How Attacker Knowledge Affects Privacy Risks: An Analysis Using Probabilistic Programming

Louise Halvorsen
IT University of Copenhagen
Copenhagen, Denmark

Siv L. Steffensen
IT University of Copenhagen
Copenhagen, Denmark

Willard Rafnsson
IT University of Copenhagen
Copenhagen, Denmark

Oksana Kulyk
IT University of Copenhagen
Copenhagen, Denmark

Raúl Pardo
IT University of Copenhagen
Copenhagen, Denmark

ABSTRACT

Governments and businesses routinely disclose large amounts of private data on individuals, for data analytics. However, despite attempts by data controllers to anonymise data, attackers frequently deanonymise disclosed data by matching it with their prior knowledge. When is a chosen anonymisation method adequate? For this, a data controller must consider attackers befitting their scenario; how does attacker knowledge affect disclosure risk?

We present a multi-dimensional conceptual framework for assessing privacy risks given prior knowledge about data. The framework defines three dimensions: distinctness (of input records), informedness (of attacker), and granularity (of anonymisation program output). We model three well-known types of disclosure risk: identity disclosure, attribute disclosure, and quantitative attribute disclosure. We demonstrate how to apply this framework in a health record privacy scenario: We analyse how informing the attacker with COVID-19 infection rates affects privacy risks. We perform this analysis using Privug, a method that uses probabilistic programming to do standard statistical analysis with Bayesian Inference.

CCS CONCEPTS

• **Security and privacy** → **Data anonymization and sanitization**; **Privacy protections**; • **Mathematics of computing** → **Bayesian computation**; **Information theory**; • **Applied computing** → *Health informatics*.

KEYWORDS

privacy, anonymization, probabilistic programming, health privacy

ACM Reference Format:
Louise Halvorsen, Siv L. Steffensen, Willard Rafnsson, Oksana Kulyk, and Raúl Pardo. 2022. How Attacker Knowledge Affects Privacy Risks: An Analysis Using Probabilistic Programming. In *Proceedings of the 2022 ACM International Workshop on Security and Privacy Analytics (IWSPA '22), April 27, 2022, Baltimore, MD, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3510548.3519380

IWSPA '22, April 27, 2022, Baltimore, MD, USA.

1 INTRODUCTION

Motivation. Privacy concerns of individuals are routinely violated by businesses and governments [2–4, 17]. With the emergence of big data analytics, these entities increasingly collect and disclose large quantities of personal data, for public-good and profit. In response, several regulations have emerged which call for appropriate measures for data protection, including anonymisation of personal data [25, 27, 28]. Unfortunately, such measures are non-trivial to implement; seemingly anonymised data frequently gets deanonymised, after being published. This has already resulted in the disclosure of medical-[2], financial-[4], travel-[3], and habit-records [17], on millions of people. This can lead to considerable harms: identity theft, discrimination, harassment, and more [8, 24].

The challenge faced by data controllers, is *attacker knowledge*. Attackers may obtain a lot of information from domain knowledge, personal observations, private correspondence, public media, and datasets available online. By matching disclosed data against their knowledge, an attacker can deanonymise poorly-anonymised data.

Many privacy protection mechanisms exist, e.g. k-anonymity [26], l-diversity [15], t-closeness [13], and differential privacy [7], which each provide strong privacy guarantees. However, these mechanisms have fixed models of attacker knowledge; if none is an perfect fit for the scenario, then the data controller must compromise on either privacy or utility. Furthermore, the use of these methods is not widespread in practice [10]; instead, data controllers tend to simply remove identifiers or aggregate data, which unfortunately makes deanonymising records easy for attackers [2–4, 8, 17, 24].

Data controllers should *quantify* the privacy risks of the anonymisation methods they use. Doing so is non-trivial, due to the variety of disclosure risks that can occur [16, 29, 30]. This is even more challenging when no concrete dataset is available for the analysis, e.g. in the case where risks need to be estimated *before* said data is collected (e.g. for *data minimisation* and *privacy impact assessment*).

Framework. We present a multi-dimensional conceptual framework (Sect. 3) for assessing privacy risks given prior knowledge about data. The framework defines three dimensions, along which a data controller performs this assessment. *Distinctness*, a quality of the input data, is the extent of which input data records are uniquely identifiable. Attribute values and combinations may be (un)common in the data set. *Informedness*, a quality of the attacker, is the extent of which the attacker is informed about the input data, prior to disclosure. An attacker might know only its structure, might know detailed public statistics about records, or might even already know

some, or most, records. *Granularity*, a quality of the disclosure program, is the extent of which data has been distorted, or reduced in resolution, during disclosure. Examples include dropping identifying attributes, adding random noise, and making values coarser.

To further aid the data controller in this assessment, we model three well-known [16, 30] types of disclosure risk, each represented by a *query*. *Identity disclosure* is when an attacker identifies which record belongs to a data subject. *Attribute disclosure* is when an attacker learns the value of an attribute of a data subject, without necessarily knowing which record belongs to the data subject. *Quantitative attribute disclosure* is when an attacker changes their belief about the value of an attribute of a data subject. These disclosures are in decreasing order of severity, from disclosing a record, to disclosing an attribute, to disclosing information about an attribute.

We demonstrate (Sect. 5 and 6) how to apply this framework to systematically perform privacy risk analysis. First, a data controller defines the range of each dimension in the framework. A choice from each of these ranges yields a concrete analysis *scenario*. The data controller then performs the analysis for each scenario, by issuing their queries on it, the result of which is the disclosure risk for that scenario. The result is a complete picture of how changes along each dimension affect disclosure risk. We analyse each scenario semi-automatically using Privug [18]. Privug (Sect. 4) is a method that uses Bayesian inference to analyse privacy risks in programs, in a *probabilistic programming* language. Note that the framework is *independent* of the tools used to analyse the scenarios; LeakWatch [1] could for instance be used instead.

Together, this constitutes a systematic approach for exploring how prior knowledge affects privacy risk. In contrast to an ad-hoc, unstructured exploration, a data controller now has concrete dimensions, queries, and methods, for making a privacy risk assessment.

Health Data Privacy. Our approach is a significant contribution to health data privacy. To show this, we (in our demonstration) analyse disclosure risks associated with publishing COVID-19 infection rates in Denmark. We focus on the public datasets that Statista (`www.statista.com`) regularly publishes on infection rates in the Danish population [6]. These datasets are aggregated into age groups. Our goal is to *quantify privacy risks for all age groups, and find out which age groups are most vulnerable.* Applying our conceptual framework, we proceed by instantiating the three dimensions. Granularity consists of two anonymisation methods: `attr_r` drops identifying attributes, and `attr_g` reduces the granularity of attributes. Informedness consists of two attackers, who vary in how much prior knowledge they have about infection rates. The first attacker, uniform, knows nothing about infection rates, whereas the second, covid19, is informed about publicly-available statistics on the first few months of the COVID-19 outbreak in Denmark. Distinctness consists of the 112 different age groups that we suppose that attacker is trying to learn about. This yields a total of $112 * 2 * 2 = 448$ scenarios, each of which we perform our three queries on using Privug. The result is an extensive and detailed analysis of privacy risks, presented in Figs. 6 to 8. Summarised:

(1) Using the uniform prior results in a worst-case overestimation of identity and attribute disclosure risks for people aged 0-79, but results in an *under*estimation of these same risks for people in the age group ≥ 80. (Sections 6.1 and 6.2).
(2) Reducing granularity of quasi-identifiers (e.g. age, birthday, zip [9, 23, 26]) reduces attribute disclosure risk by at least 20% in the covid19 prior and 10% for uniform. But doing so is insufficient for protecting age group ≥ 80 (up to 80% attribute disclosure risk, and only offers low protection for people aged 0-79 (up to 50% attribute disclosure risk). (Section 6.2).
(3) Using the uniform prior results in a worst-case overestimation of quantitative disclosure risks—both for ordinary records (30 year old people) and outliers (90 year old people). However, for the covid19 prior, the attacker learns proportionally—the risk increases by up to 30%. (Section 6.3).

As evident by the comprehensiveness of the above results, our approach is a significant step forward in the area of privacy risk analysis of health records in general, especially due of pt. iii) below.

Contribution. Our contributions include:

(1) a multidimensional conceptual framework for assessing how attacker knowledge affects privacy risk:
   (a) three dimensions along which assessment is performed,
   (b) model of three well-known types of disclosure risk;
(2) demonstration of how to apply this framework to perform privacy risk analysis using probabilistic programming;
(3) approach to obtain comprehensive privacy risk results in health data privacy using a Bayesian model of input data.

Our approach stands out in three important ways. i) We provide an semi-automatic way to quantify well-known notions of disclosure risks. This is in contrast to to existing works, which require a non-negligible level of expertise to be carried out. ii) We are analysing the performance of a *anonymization method*, not how vulnerable a specific dataset is. iii) The model of the input dataset is Bayesian; while we can model a specific input dataset, we can furthermore model varying degrees of uncertainty about the input dataset, corresponding to what an attacker could reasonably know. Notably, this enables us to analyse data protection measures *before* we collect data, à la uniform. This facilitates data minimisation, included in several privacy regulations and guidelines (see e.g. GDPR [27], Article 5), and reduces disclosures resulting e.g. from server breaches. This is in stark contrast with existing risk estimation methods, which have to be applied on a particular dataset.

All data relevant to our demonstration, i.e. public datasets of Danish demographics and COVID-19 infection, probabilistic models, anonymisation programs, and the complete set of results in this paper, are available in our public repository [21].

2 MOTIVATION (PRIVACY VIOLATIONS)

The overall problem is privacy violations stemming from insufficient anonymisation of disclosed records. This problem is very broad; to illustrate this, we present three scenarios (one in detail) showcasing varied privacy concerns: health, personality, and location. These examples furthermore highlight recurring elements.

Health. Consider the anonymisation of medical records. The input data has five columns: name, zip code, birthday, sex, and diagnosis. Consider an anonymisation algorithm (`hea`) which simply drops the name column, as illustrated in Fig. 1a. Suppose that a data controller applies this algorithm on the input data, and discloses the result.

name	zip	birthday	sex	diagnosis
Alice	2300	15.06.1956	F	ill
Bob	2305	15.06.1956	M	healthy
Carol	2300	09.10.1925	F	healthy
Dave	2310	01.01.2000	M	ill

$\xrightarrow{\texttt{hea}}$

zip	birthday	sex	diagnosis
2300	15.06.1956	F	ill
2305	15.06.1956	M	healthy
2300	09.10.1925	F	healthy
2310	01.01.2000	M	ill

$\oplus$

name	zip	birthday	sex
Mark	2450	30.09.1977	M
Rose	2870	24.12.1985	F
Alice	2300	15.06.1956	F
Dave	2310	01.01.2000	M

$\rightsquigarrow$ Alice is ill.

(a) Health records anonymized by dropping an identifying attribute (name). **(b) Linking attack reveals Alice's record.**

Figure 1: Anonymized health record reidentified by means of a linking attack.

name	zip	birthday	sex	diagnosis
Alice	2300	15.06.1956	F	ill
Bob	2770	15.06.1956	M	healthy
Carol	2300	07.03.1997	F	healthy
Dave	2770	01.01.2000	M	ill
Eve	2300	22.05.1968	F	ill
Frank	2770	23.12.1961	M	healthy
Grace	2300	19.11.1991	F	healthy
Harry	2770	29.08.2002	M	ill

$\xrightarrow{\texttt{hea'}}$

zip	sex	age	diagnosis
2770	M	M	healthy
2770	M	M	healthy
2770	M	Y	ill
2770	M	Y	ill
2300	F	Y	healthy
2300	F	Y	healthy
2300	F	M	ill
2300	F	M	ill

$\oplus$ Alice is a middle-aged female living in 2300. $\rightsquigarrow$ Alice is ill.

(a) Health records anonymized by grouping values, satisfying 2-anonymity. **(b) Homogeneity attack reveals Alice's diagnosis.**

Figure 2: Anonymized health record's diagnosis revealed by means of a homogeneity attack.

Suppose that users have not consented to their diagnosis being disclosed. Despite the disclosed records being anonymised, the diagnosis of individuals might be revealed. Suppose that you already possess a dataset with zip codes, birthdays, sex, and, crucially, names as shown after the "⊕" in Fig. 1b. Since zip code, birthday, and sex, form a quasi-identifier in both datasets, a simple join of the datasets could reveal the names of the individuals from the disclosed medical records, thus revealing whether or not they are ill, as per the "⇝".

Sweeney [26] famously joined medical records disclosed by the Group Insurance Commission (GIC) in Massachusetts, with a voter registration list (which she acquired for $20), to reveal the health record of the then-governor of Massachusetts. Our example is based on this incident. This is a *linking attack*. A high percentage of the population can be uniquely identified by a combination of common demographics in this manner [9, 23]; 87% of the US population can be uniquely identified by zip code, sex, and date of birth [26].

Sufficiently anonymising data is hard. Suppose a data controller wishes to publish medical records with high utility, yet where uniquely identifying a record is impossible. She applies `hea'` on the input data, which drops names, replaces birthdays with age groups, and discloses the result, as illustrated in Fig. 2a. The released data satisfies k-anonymity ($k = 2$); no record can be uniquely identified. The disclosed data has high utility; it says that all young adult (Y) males in 2770 are ill, and all middle-aged (M) females in 2300 are ill. However, this dataset is vulnerable. Suppose an attacker knows that Alice is a middle-aged female living in 2300, and that her record is in the data set. (The attacker may be a neighbour who knows that Alice got tested.) All records matching these constraints are ill. The attacker concludes that Alice is ill. Machanavajjhala et al. [14] demonstrated this shortcoming of k-anonymity (and proposed l-diversity as a remedy). This is a *homogeneity attack*.

Personality. Narayanan and Shmatikov [17] linked the Netflix prize dataset, containing anonymised movie ratings of 500,000 Netflix subscribers, with (public) profiles from IMDB. They succeeded despite Netflix injecting noise into the data during the anonymisation process. Movie watching habits and ratings reveal intimate details of people, such as *political orientation* (e.g. the movie "Fahrenheit 9/11"), *religious views* ("Jesus of Nazareth"), and *sexual orientation* ("Queer as Folk") [17]. This is in line with results from sociology and psychology research [12], which concludes that our movie watching habits and ratings are indicators of how we rank on the Big Five personality factors [11]. For instance, Alice may like science-fiction and fantasy *because* she is creative and adventurous, albeit reserved. Inferring such facts from released data is an *inference attack*.

Location. Culnane et al. [3] showed that it only takes two datapoints to identify an individual in anonymised travel records for 15 million travel cards, disclosed by Public Transport Victoria (PTV), Melbourne, Australia. This can be done by linking travel records with event participation information from Facebook or `meetup.com`. Revealing travel information on individuals can reveal intimate details of people; where they work, where they sleep, whom they travel with. Travelling frequently to the same location may reveal *health information* (area has a hospital), *religious affiliation* (church), *sexual orientation* (gay bars), or *substance abuse* (pusher street).

Disclosing records that depends on personal information poses a privacy risk: an attacker may relate them with information they already possess, and deanonymise them. Suppose you are the data controller, faced with the task of anonymising data. How do you know that using a certain anonymisation algorithm poses a privacy risk? Crucially: What is a good model for attacker knowledge, and how does informing it with realistic data influence risk analysis?

3 FRAMEWORK (DIMENSIONS, QUERIES)

How do you approach the above problem? Where do you start your analysis, and how will you know that you have high-confidence results? For this, we present a conceptual framework, which serves as an analytical tool by introducing concepts which help the data controller think about and approach this problem. Later, we will see a concrete manner in which the data controller can solve this problem in an semi-automated, systematic manner, using Privug.

Before we can make statements about privacy, we need to know what kind of attackers we consider in our risk assessment. In our conceptual framework, we assume the following threat model.

Threat model. We consider an attacker that i) has some prior knowledge about the input dataset, ii) who knows the program, and iii) who observes the output of the program. The goal of the attacker is to *infer additional information about the input dataset*.

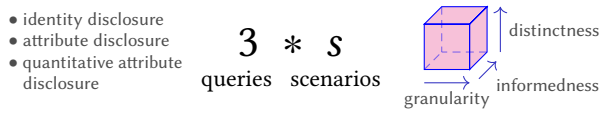


Figure 3: Overview of our framework ($s = |D| * |I| * |G|$).

Pt. i) encompasses a broad space of attackers, ranging from ones that know nothing, to ones that know everything, about the input dataset, with varying degrees of certainty. It thus subsumes attackers which perform linking or homogeneity attacks. Pt. ii) is inspired by Kerckhoff's principle; we should not rely on the program being obscure for privacy to be preserved, as an attacker might know (or infer) the program. Pt. iii) similarly does not rely on the disclosed output being obscure for privacy to be preserved (it is disclosed).

Dimensions. Our conceptual framework introduces three dimensions to aid the data controller in thinking about the above problem.

Distinctness. A quality of the input data, this dimension represents the extent of which records in the input data are uniquely identifiable. A record might be similar to many records in the input dataset, might be unique, or anywhere in between. For instance, in the input data of Fig. 1b, Carol is more uniquely identifiable than Alice with respect to age, since Alice shares her age with Bob (whereas Carol shares an age with no-one). In the personality example, viewers that watch really obscure movies are more identifiable than viewers that watch mainstream movies. Finally, in the location example, travellers that travel to odd locations at odd hours are more identifiable than travellers that travel on main transport routes in rush hour. Analysing with respect to varying degrees of distinctness enables the data controller to evaluate the extent of which the disclosure program protects *outliers* in the input data.

Informedness. A quality of the attacker, this dimension represents the extent of which the attacker is informed about the input data, prior to disclosure. For instance, in Fig. 1b, the attacker knew the zip, birthday and sex of Alice and Dave, whereas in Fig. 2, the attacker knew the zip and sex of Alice, and that she's middle-aged. In the personality example, the attacker knew which movies users had rated on IMDB, and in the location example, the attacker knew which events people had participated in. Whereas these are examples of the attacker knowing things with certainty, the attacker might also possess knowledge with varying degrees of uncertainty. For instance, the attacker might know that sexes in the input dataset are distributed with a slight bias towards females. In other words, the attacker might know nothing besides the structure of input records, might know detailed public statistics about records, or might even already know some, or most, of the records. Analysing with respect to varying degrees of informedness enables the data controller to assess how attacker knowledge affects privacy risk.

Granularity. A quality of the disclosure program, this dimension represents the extent of which data has been distorted, or reduced in resolution, during disclosure. For instance, in Fig. 1b, the disclosure program drops identifying attributes, whereas in Fig. 2, the disclosure program furthermore makes values more coarse. The disclosure program can also add random noise, as in the personality example. Analysing with respect to varying degrees of granularity enables the data controller to evaluate the effect of different anonymisation approaches, to assess which approach strikes the desired balance between privacy and utility.

Queries. To further aid the data controller, we highlight three known [16, 30], relevant types of disclosure risk, represented by queries.

Identity disclosure. Can an attacker identify precisely which record belongs to a data subject? For instance, in Fig. 1b, the attacker can identify which row belongs to Alice (the first row). This is the most severe disclosure; if the attacker pinpoints Alice's record, then the attacker learns all of Alice's attributes in the disclosed dataset.

Attribute disclosure. Does the attacker learn the value of an attribute of a data subject (without necessarily knowing which record belongs to said data subject)? For instance, in Fig. 2, while the attacker cannot pinpoint which of the rows belong to Alice, the attacker learns the diagnosis since all candidate rows are diagnosed ill. Suppose, for instance, that the input dataset furthermore contained a street address, which is preserved by the disclosure program. Then the attacker does not learn Alice's street address exactly (but rather, reduces it to two possible addresses).

Quantitative attribute disclosure. Does the attacker change their belief about the value of an attribute of a data subject? With belief being (un)certainty of knowledge, we think of these as probabilities. For instance, consider a variation of the input data in Fig. 2, where Carol and Grace's record are not present. The disclosed table then does not contain the $\langle 2300, \mathsf{F}, \mathsf{Y}, \mathsf{healthy}\rangle$ rows. Suppose that 2% of the general population is ill, and that this is known by the attacker prior to disclosure. Suppose further that the attacker knows nothing about Alice (not even her name) except that her record is in the input dataset. Then, upon seeing the disclosed dataset, the attacker's belief that Alice is ill becomes $2/3$—a huge increase. This is the least severe disclosure of the three; the attacker neither pinpoints Alice's record, nor learns an attribute with certainty.

Scenarios. Each choice along each of the three dimensions gives rise to a *scenario*, to be analysed for disclosure risk using the above queries. This is summarised in Fig. 3. Here, D, I, and G are the record types, attackers, and disclosure programs considered by the data controller. Large sets yield many scenarios, each of which needs to be analysed with the three queries. In the following, we show how to perform these analyses semi-automatically using Privug.

4 METHOD (PRIVUG)

To reason about privacy risks stemming from the use of an anonymisation algorithm, we use Privug [18]. Privug semi-automates the process of quantifying risk. We briefly recall the Privug method here, and refer the reader to [18] for further details.

In Privug, we define attacker knowledge as a probabilistic model, and use Bayesian inference to reason about what the attacker learns.

Probabilistic model. A probabilistic model describes a stochastic phenomenon in terms of random variables and their relationships. We express our probabilistic models in a programming language. Probabilistic programming is the use of programming languages for probabilistic modelling and reasoning. We use Figaro [19], a probabilistic programming language embedded in Scala. We can e.g. write `val x = Uniform(0,10)` to specify that $x \sim \mathbf{Uni}(0, 10)$, define y in terms of x, define a distribution over datasets, etc.

The steps of Privug (see Figure 4) are divided in two phases.

Modeling Phase. First, a data controller defines the scenario as a probabilistic model. The model is created in three steps:

(1) Prior. *Define the attacker's knowledge about the input of the program before observing the output.* This is modelled as a distribution over (input) *datasets*. This model is very general. With it, we can model how much the attacker knows—with varying degrees of certainty—about the size of the dataset, which values attributes can take, which records are in the dataset, and dependencies between all of these. In Sect. 5.1 we give two priors on COVID-19 infection rates for the Danish population: an uninformed attacker making reasonable guesses (uniform), and an attacker informed about publicly-released statistics on the matter (covid19).

(2) Disclosure program. *Define the probabilistic version of the anonymisation program.* The probabilistic version maps a *distribution* over inputs to a *distribution* over outputs. We invoke it on the prior to compute the attacker's prediction of the output of the anonymisation program. The probabilistic version is trivially obtained from the original anonymisation program, usually by simply updating its type signature. In Sect. 5.2 we show the probabilistic version of two anonymisation programs: one that drops a column (`attr_r`), and one that reduces granularity of data (`attr_g`).

(3) Observation. *Condition the program outputs.* This asserts evidence that the attacker has after observing the output of the anonymisation program. This is required for some types of analysis. In Sect. 5.3, for quantitative attribute disclosure, we assert that k records in the output share a distinguished record's age group.

Analysis Phase. The data controller then computes & analyzes the knowledge that the attacker obtains after disclosure. In two steps:

(4) Posterior. *Use Bayesian inference to compute the knowledge of the attacker after observing the output, i.e., posterior distribution.* Figaro includes several inference algorithms to estimate the posterior distribution automatically. We use Importance Sampling (IS) [22]. Intuitively, IS estimates the posterior by repeatedly generating sample datasets according to the prior, and running the disclosure program on each of the samples to estimate the distribution of output datasets. If observations are defined, IS rejects samples that do not satisfy the observation. For our analyses, we generate 5000 samples. This number of samples has been shown to provide a good estimation for the type of programs we analyse here [18].

(5) Posterior Analysis. *Query the prior and posterior.* By doing so, the data controller can analyse what the attacker learns about the input from the output. All statistically defined queries and leakage measures can, in principle, be used [18]. In Sect. 6, we show and evaluate the result of issuing the disclosure queries from Sect. 5.3.

5 DEMONSTRATION: SETUP

In our demonstration of our framework, we evaluate the privacy risk of two anonymisation programs intended to protect the privacy of the individuals in a COVID-19 dataset. Consider a data analyst responsible for releasing COVID-19 data (of Danish citizens). The non-anonymised dataset contains an identification number for each data subject, the COVID-19 information (a binary diagnosis: "ill" or "healthy"), and demographic data, namely, birthday, age, zip code and sex. The analyst wishes to assess the privacy risks of two anonymisation methods: only removing the identifier, or furthermore decreasing the granularity of attributes (generalise). In particular, the analyst is interested in three types of analyses: i) determine the probability that an attacker uniquely identifies a record in the dataset (identity disclosure); ii) determine the probability that an attacker learns the diagnosis (attribute disclosure); and iii) determine how certain an attacker is about the diagnosis (quantitative attribute disclosure). The analyst will do this in the presence of two attackers: a uniform prior, and a prior that is informed with publicly available demographic data and COVID-19 data from Denmark.

This scenario is the same as in the health example in Sect. 2. However, our experiment differs in three crucial ways from Sweeney's work [26]. First, we consider COVID-19 infection rates in Denmark, whereas Sweeney considered hospital records disclosed by the GIC in Massachusetts. Second, we assess the privacy risk of *anonymisation programs*, not the vulnerability of a given dataset. We do not need the original dataset—in fact, we do not possess it. Instead, we create Bayesian models of what the attacker might know about the input data. Third, instead of looking for a linking [26] or homogeneity [14] attack on a dataset, we analyse anonymisation programs broadly, against attackers with varying degrees of informedness.

We refer to the distinguished record that the attacker is attempting to infer information about (Alice) as *the victim*.

5.1 Priors (informedness, distinctness)

We investigate two different priors built with the attributes described above. The uniform prior assumes a uniform distribution of all the demographic attributes, modeling a complete lack of knowledge of how these attributes can be distributed in the datasets. This prior corresponds, for instance, to the situation at the beginning of the COVID-19 pandemic when governments did not publish infection rates yet. The COVID-19 prior takes into consideration the actual distribution of the demographics as well as COVID-19 infections in May 2020 among the Danish population, based on published statistical data, thus modelling a stronger attacker that makes use of publicly available data. In this prior, the probability of a person being ill depends on the person's age and sex.

Both priors assume the dataset has a fixed size, and that records are independent from one another. We do not consider knowledge e.g. of the form "if Bob is in the dataset, then so is Carol", or "if everyone is male, then the dataset is small" (modellable in Privug).

uniform. In the uniform prior, the attributes of sex, birthday, age and zip codes are uniformly distributed. More specifically, birthdays are uniformly distributed from a range of 0 to 364 denoting the 365 days in a year. (For the sake of simplicity, we ignore leap years in our experiment.) Age is defined in a similar way and is uniformly distributed from a range of 0 to 112. Furthermore, names are uniformly distributed from a list of 54 names. As for zip codes, around 250 different zip codes are uniformly distributed. Diagnosis has a probability of being ill at 0.2, i.e. a Bernoulli distribution. This is a (uneducated) guess, which is independent of demographics.

covid19. In the COVID-19 prior, distributions are defined from real-life, publicly-available demographics data on the Danish population.

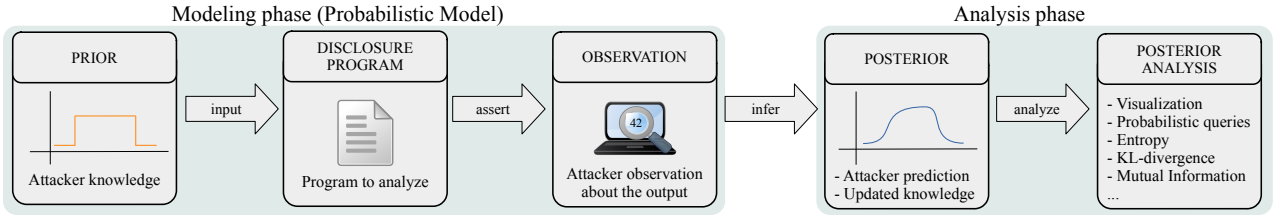
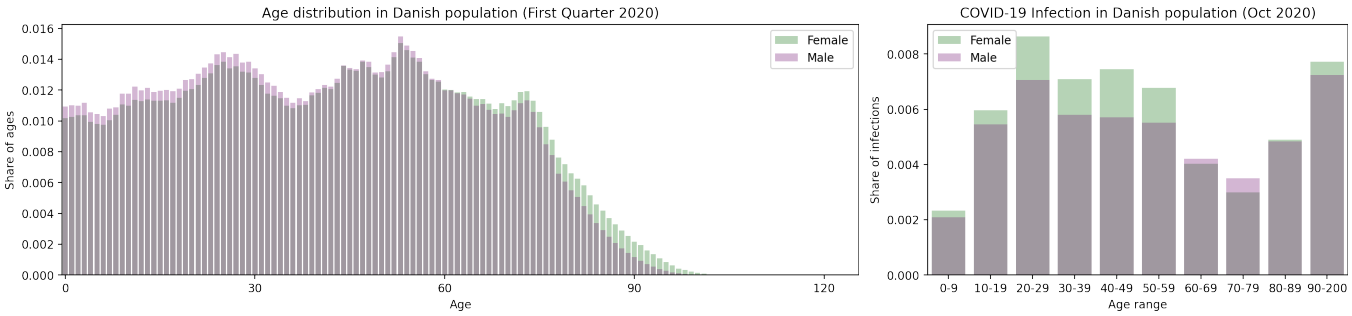


Figure 4: PRIVUG overview. Boxes represent elements of the method. Arrows mark the sequence of steps from prior to posterior.



(a) Age distribution in Denmark. **(b) COVID-19 infection rates in Denmark.**

Figure 5: Distribution of age and COVID-19 infection rates in Denmark (from publicly-available sources).

For the distribution on age and sex, we use data from Statbank [5] maintained by Statistics Denmark, the central authority on statistics on Danish society. The data is a mapping from ages (0 to 111 years), to the share of people with that age living in Denmark in the first quarter of 2020. Figure 5 (left) shows the distribution. For sex, the data is almost evenly distributed between female and male in Denmark; 51 percent female, and 49 percent male. The distribution on age depends on gender; there are more elderly females than there are elderly males, and there are more young males than there are young females. For the distribution on diagnosis, we use (freely available) data [6] from Statista, a German private company specialising in market and consumer data. This data consists of the number COVID-19 cases as of October 6, 2020, by age (grouped in intervals of 10 years) and sex. In Denmark, females have a higher infection rate than males in all age groups, save for 60-70 and 70-80 years. Zip codes and birthdays remain uniformly distributed.

Data. Figure 5a shows the distribution over age and sex in Denmark. For instance, out of all males, ~0.011 of them are 0 years old, and out of all females, ~0.010 of them are 0 years old. This corresponds to two categorical distributions (for females & males resp.). Figure 5b shows the infection probability for each age interval in Denmark. For instance looking at the first column, the share of the Danish population, who are under the age of 9, female and are infected with COVID-19 is 0,0023. That is, there is a probability of 0,23 % of being infected with COVID-19, if you are female under age of 9.

Implementation. Figure 9 shows the specification of our two priors in Figaro. The function `VariableSizeArray(size,...` defines a probability distribution over datasets w/ 500 rows (`size` ~ **Con**(500), a constant distribution). Intuitively, it defines how to sample a dataset, in terms of how to sample a record. In Fig. 9a (UNIFORM), the attributes in each row are distributed as described above. In particular, `Flip` is a Bernoulli distribution on Booleans, and `If` branches on the Boolean to map it to a diagnosis. In Figure 9b (COVID19), names, zip code and birthdays are distributed the same way as in UNIFORM. However, sex is now a Bernoulli distribution, informed by demographics data from the Danish population. Similarly, which distribution over ages to draw from when sampling a record depends on the sampled sex (`distAge`), and which distribution over diagnosis to draw from depends on the sampled sex and age (`distIll`). These functions `distAge` and `distIll` are learned from the data given in Fig. 5, described above.

Victim record. In both priors, we inject a distinguished record—the victim—that the attacker is attempting to infer information about. Her name is Alice, she's born on February 13, lives in zip code 2300, and has diagnosis "ill". However, for her age, we consider all possible ages in the Danish population, i.e., from 0 to 112.

We do this to analyse the privacy risk of all possible age groups in the Danish population, and to be able to find out the more vulnerable age groups. With this, we model the *distinctness* dimension; by performing the analysis for each of the age groups, we consider records in the dataset of all kinds of distinctness. Injecting this record ensures the attacker's victim is included in the dataset. We remark that not injecting this record would not change the results of our analysis. Rather, the absence of this record may reduce the accuracy of our results. The sampling process would generate datasets without the victim record that are not relevant in our experiment.

5.2 Disclosure programs (granularity)

In order to test the impact of different anonymisation techniques on the outputs of our risk estimation, we include some of these techniques in our experiments. Namely, we consider *attribute removal* (`attr_r`) and *attribute generalisation* (`attr_g`), described below.

Attribute removal (`attr_r`). The first anonymisation program drops the name column from each record. The resulting dataset has high utility, since all other attributes are left intact. The probabilistic version, `attr_r`, is given in Fig. 10a. It takes a distribution over datasets

as input, and produces a distribution over datasets (with names removed) as output. In Figaro, `Element[T]` is a distribution over `T`. Thus, `ContainerElement[I,T]` is a distribution over containers of `T`, indexed by `I`. Likewise, in `FixedSizeArrayElement[T]`, the index of the container (a fixed-size array) is always `Int`. The original (non-probabilisic) version of `attr_r` is the same as `attr_r`, but with the element types replaces with `List`. It is thus quite easy for a data analyst to obtain `attr_r` from an anonymization program.

Attribute generalisation (`attr_g`). The second anonymisation program, in addition to dropping the name column, furthermore reduces granularity of data. The resulting dataset is less vulnerable to a linking attack, at the cost of some utility. The probabilistic version, `attr_g`, is given in Fig. 10b. We generalise three attributes: zip code, birthday, and age, using (rather simple) generalisation functions `zG`, `bG` and `aG`, respectively, given in Fig. 11. Concretely, zip codes are divided into four bins, birthdays into 12 bins (month), and ages into five bins with approximately twenty years in each.

5.3 Posterior analysis (disclosure queries)

To analyse privacy risks, we consider three types of disclosure, in decreasing order of severity. The first type, *identity disclosure*, estimates the risk that an adversary can identify a record in the anonymised dataset that belongs to a specific data subject—a linking attack. The second type, *attribute disclosure*, considers the risks of an adversary learning the diagnosis of a data subject, without necessarily identifying their record in the dataset—a homogeneity attack. The third type, *quantitative attribute disclosure*, furthermore considers the change of beliefs of the adversary about the probability of a particular data subject being sick based on the released dataset. The disclosures queries are explained in more details below.

Identity disclosure. For quantifying identity disclosure risks, similar to the original Sweeney reidentification, we look at whether the victim is uniquely identified by a given set of attributes—i.e., quasi-identifier analysis. To this end, we infer a distribution over the number of records that share the given set of victim attributes.

Let "$\#(p)$" be the random variable denoting the number of records satisfying predicate p. Predicates are defined on output records, i.e. tuples of the form (z, b, s, a, d). Suppose, now, that we conduct a quasi-identifier analysis for the singleton set of attributes, age. We are then estimating the distribution $P(\#(a = AGE))$, that is, the distribution of the random variable denoting the number of records in the output dataset with the victim's age (AGE).

Attribute disclosure. To quantify attribute disclosure risk, we look at the probability that, in a given dataset, all of the records with the same quasi-identifier as the victim, have the same diagnosis. When all said records have the same diagnosis, we will have learned the victim's diagnosis, despite being (perhaps) unable to determine exactly which row in the input belongs to the victim.

Let $(z_n, b_n, s_n, a_n, d_n)$ denote the nth record in the output dataset. In case of the age quasi-identifier, the probability that we then estimate is $P(\forall n \in (1, N)\ .\ a_n = AGE \implies d_n = \texttt{Ill})$, where N is the total number of records in the output dataset. In other words: the probability that each record in the dataset matching the victim's age, is `ill`. By changing the left-hand side of the implication, we obtain probabilities for other quasi-identifiers.

Quantitative attribute disclosure. Here we measure the attacker's (un)certainty of the diagnosis. We consider how the adversary answers the question "What is the probability of the victim being infected with COVID-19?" given the demographic attributes of the victim as well as the dataset after anonymisation.

More precisely, assuming there is some number k of records sharing the victim's age, we look at the distribution over the number of those records that are `ill`. In other words, the distribution $P(\#(a = AGE \wedge d = Ill) \mid \#(a = AGE) = k)$. For example, $P(\#(a = AGE \wedge d = Ill) = 2 \mid \#(a = AGE) = 10)$ gives the probability of the attacker learning the victim diagnosis with certainty $2/10 = 0.2$, i.e., two out of ten people with the victim's age are ill.

6 DEMONSTRATION: RESULTS

We analyse how each *dimension* affects disclosure risk. To this end, we run each *query* on each *scenario*, and compare the results along each dimension (cf. Fig. 3). For the sake of brevity, for identity disclosure and quantitative attribute disclosure, we opt to show only query result for the age attribute, and, for `attr_g`, to only show results of generalising age. Results of queries and generalisations for other attribute combinations (i.e. those listed in the x-axis of the in lower plots in Fig. 7), can be found in our public repository [21].

6.1 Identity disclosure analysis

For each scenario, we infer a distribution over the number of records that have the same age as the victim. Figure 6 shows the results. The goal is to evaluate how each dimension affects identity disclosure risk—i.e., the extent of which the victim is uniquely identifiable—and determine what age group is more vulnerable to identity disclosure risks. To this end, we first compare how changes along a dimension affect the distribution (i.e., its mode and variance).

Informedness. Informing the prior increases the mode for ages 0-70, and decreases the mode for ages ≥ 80. This is specially pronounced for ages ≥ 80 when using `attr_g` for anonymisation. The variance of the distributions *decreases* for people in the age groups ≥ 80, meaning that most of the probability is concentrated at the mode.

Granularity. Generalising attributes changes the support of the distributions: for `attr_r` is ca. $(0, 25)$ whereas for `attr_g` is ca. $(50, 175)$ (except for the COVID19 prior). This change also increases the variance of the distributions. Generalising attributes *increases* the mode of the distributions, by a factor of ~20 for age groups 0-70 and up to a factor of 30 for ≥ 80. These results indicate that increasing granularity notably *decreases* identity disclosure risks.

Distinctness. Increasing distinctness does not significantly modify the mode of the distributions of age groups 0-70, but notably *decreases* the mode for ≥ 80. The variance of the distributions is not affected for age groups 0-70, but *decreases* for ≥ 80 and COVID19 prior. Note that, for COVID19 prior, `attr_r` anonymisation and the age group ≥ 80, the probability is concentrated in low values (concretely values < 2); indicating a high risk of re-identification.

Findings. From this analysis, we present the following findings.

For age groups 0-70, UNIFORM constitutes a worst-case analysis for identity disclosure. This is because the mode of UNIFORM is lower than that of COVID19 in Fig. 6; fewer people share the victim's age

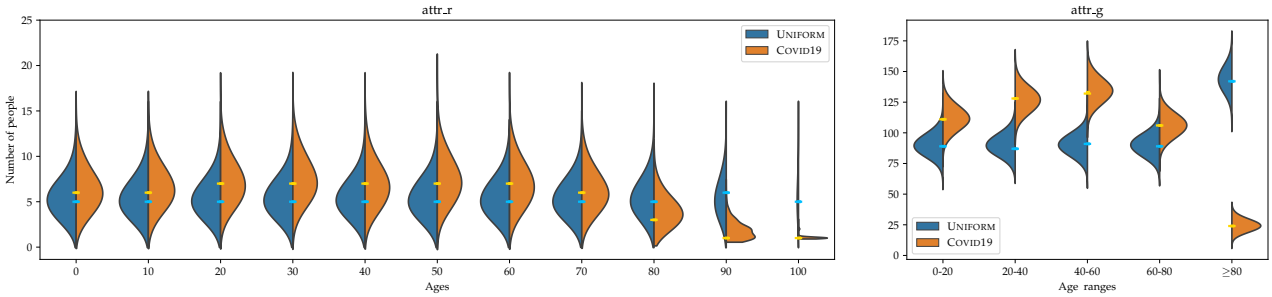


Figure 6: Identity disclosure results. Probability that n people are in the same age group as the victim. The left side of the violin plots shows the distributions for the UNIFORM prior (blue), and the right side the distributions for the COVID19 prior (orange). The markers show the mode of the distributions for UNIFORM (light blue) and COVID19 (light orange) priors.

in UNIFORM, making the victim more identifiable (i.e. higher risk) with UNIFORM. Further, the probability that the victim is uniquely identifiable is (near) zero, indicating that age groups 0-70 are not at risk of being identified based on their age. In Fig. 6, the UNIFORM peak is 5, meaning $5/500 = 1/100$ people match the age of the victim (slightly fewer than what Fig. 5 says about the Danish population).

For age groups ≥ 80, however, the *opposite* holds; using UNIFORM leads to a profound underestimation of risk. This is because the mode of COVID19 is lower than the mode of UNIFORM. In the last column of `attr_g` in Fig. 6, the COVID19 mode at ca. 24 is much smaller than the UNIFORM mode at ca. 145, and the COVID19 mode for in the other age groups. Worse, in `attr_r` and COVID19 prior, the age groups ≥ 90 are very likely unique (> 50% probability). This means that people in these age groups are at a high risk of being identified based on their age when data is anonymised with `attr_r`.

6.2 Attribute disclosure analysis

For each scenario, we infer the probability that every record, sharing a given set of attributes with the victim, is ill. Figure 7 shows the results. We consider combinations of age, birthday, zip code, and sex, shown on the x-axis on the bottom plots. For granularity, we consider removal of name ("attr_r(name)" in Fig. 7), generalisation of age ("attr_g(age)"), and generalisation of age, zip and birthday ("attr_g(age,zip,day)"). The goal is to evaluate how each dimension affects attribute (i.e. diagnosis) disclosure risk—the extent of which the victim's diagnosis can be inferred with certainty by the attacker—and determine what age group is more vulnerable to attribute disclosure. We do this by first comparing how changes along a dimension affect the inferred probabilities.

Queries. Singleton attribute sets (age-row) pose a near-zero disclosure risk for age groups 0-79. However, for age groups ≥ 80 and COVID19 prior, we see an exponential growth in disclosure risk, with age groups ≥ 90 having over 90% risk of an attacker learning the victims' diagnosis. Increasing the attribute set increases the disclosure risk. This is because the larger the attribute set, the fewer records there will be that match those attributes with the victim; the victim is more likely to be uniquely identified by the attribute combination. Adding zip or birthday greatly increases the risk (birthday slightly more), whereas adding sex barely increases it. For instance, in the case of `attr_r`, for attribute set age, zip and sex, disclosure risk is ~100% irrespective of distinctness and informedness. The same holds for age, zip, and birthday, for `attr_r` and `attr_g(age)`.

Informedness. Informing the prior, for age groups 0-79, *decreases* the probability for each query and each granularity. Informing the prior, for age groups ≥ 80, *increases* the probability for each query and each granularity. There are several cases where this increase is especially profound: i) when using `attr_r` and attribute sets age and age & sex, and ii) when we generalise age, zip & day for attribute sets age, zip & day and age, zip, day & sex.

Granularity. Generalising attributes *decreases* the probability for all scenarios, except for the attribute sets {age, zip, day} and {age, zip, day, sex} for which the probability remains the same when comparing "attr_r" and "attr_g" plots. For age groups 0-79, when we generalise age, risk for attribute set age & zip decreases from $\langle 98\%, 98\% \rangle$ (UNIFORM prior, COVID19 prior) to $\langle 71\%, 60\% \rangle$. For age groups ≥ 80, however, the risk decreases from $\langle 98\%, 98\% \rangle$ to $\langle 58\%, 89\% \rangle$. Note the small decrease for COVID19. Disclosure risk does decrease dramatically if we furthermore generalise birthday and zip. For age groups 0-79, the highest disclosure risk for COVID19 is 45% and 52% for UNIFORM. For age groups ≥ 80, the risks are greatly reduced; however, the highest disclosure risk for COVID19 is still 75%.

Distinctness. Increasing distinctness *decreases* mildly the risks for age groups ≥ 80 for UNIFORM prior. However, doing so *increases* the risk for COVID19 for most queries (except of course for queries that already had ~100% probability). For instance, for attribute set {age, zip, day, sex} (last column), for generalisation of age, zip & birthday ("attr_g(age,zip,day)"), for UNIFORM, the risk decreases from 52% to 40%, whereas for COVID19, the risk doubles from 40% to 80%.

Findings. From this analysis, we present the following findings.

We observe the same effect of distinctness on the UNIFORM prior, that we found during identity disclosure analysis. For age groups 0-79, UNIFORM is a worst-case analysis for attribute disclosure analysis. This is seen by comparing plots vertically in Fig. 7; values in bottom plots are less than those in the top plots. For age group ≥ 80, the *opposite* holds; UNIFORM leads to an underestimation of risk.

Furthermore, for the group ≥ 80 and COVID19, `attr_g` is ineffective. However, for age groups 0-79, `attr_g` offers low but better privacy protection (ca. 40% attribute disclosure risk). Our results show the importance of generalising these attributes, but also indicate that these generalisation mechanisms may be insufficient.

With these results, data controllers make better, informed decisions on disclosing data, or can inform people in different age groups of risk of disclosing their data (and seek their consent).

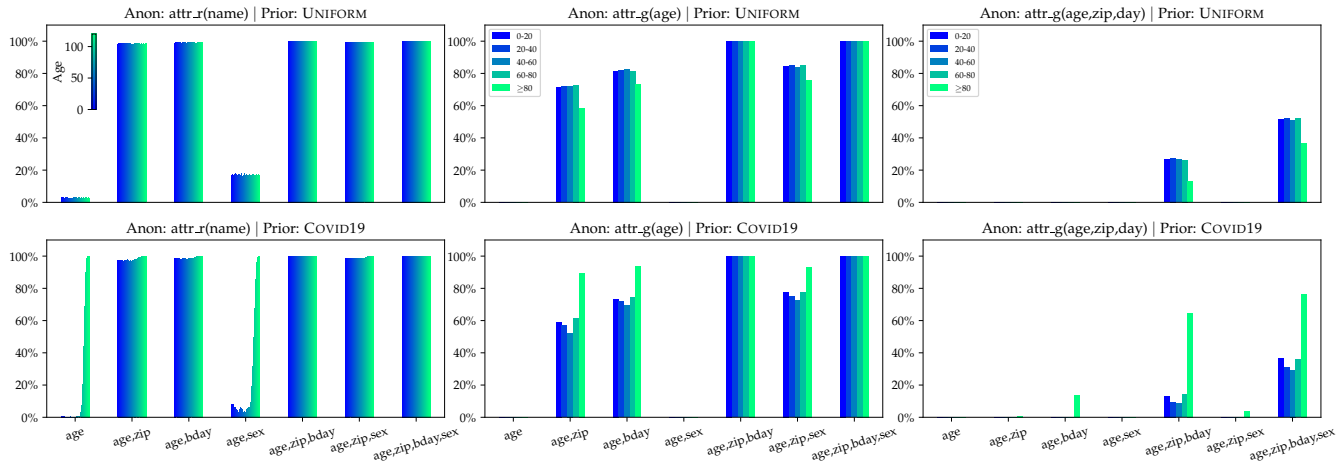
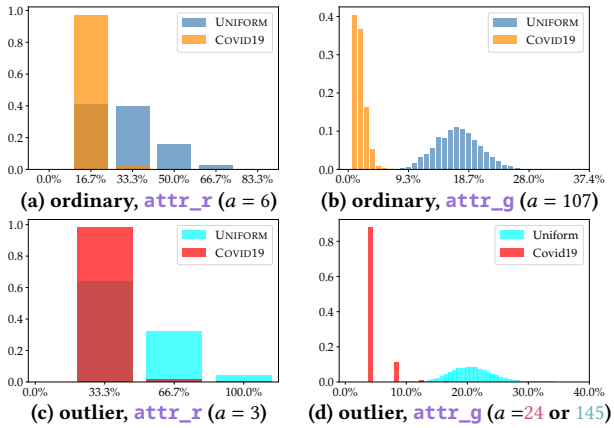


Figure 7: Attribute disclosure results. Probability that every record sharing the victim's attribute values (x-axis) is ill. Each column corresponds to an anonymisation program, and each row corresponds to a prior. Bars are color coded, from dark blue (young age groups) to light green (old age groups).



(a) ordinary, `attr_r` ($a = 6$) **(b) ordinary, `attr_g` ($a = 107$)**

(c) outlier, `attr_r` ($a = 3$) **(d) outlier, `attr_g` (a =24 or 145)**

Figure 8: Quantitative attr. discl. results. Probability that n/a of people in the same age group as the victim are ill.

6.3 Quantitative attribute disclosure analysis

For each scenario, we infer a distribution, over the number of those records that are ill, with the condition (observation) that a records have the same age as the victim. Figure 8 shows the result. The goal is to evaluate how each dimension affects quantitative attribute disclosure risk, i.e. the extent of which the attacker learns something about the probability that the victim is ill (i.e., how *certain* the attacker is that the victim is ill). Further, we compare how probabilities differ before and after the observation, by comparing what changes along a dimension does to the attacker certainty on whether the victim is ill.

Choice of a. For several distributions in Fig. 8, we picked a to be a value that is roughly equally probable for UNIFORM and COVID19 in the corresponding plot in Fig. 6 and has probability ≥ 0 in both distributions. Note that there is no value a meeting these conditions across all age groups. Consequently, we have selected two illustrative type of records: i) a record with $age = 32$ which we refer to as *ordinary* (people in this age are well represented in the Danish population, cf. Fig. 5a), and ii) a record with $age = 90$ which we refer to as *outlier* (people in this age are underrepresented in the Danish population). Furthermore, since there exists no equiprobable value for the age group ≥ 80 (outlier) when generalising age (last column in Fig. 7), we pick the most likely outcome for each prior. As a result, we cannot draw conclusions from informing the prior in Fig. 8d, as the histograms are not comparable along that dimension. These choices are for the sake of presentation, they are not a limitation of our framework. Distributions for any choice of a and age group can be analyzed using the program in our public repository [21]. In Fig. 8, the values of the x axis represent the portion of the a records that share the victim's age, that are ill. In Fig. 8a, 33% represents $0.33 * 6 = 2$, i.e. the probability that 2 out of the 6 records are ill.

Certainty increase. We compute how much the certainty of the attacker (of victim being ill) increases after making an observation.

First, what was the certainty prior to observation? For UNIFORM, prior certainty is 20%. For COVID19, it is 0.7% for ordinary, and 0.77% for outlier (cf. column 30-39 resp. 90-200 in Fig. 5b).

Next, what is the certainty after observation? We see this by looking at the mode of each of the eight histograms in Fig. 8. In Fig. 8a in COVID19, the mode is 17% (i.e. it is most probable that $0.17*6 = 1$ out of $a = 6$ are ill). Thus, after observation, 17% becomes the new certainty of whether the victim is ill in that scenario.

Finally, we compute the ratio of the two, to discover how much the certainty has increased (Table 1). For the `attr_g`-UNIFORM scenarios, certainty does not increase; it remains at 20%. In all other scenarios, certainty increases, by varying amounts. For instance, in the ordinary-`attr_r`-UNIFORM scenario, certainty increases slightly (by a factor of 1.65), whereas in the outlier-`attr_r`-COVID19 scenario, certainty increases enormously (by a factor of 42.86).

We investigate how changes along dimensions affects the attacker's certainty. We do this by comparing increases in Table 1. We compare certainties after observation across scenarios, and we compare how much certainty increased upon making the observation. Since the values of a differ in the plots, we are unable to draw conclusions from comparisons along the distinctness dimension.

distinctness	granularity	informedness	prior	mode	increase	ratio
ordinary	attr_r	UNIFORM	20.00%	33%	13.00%	1.65
ordinary	attr_r	COVID19	0.70%	17%	16.30%	24.29
ordinary	attr_g	UNIFORM	20.00%	20%	0.00 %	1.00
ordinary	attr_g	COVID19	0.70%	1%	0.30 %	1.43
outlier	attr_r	UNIFORM	20.00%	33%	13.00%	1.65
outlier	attr_r	COVID19	0.77%	33%	32.23%	42.86
outlier	attr_g	UNIFORM	20.00%	20%	0.00%	1.00
outlier	attr_g	COVID19	0.77%	4%	3.23%	5.19

Table 1: Increase in illness certainty after observation.

Informedness. Informing the prior *decreases* the attacker's certainty. The certainty decreases more for ordinary than for outlier. This can be seen by comparing mode values of rows 1 & 2, 3 & 4, etc.; for ordinary-attr_r, certainty decreases from 33% to 17%, whereas for outlier-attr_r, certainty does not decrease (33% cf. 33%).

If we instead consider *how much* certainty increases, we get a different picture. For the informed prior, the attacker's certainty *increases proportionally more.* This increase is profound for attr_r, but slight for attr_g. This can be seen by comparing increases (or ratios) of 1 & 2, 3 & 4, etc.; for outlier-attr_r, the certainty increase is 32.23% (a profound factor 42.86 increase) cf. 13.00% (a meager factor 1.65 increase), whereas for outlier-attr_g, the certainty increase is 3.23% (a factor 5.19 increase) cf. 0.00% (no increase). The increase is not significantly different for outlier compared to ordinary. The largest such difference is ordinary-attr_r compared to outlier-attr_r; in the former, certainty increases from 13.00% (factor 1.65) to 16.3% (factor 24.29), whereas in the latter, certainty increases from 13.00% (factor 1.65) to 32.23% (factor 42.86).

Granularity. Generalising attributes reduces the attacker's certainty. For COVID19, this reduction is profound, whereas for UNIFORM, it is slight. This can be seen by comparing mode values of rows 1 & 3, 2 & 4, etc.; for ordinary-COVID19, certainty reduces from 17% to 1%, whereas for ordinary-UNIFORM, certainty reduces from 33% to 20%.

If we instead consider *how much* certainty increases, we get a similar picture. For generalised attributes, the attacker's certainty *decreases proportionally more.* This decrease is profound for UNIFORM, but slight for COVID19. This can be seen by comparing increases (or ratios) of 1 & 3, 2 & 4, etc.; for outlier-COVID19, the certainty increase is 3.23% (a factor 5.19 increase) cf. 32.23% (a profound factor 42.86 increase), whereas for outlier-UNIFORM, the certainty increase is 0.00% (no increase) cf. 13.00% (factor 1.65 increase).

Findings. From this analysis, we present the following findings.

Informing the prior decreases the attacker's certainty. However, for the informed prior, the attacker learns proportionally more than for the uninformed prior. This nuance is crucial; in some cases, a data controller cares about the attacker's prediction of the victim's illness, whereas in other cases, a data controller may care about how much information the attacker learns.

Generalising attributes reduces the attacker's certainty. Likewise, for generalised attributes, the attacker learns proportionally less than for removed attributes. For UNIFORM, generalising attributes reduces what the attacker learns to 0, whereas for COVID19, doing so reduces what the attacker learns from a lot to a little.

7 RELATED WORK

In [18], Pardo *et. al.* introduced PRIVUG and evaluated its accuracy, scalability and applicability for a wide range programs (e.g. differential privacy [7] and k-anonymity [26]). Our main novelty cf. [18] is the *conceptual framework.* In PRIVUG, the result of analysis depends on the prior. While expressive and fine-grained, great care must be exercised when choosing a prior during risk analysis (it must model a realistic attacker). Our conceptual framework exists to help the data controller navigate the problem space—the data, the attackers, and the mechanisms—to then organise approaches for solving the (risk assessment) problem. While motivated by [18], our framework is *independent* of PRIVUG (see Sect. 1). Finally, no previous work evaluates how attacker knowledge affects privacy risk in our level of detail (see findings Sect. 6); [18] only evaluates informedness (not distinctness & granularity). Though [18] introduced Sweeney's experiment using PRIVUG (the authors study attribute removal and k-anonymity on a uniform prior), our paper presents significant novelties. First, we consider UNIFORM and COVID19 priors. As we have shown, COVID19 has revealed insights not detected in UNIFORM. Second, we have extended the structure of the dataset to include the age attribute—this was required as the attribute is present in the COVID-19 dataset. Finally, we study attribute generalisation.

Other tools and methods to analyse re-identification risks hav been proposed in the last decade, see [20] and references therein. Most of these tools can be used to perform all (or a subset of) the analyses we study here (cf. Sect. 5.3). The main difference between these tools and the experiment in this paper is that these tools evaluate re-identification risks on concrete datasets, as opposed to re-identification risks on anonymisation algorithms. We remark also that PRIVUG, can be used in the absence of a concrete dataset (e.g., our uniform prior), which makes it possible to analyse anonymisation algorithms before real accurate data is available. This is especially important in our experiment, as accurate COVID-19 data was available only a few months after the pandemic started.

In [23], Rocher *et. al.* use a probabilistic model to evaluate re-identification in anonymised datasets (removing identifiers as in attr_r) that contain only a fraction of the records of the complete dataset. Since datasets are incomplete, the authors use publicly available demographic data to assess re-identification risks (identity disclosure in our paper). The goal of the model is to estimate the risk of identity disclosure given a set of demographic attributes. Instead, we focus on evaluating different anonymisation programs (attr_r and attr_g); [23] does not support this. We have further studied two more privacy metrics: attribute disclosure and quantitative attribute disclosure. However, our attacker model is more limited; we assess privacy risks on a pre-defined record (the victim), and assume that the attacker knows that the victim is included in the dataset. The attacker model in [23] extrapolates to any member to the population, and makes no assumptions about the presence of the victim's record in the dataset.

8 CONCLUSION

In this paper, we presented a multi-dimensional conceptual framework for assessing privacy risks in the presence prior knowledge about data, as well as a systematic method of conducting the analysis using PRIVUG. We have demonstrated the effect of attacker

knowledge on different types of privacy risks and on different anonymisation programs, in a health data privacy setting. The demonstration illustrates the process that data controllers may follow to perform privacy risk analysis on anonymisation programs, and refine their results by using publicly available data. We have focused on two anonymisation programs working on COVID-19 infection data: direct identifier removal and attribute generalisation. We have studied three common types of privacy risks: identity disclosure, attribute disclosure, and quantitative attribute disclosure. We have studied two priors (attacker knowledge). A UNIFORM prior modelling an attacker with no side-knowledge, and a COVID19 prior which includes side-knowledge about COVID-19 infections and demographics of Danish citizens. We have identified the most vulnerable age groups in the Danish population.

Our framework makes it easier for data controllers to explore privacy risks in health care systems (or similar), thus enabling them to make informed decisions when anonoymising data.

Acknowledgement. Work partially supported by the Danish Villum Foundation through Villum Experiment project No. 00023028 'Assessment of Reidentification Risks with Bayesian Probabilistic Programing'

REFERENCES

[1] T. Chothia, Y. Kawamoto, and C. Novakovic. Leakwatch: Estimating information leakage from java programs. In *European Symposium on Research in Computer Security*, pages 219–236. Springer, 2014.

[2] C. Culnane, B. I. Rubinstein, and V. Teague. Health data in an open world. *arXiv preprint arXiv:1712.05627*, 2017.

[3] D. Culnane, A. Rubinstein, I. Benjamin, A. Teague, et al. Stop the open data bus, we want to get off. *arXiv preprint arXiv:1908.05004*, 2019.

[4] Y.-A. De Montjoye, L. Radaelli, V. K. Singh, et al. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.

[5] Denmark Age Demographics. https://www.statbank.dk/, retrieved 12.02.2021.

[6] Denmark COVID-19 Infections. https://www.statista.com/statistics/1102237/coronavirus-cases-development-in-denmark/, retrieved 06.10.2020.

[7] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[8] S. Garfinkel, J. M. Abowd, and C. Martindale. Understanding database reconstruction attacks on public data. *Commun. ACM*, 62(3):46–53, 2019. ISSN 0001-0782.

[9] P. Golle. Revisiting the uniqueness of simple demographics in the us population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, pages 77–80, 2006.

[10] V. Hargitai, I. Shklovski, and A. Wasowski. Going beyond obscurity: Organizational approaches to data anonymization. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

[11] A. J. Hendriks, W. K. Hofstee, and B. De Raad. The five-factor personality inventory (ffpi). *Personality and individual differences*, 27(2):307–325, 1999.

[12] G. Kraaykamp and K. Van Eijck. Personality, media preferences, and cultural participation. *Personality and individual differences*, 38(7):1675–1688, 2005.

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

[14] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):1–52, 2007.

[15] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[16] G. J. Matthews and O. Harel. Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5:1–29, 2011.

[17] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[18] R. Pardo, W. Rafnsson, C. Probst, and A. Wasowski. Privug: Using probabilistic programming for quantifying leakage in privacy risk analysis. In *Proceedings of 26th European Symposium on Research in Computer Security (ESORICS'21)*, pages 417–438. Springer International Publishing, 2021. ISBN 978-3-030-88428-4.

[19] A. Pfeffer. *Practical probabilistic programming*. Manning Publications Co., 2016.

[20] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn. Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*, 2020.

[21] Public Repository with: Figaro model, Scala programs, experiments, and additional results. https://bitbucket.org/itu-square/privug-covid-19-danish-citizens/.

[22] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, 2004. ISBN 978-1-4419-1939-7.

[23] L. Rocher, J. M. Hendrickx, and Y.-A. De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9, 2019.

[24] D. J. Solove. *Understanding privacy*. Harvard University Press, May, 2008.

[25] State of California. Assembly Bill No. 375 California Consumer Privacy Act (CCPA), 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375, last accessed on 19.02.2021.

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

[27] The European Parliament and Council of European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj, last accessed on 10.11.2020.

[28] The United States Congress. Health Insurance Portability and Accountability Act of 1996 (HIPAA), 1996. https://www.govinfo.gov/link/plaw/104/public/191, last accessed on 19.02.2022.

[29] V. Torra. *Data privacy: Foundations, new developments and the big data challenge*. Springer, 2017.

[30] I. Wagner and D. Eckhoff. Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):1–38, 2018.

A DEMONSTRATION: IMPLEMENTATION

```
VariableSizeArray( size, i =>
 for {
  n <- Uniform(names:_*)
  z <- Uniform(zips:_*)
  b <- Uniform(0, 364)
  s <- Uniform(Male, Female)
  a <- Uniform(0,112)
  d <- If(Flip(0.2),
          Ill, Healthy)
 } yield (n, z, b, s, a, d))
```

(a) UNIFORM

```
VariableSizeArray( size, i =>
 for {
  n <- Uniform(names:_*)
  z <- Uniform(zips:_*)
  b <- Uniform(0, 364)
  s <- If(Flip(0.49),
          Male, Female) // DK
  a <- distAge(s)       // DK
  d <- distIll(s,a)     // DK
 } yield (n, z, b, s, a, d))
```

(b) COVID19

Figure 9: Priors (attacker knowledge)

```
def attr_r (
 rs: FixedSizeArrayElement[(Name,Zip,BDay,Sex,Age,Diagnosis)]
 ) : ContainerElement[Int,(Zip,BDay,Sex,Age,Diagnosis)]
 = rs.map { case (n,z,b,s,a,d) => (z,b,s,a,d) }
```

(a) Attribute Removal

```
def attr_g (
 rs: FixedSizeArrayElement[(Name,Zip,BDay,Sex,Age,Diagnosis)]
 ) : ContainerElement[Int,(Zip,BDay,Sex,Age,Diagnosis)]
 = rs.map { case (n,z,b,s,a,d) => (zG(z),bG(b),s,aG(a),d) }
```

(b) Attribute Generalisation

Figure 10: Anonymisation Programs

```
def zG (z: Zip ) : Zip  = ( (z - 1000) min 3000) / 1000
def bG (b: BDay) : BDay = (b min 330) / 30
def aG (a: Age ) : Age  = (a min 80) / 20
```

Figure 11: Auxiliary functions of `attr_g`