[100, 210]	179.5	7.2	[30, 200]	87.9	15.0
	[55, 60)	[120, 219]	178.7	7.0	[32, 197]	87.8	15.0
	[60, 65)	[105, 208]	177.7	6.8	[34, 200]	87.4	14.6
	[65, 70)	[120, 225]	176.5	6.7	[35, 200]	86.6	14.4
	[70, 75)	[105, 200]	175.8	6.7	[33, 200]	84.9	13.8

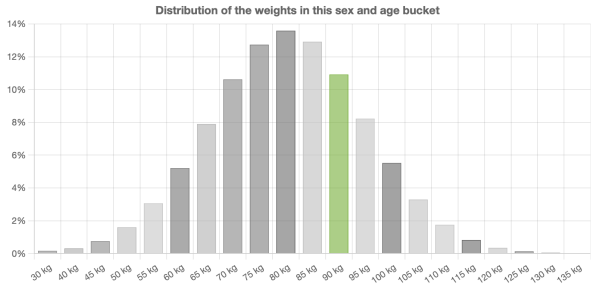
Table 5: Minimum, maximum, mean and standard deviation for table *12211-9018* provided by the German Federal Statistical Office [5].



(a) Population in the district of Bristol, UK.



(b) Height distribution.



(c) Weight distribution.

Figure 8: Example evaluation of VisualAnon for a 25–29-year-old male from Bristol, UK, with a height between 180–184cm and a weight between 90–94kg.

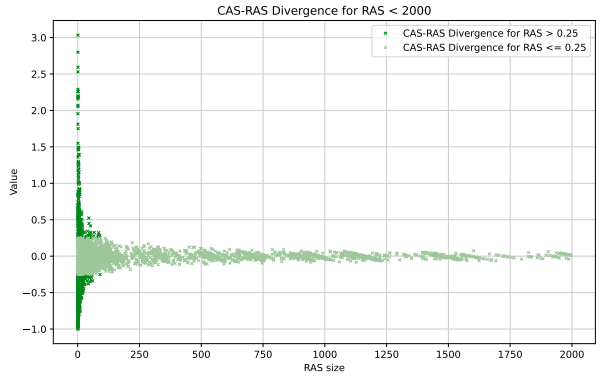


Figure 9: CAS-RAS divergence for RAS sizes below 2000.

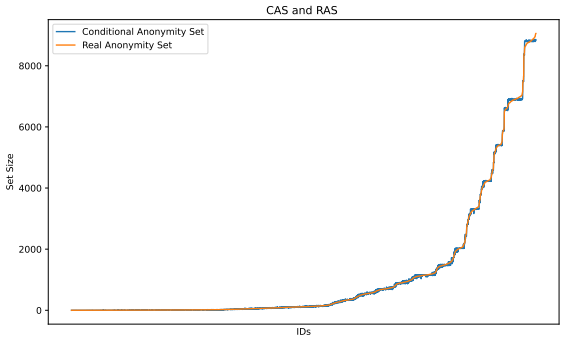


Figure 10: CAS and RAS for each user, showing the relationship and distribution across the full range of values.

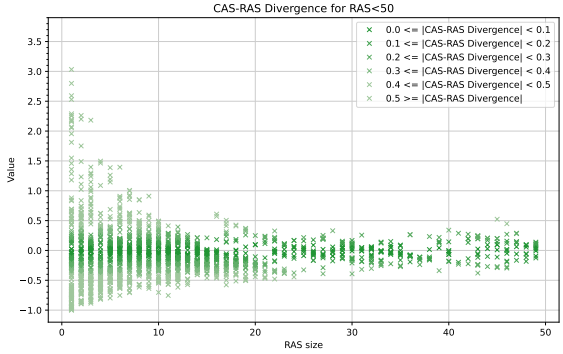


Figure 11: CAS-RAS divergence for RAS $\leq$ 50, highlighting the differences in smaller RAS values.

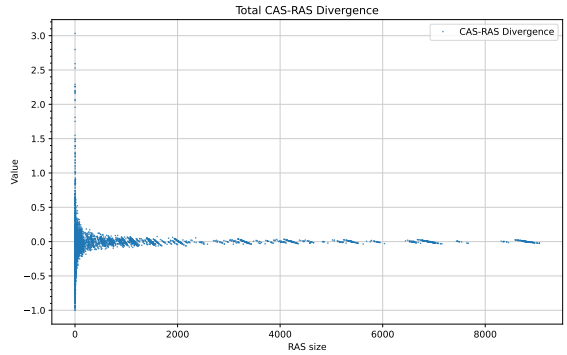


Figure 12: Total CAS-RAS divergence, illustrating how the absolute divergence decreases as RAS increases.

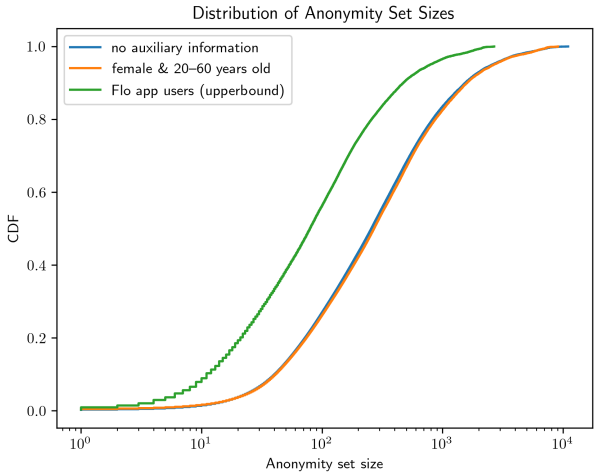


Figure 13: CDF of anonymity set sizes for different auxiliary information for the US data.

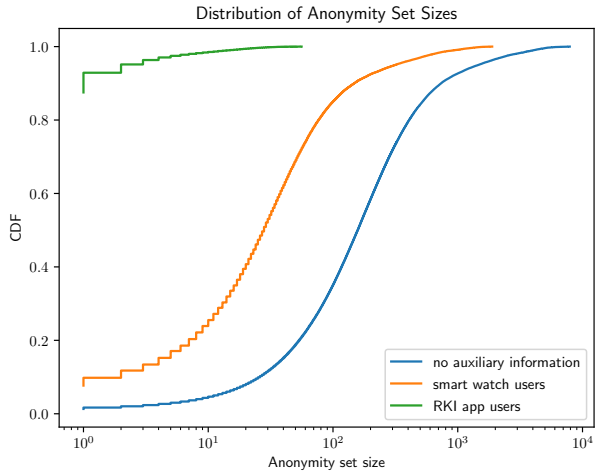


Figure 14: CDF of anonymity set sizes for different auxiliary information for the German data.

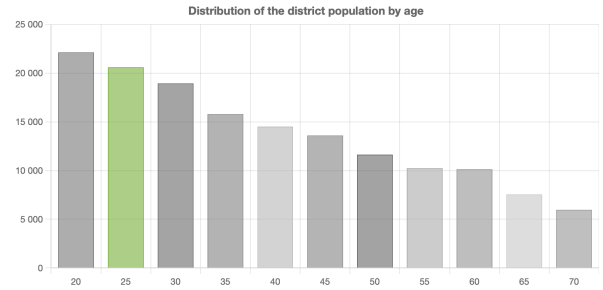


Figure 15: Population in the district of Bristol, UK.

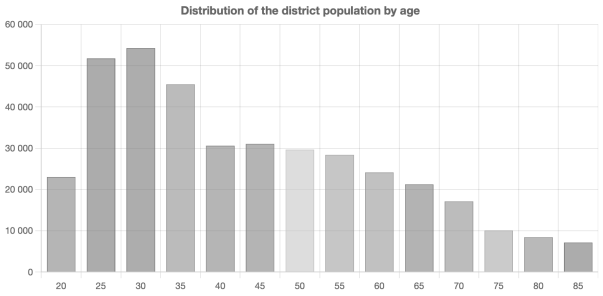


Figure 16: Male population distribution in San Francisco.

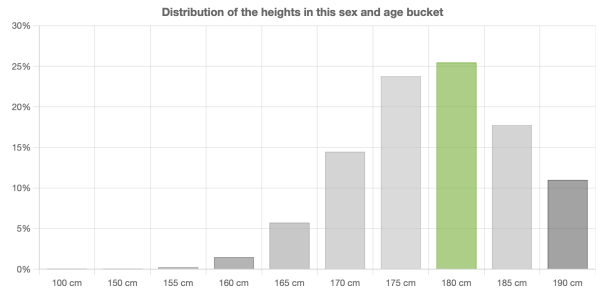


Figure 17: Height distribution for a 25–29-year-old male from Bristol, UK, with a height between 180–184cm.

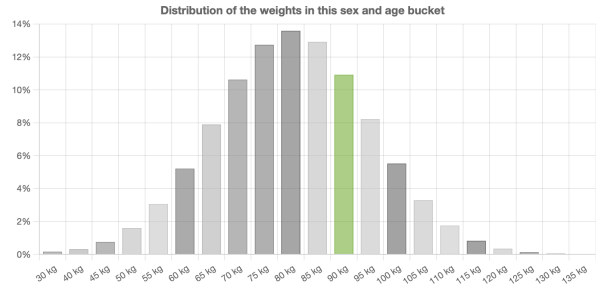


Figure 18: Weight distribution for a 25–29-year-old male from Bristol, UK, with a weight between 90–94kg.

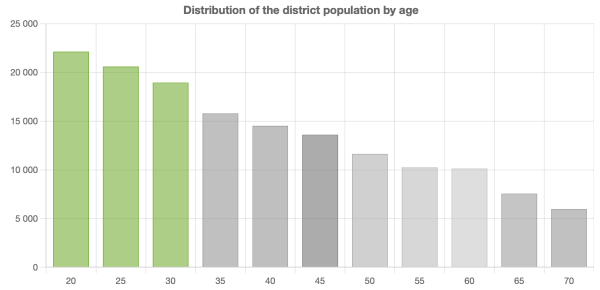


Figure 19: Selecting Multiple ages in VisualAnon, at the example of Bristol, UK.

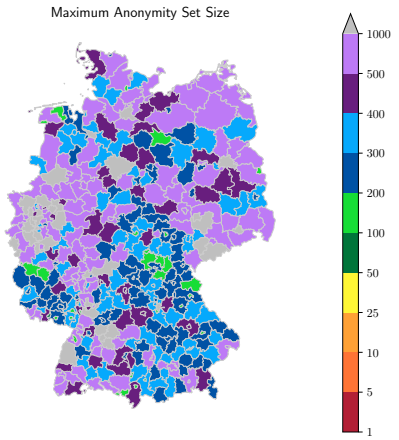


Figure 20: Maximum set sizes of districts in Germany, no background knowledge.

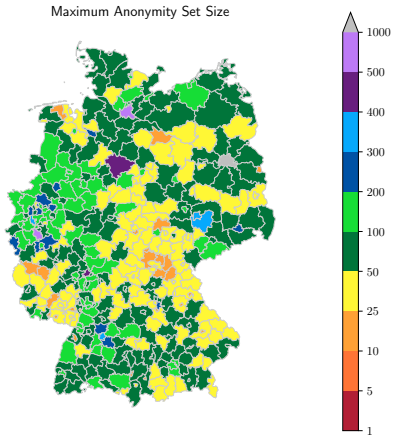


Figure 21: Maximum set sizes of districts in Germany, smartwatch users.

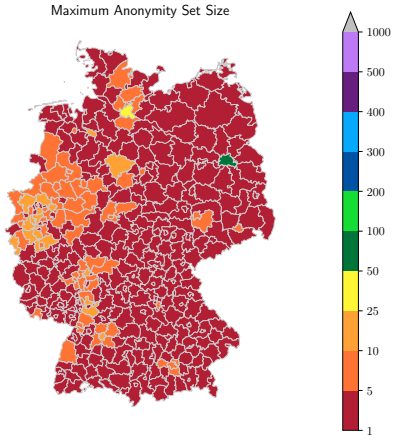


Figure 22: Maximum set sizes of districts in Germany, RKI participants.

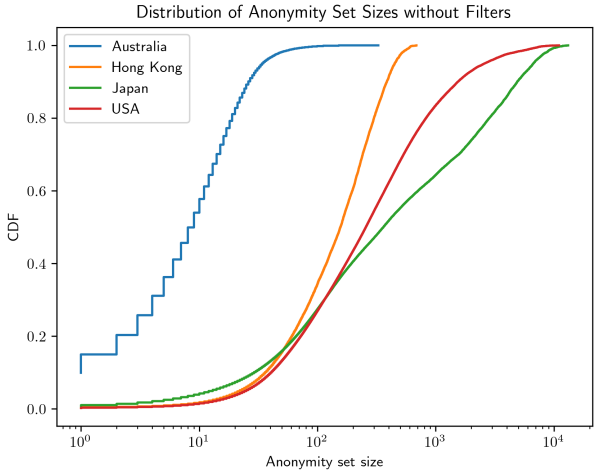


Figure 23: Cumulative distribution function comparison across Australia, Hong Kong, Japan, US, and Germany without any background knowledge.

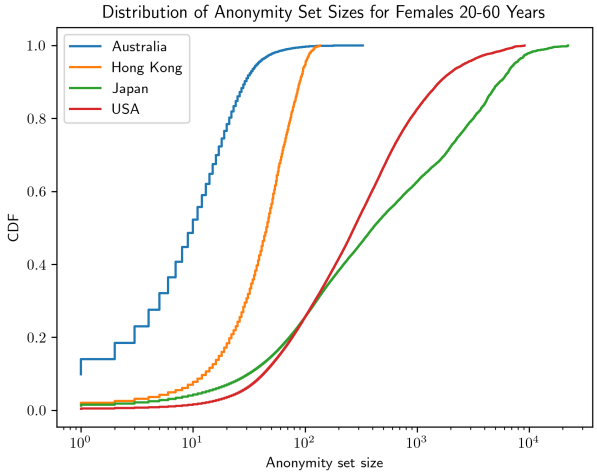


Figure 24: Cumulative distribution function for females aged 20–60 years in Australia, Hong Kong, Japan, US, and Germany.

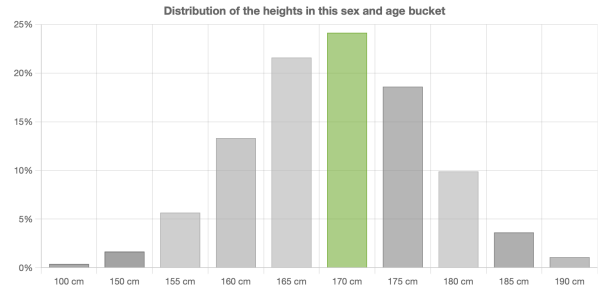


Figure 25: Height distribution of Japanese males aged 20-30 years from Akaiwa-shi, Japan, with a height between 170-175cm.

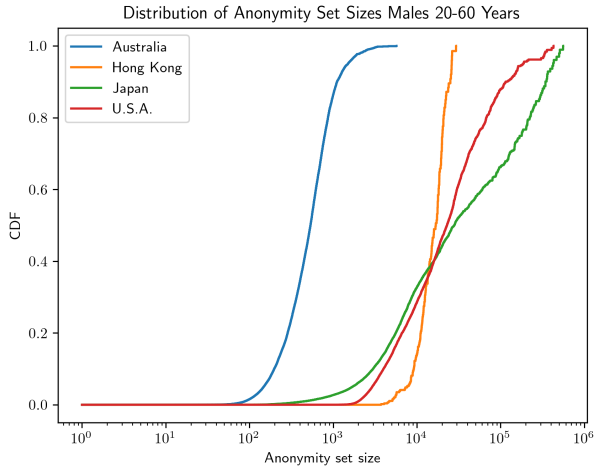


Figure 26: Comparison of the cumulative distribution functions of Australia, Hong Kong, Japan, and the US based on the attributes district, gender, and age for males, 20–60 years old.

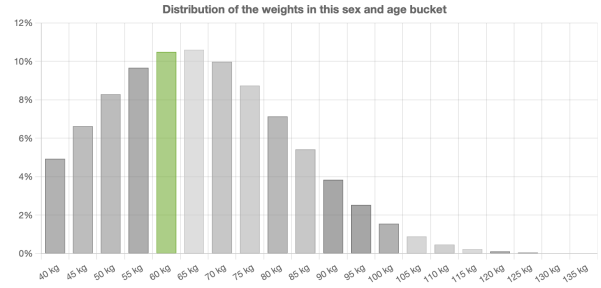
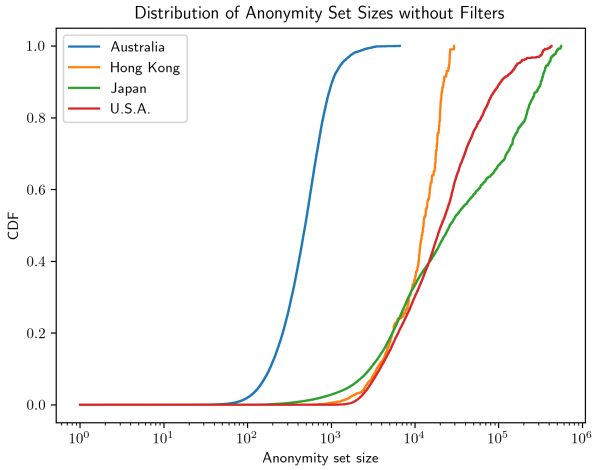
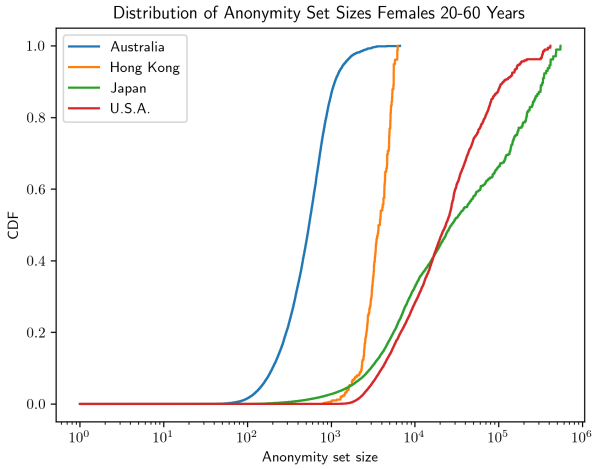


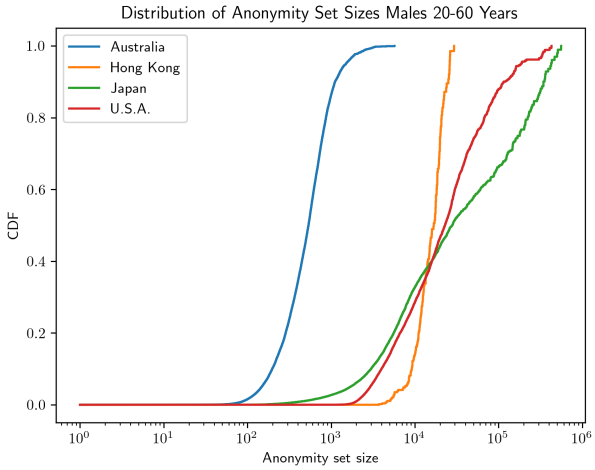
Figure 27: Weight distribution of Japanese males aged 20-30 years from Akaiwa-shi, Japan, with a weight between 60-65kg.



(a) No background knowledge.



(b) Females, 20–60 years old.



(c) Males, 20–60 years old.

Figure 28: Comparison of the cumulative distribution functions of Australia, Hong Kong, Japan, and the US based on the attributes district, gender, and age with different background assumptions.