Testing the robustness of anonymization techniques: acceptable versus unacceptable inferences - Draft Version

Gergely Acs, Claude Castelluccia, Daniel Le Métayer

1 Introduction

Anonymization is a critical issue because data protection regulations such as the European Directive 95/46/EC and the European General Data Protection Regulation (GDPR) explicitly exclude from their scope "anonymous information" and "personal data rendered anonymous"[1]. However, turning this general statement into effective criteria is not an easy task. In order to facilitate the implementation of this provision, the Working Party 29 (WP29) has published in April 2014 an *Opinion on Anonymization Techniques*[2]. This Opinion puts forward three criteria corresponding to three risks called respectively "singling out", "linkability" and "inference". In this paper, we first discuss these criteria and suggest that they are neither necessary nor effective to decide upon the robustness of an anonymization algorithm (Section 2). Then we propose an alternative approach relying on the notions of acceptable versus unacceptable inferences (Section 3) and we introduce *differential testing*, a practical way to implement this approach using machine learning techniques (Section 4). The last section discusses related work and suggests avenues for future research (Section 5).

2 Analysis of the criteria of the Working Party 29

The WP29 recommends that data controllers consider three risks to assess the robustness of their anonymization algorithm:

- *Singling out*, which is the "possibility to isolate some or all records which identify an individual in the dataset".
- *Linkability*, which is the "ability to link at least two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases)".
- *Inference*, which is the "possibility to deduce, with significant probability, the value of an attribute from the values of other attributes".

As discussed in the introduction, the goal of the WP29 was to provide an interpretation of the Directive 95/46/EC with regard to anonymization and to facilitate its implementation. The main issue to be addressed is therefore the relevance and usefulness of the criteria for this purpose. A first observation concerning the legal sources is that both the Directive and the GDPR heavily rely on the notion of "identifiability". A personal data is defined in the GDPR as "any information related to an identified or identifiable natural person" and a dataset is considered as properly anonymized if "the data subject is no longer identifiable". Taking a broader perspective, singling out, linking and inference can actually be seen as three different ways of deriving new information (or attributes in the database terminology) about individuals.

First, we want to stress that the WP29 criteria are neither necessary conditions nor effective means to assess anonymization algorithms. They are not necessary because they do not take into

[1] Recital 26 of the European General Data Protection Regulation.

[2] Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques, adopted 10 April 2014.

account the type of information that can be derived. In some cases, this information may actually be insignificant, noisy or even useless. As an illustration, the RAPPOR technology ensures that individual data are randomized in such a way that they keep a global utility without jeopardizing individual privacy [8]. Even though individual data do not pass the "singling out" and "linkability" criteria, it can be shown that RAPPOR offers a high level of privacy (in particular, it provides strong deniability and ε-differential privacy guarantees).

As far as effectiveness is concerned, it depends very much on the precise meaning of "inference" in the third criterion. If inference is taken in a very general sense, considering any type of attribute and inference technique, then it can be argued that this criterion is so powerful that it encompasses all possible ways to identify individuals (in the sense of associating and individual with an attribute). However, this interpretation would also be meaningless as it would accept only datasets without any utility at all. Indeed, the ultimate usefulness of a dataset is always to infer new information, for example by discovering new links or correlations between attributes. Therefore, a dataset passing the third test in this strong sense would necessarily be useless. The only way to make this criterion meaningful would be to qualify it and consider inferences of attributes about "specific" individuals with "sufficient" accuracy. But then we face the threshold issue: where should the red line be put to decide upon "specific" and "sufficient". For example, inferring an attribute about the population of a city, or a rule like "a man smoking between 1 and 4 cigarettes per day is 3 times more likely to die from lung cancer than a non-smoker" should clearly be acceptable. However, deriving attributes about the inhabitants of a building may or may not be acceptable depending on the size of the building. More generally, deriving information about a person should be considered in a different way when this information also applies to a population as a whole[3] (e.g. population statistics). Therefore, inference as such is not a clear-cut criterion. It should involve other fine tuned parameters which are not necessarily objective (i.e. they may result from political or collective decisions about what is considered as acceptable or not in society).

The second point that we would like to emphasize is that the concept of data re-identification of anonymized datasets is misleading: attribute inference should be the primary concern. El Emam and Alvarez [6] argue that, instead of attribute inference, identity disclosure should be mitigated. Their argument is based on the observation that automatically preventing attribute inferences usually lead to useless datasets. According to them, it is not the inference per se, but rather its *usage* that can be "discriminatory, creepy, surprising, or stigmatizing". They recommend that a privacy ethics council should advise whether the resulting "anonymized" dataset can be safely released or not and under which conditions.

Although we agree that the release of anonymized datasets should be reviewed and controlled by a review board, we believe that the situation is a bit more complex. First, we argue that while the mitigation of "identity disclosure" is the primary goal of pseudo-anonymisation schemes, it is not relevant for data anonymisation schemes. Indeed, for an adversary, identity disclosure is the assignment of a correct identity to an anonymized record. However, if a dataset is correctly anonymized, its records are very likely to be highly noised or aggregated (as opposed to pseudo-anonymisation schemes where the identifiers are just removed, and the data are published without transformation). Therefore, re-identification becomes pointless, as the adversary would then potentially re-identify some noisy records that would certainly be useless.

We should avoid the fetichization of the notion of identity and rather see identity disclosure as one way among others to infer information about individuals (by identifying or singling out their records). Even though in some places the wording of the Directive 95/46/EC and the GDPR seem to focus on identity disclosure, they are not entirely consistent to this respect[4] and we believe

[3]This comment is in line with the distinction between "personal information" and "private information" made by Franck McSherry [1]. Private information is seen as "secrets that you can keep by withholding your data" whereas "personal data" could be derived by inference from datasets in which you are not necessarily involved.

[4]For example, Recital 26 stresses the fact that "personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person", which entails that protection against identity disclosure is not sufficient to make a dataset anonymous.

that the WP29 is right in taking a more general interpretation. However, as discussed above, we are not convinced that its three criteria can be really effective in practice. Considering that inference is the key issue, we believe that anonymized datasets should be assessed by the yardstick of inference techniques, in order to assess both

- privacy risks, which can be seen as the risks of inferring privacy intrusive information and
- benefits, which can be seen as the possibility of inferring useful (and legitimate) information.

The heart of the matter is therefore a risk-benefit analysis relying on a precise study of the inferences that can be drawn from the dataset followed by decisions about where the line should be put between acceptable and unacceptable inferences. In the next sections, we successively refine this notion of acceptable and unacceptable inferences (Section 3) and propose an approach to implement it using machine learning techniques (Section 4).

3 Beyond de-identification: private versus public inferences

As argued in the previous section, the ability to perform inferences is the key issue with respect to both privacy and utility. It is a well-known fact, however that automatically preventing the inference of attribute values is difficult and often leads to useless or at least very noisy or aggregated datasets. Hence, we are facing the following dilemma: how can anonymized datasets be useful and still protect against undesirable inferences?

To solve this dilemma, we consider the following notion of privacy in this paper. The intuition is that if an adversary cannot prove that the record of a user was used to generate the anonymized dataset, then by definition this record is "protected". In other words, a dataset is deemed "anonymized" according to our model if it can be shown that, for any user, the resulting inferences based on this dataset do not depend on the user's contribution (or record) but on the contribution of other users (which may be correlated): the inference accuracy and certainty should be about the same whether the user's record is included or not in the dataset. This model therefore protects against "private" inferences while still allowing "group/public" inferences.

It might happen that the properties of the population can be used to build a model that can be applied to individuals with high accuracy [2]. However, we do not consider this to be a privacy breach as long as the group/population size is large enough. Instead, as in [7], we believe that there are acceptable and unacceptable disclosures: "learning statistics about a large population of individuals is acceptable, but learning how an individual differs from the population is a privacy breach".

We acknowledge that, because of their nature, certain "group" inferences can still be harmful, which means that the release of the resulting anonymized dataset should still be reviewed and controlled by a privacy ethics committee. Generally speaking, the decisions to release a dataset should always be part of a rigorous privacy risk analysis, which systematically identifies the risks and the potential benefits of publishing the datasets [4].

To summarize, we argue that the challenge is to provide criteria to distinguish between acceptable and unacceptable inferences. We believe that acceptability can be based on two criteria:

1. The *basis* of the inference: is the inference performed on the basis of the records of one (or a small group of) individual(s) or on the basis of the records of a large group of individuals, i.e. a "population"?
2. The *nature* of the inference: can the inference be used to discriminate users? Can it have a very negative (for example social or financial) impact?

The second criterion is partly subjective and involves ethical and legal considerations. In this paper, we focus on the first criterion and introduce, in the next section, a scheme called differential testing to assess the basis of the inference.

The main idea of our scheme is to use Machine Learning to predict the sensitive attribute of users (attributes that are usually not quasi-identifiers but rather represent some information

not to be revealed about the user such as medical diagnosis, salary, locations, etc.). We assume we have a dataset composed of several users and its anonymized version (and each user has a sensitive attribute). For each user, we infer his sensitive attribute value using the anonymized dataset as input to our machine learning algorithm. We then remove the user record from the original dataset and generate another anonymized dataset. We again infer the user's sensitive attribute value using the new anonymized dataset. In both cases, the output of the inference is a distribution on all possible values of the sensitive attribute, i.e., the probability that the user's sensitive attribute has a particular value in the original dataset according to the machine learning algorithm. If these distributions are similar, we consider that the inference was based on a group since the inclusion of the user's record does not impact the prediction. On the other hand, if the distributions are different, we conclude that the inference is "individual" since a user's record can have a substantial impact on the prediction distribution[5].

4 Differential testing: a machine learning based process

We first introduce some technical assumptions (Section 4.1), then describe (adversarial) inference of attribute values more precisely (Section 4.2). We also provide some intuition about the differential testing procedure and present its more formal definition along with a discussion (Section 4.3).

4.1 Assumptions

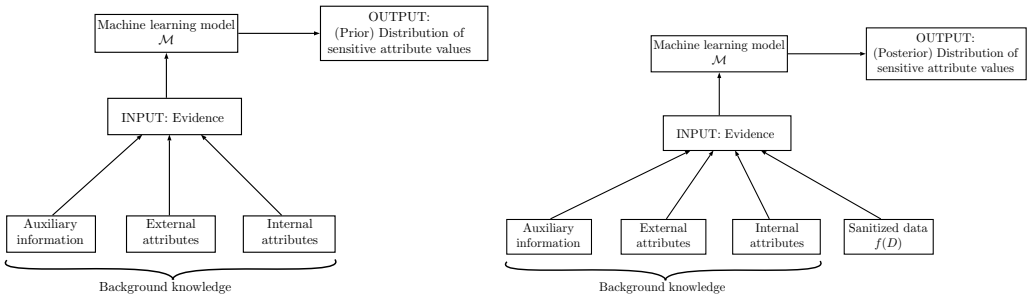
In general, the goal of an adversary is to learn new information about an individual or group of individuals by combining his (prior) background knowledge about the targets and the anonymized (or sanitized) dataset $f(D)$, where f is the sanitization (anonymization) mechanism. We assume that the target individuals are always part of D, which has microdata format (i.e., each row of D contains a set of attributes of a single individual). We distinguish between the attributes in D, called *internal attributes*, and the *external attributes* which represent any information about an individual that is not explicitly inside D. An external attribute may strongly correlate with any internal attribute. The background knowledge of the adversary can come from internal or external attributes of the target (or other individuals in D) and any other auxiliary information which represent *common knowledge*, such as public statistics of sensitive attributes. We also assume that f's output has the same microdata format as its input.

4.2 Towards Differential Testing

The learning of any attribute is modeled as a statistical inference process executed by the adversary and illustrated in Figure 1. It takes any internal/external attributes (i.e., background knowledge) of its targets and combines them with $f(D)$ in a statistical or machine learning model in order to infer further attributes. The output of the machine learning model is a distribution on the domain of the attribute to be inferred. For example, the adversary can train a Naive Bayes classifier to predict the nationality of a person. If there are 196 countries worldwide, the model will output a probability value assigned to each possible country, which represents the certainty of the adversary that the person is from a particular country. More precisely, the adversary has already some confidence in each possible attribute value a priori when it has no access to $f(D)$ (Figure 1a), and it aims to increase this confidence by using $f(D)$ as an extra evidence in the inference process (Figure 1b). In general, any inference or learning is characterized by the difference between the prior and posterior distributions of the inferred attribute values.

However, as discussed in the previous sections, the above inference is the core of any learning process to derive useful information about the dataset, which is the ultimate goal of the release. In fact, we should rather focus on inferences that can differentiate a single individual *from the rest of the dataset*. These inferences are potentially privacy-invasive, as they represent private

[5]The choice of the ML algorithm(s) is important. We will address in future work what should be its properties.



(a) Inference without sanitized data (b) Inference with sanitized data

Figure 1: Adversarial inference. $\mathcal{M}$ can be any machine learning model such as Naive Bayes, decision trees, neural networks, etc. In general terms, absolute privacy is defined as the difference between the prior and posterior distributions of the inferred attribute.

information which is specific to the individual, that is, not shared with anybody else in the dataset. The following definition translates this intuition in more technical terms.

Definition 1 (Differentiator learning model) Let D_i^- denote a dataset obtained from D by removing individual $i \in D$ from D. Let $\mathcal{M}$ be a machine learning model which induces a probability distribution $\mathcal{M}_s$ on the domain of attribute s. $\mathcal{M}$ is a δ-differentiator of attribute s in D, if

- *Average-case differentiation:* $(1/|D|) \sum_{i \in D} \mathsf{sim}(\mathcal{M}_s(f(D)), \mathcal{M}_s(f(D_i^-))) \leq \sigma$
- *Worst-case differentiation:* $\max_{i \in D} \mathsf{sim}(\mathcal{M}_s(f(D)), \mathcal{M}_s(f(D_i^-))) \leq \sigma$

where sim is a similarity measure on the output distribution $\mathcal{M}_s$.

Intuitively, a differentiator inference represents the individual's own secret which can *only* be revealed if the individual participates in the dataset. Therefore, such secret should not be inferred from an anonymized (or sanitized) dataset. To verify whether a data release is sanitized prudently, one should measure the differentiation of all inferences with respect to all attributes in the dataset, which is infeasible. Computing the differentiation of state-of-the-art learning models (e.g., using a standard machine learning library[6]) or design new models which are believed to have the largest differentiation may be sufficient to assess privacy in practice.

In Definition 1, we deliberately did not define the exact similarity measure between the distribution on the attribute domain. One can use any similarity metric for this purpose such as KL-divergence, Earth-Mover Distance, or the maximum divergence used by differential privacy [5].

For simplicity, we omitted the background knowledge from the formulation of the differentiator $\mathcal{M}$ in Definition 1. In practice, we assume that the differentiator uses all internal attribute values of the targeted individual except the sensitive attribute to be inferred. However, it should not be constrained to use only internal attributes and potentially be provided with additional auxiliary information such as public statistics about the sensitive attribute.

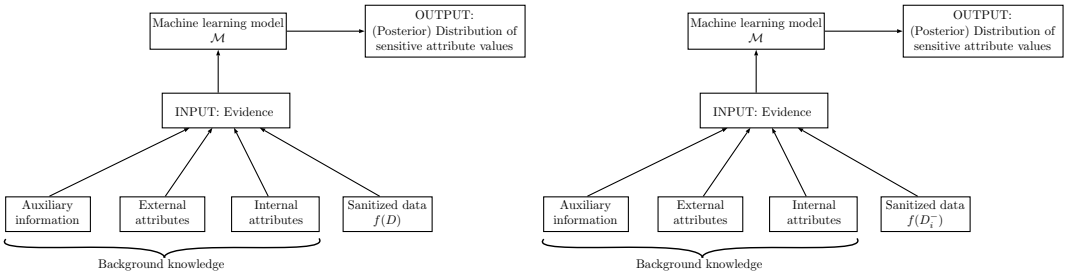
Our approach is illustrated in Figure 2, and can be be summarized as follows:

1. For each record, r_i, of the original dataset D:
   - Step 1:
     - Anonymise the dataset to obtain the anonymized dataset $f(D)$
     - Predict from $f(D)$ the sensitive attribute (s) distribution using a machine learning algorithm.
   - Step2:

[6] `http://scikit-learn.org/`

- Remove r_i from the D and anonymize it to obtain $f(D_i^-)$.
- Predict from $f(D_i^-)$ the sensitive attribute distribution (s') using the same machine learning algorithm as in Step1.

2. Testing step: If the two distributions s and s' are similar, according to a similarity measure to be defined, then the test is successful since the prediction does not depend on the actual record of the target data subject but rather based on other users.



(a) Inference with all sanitized records

(b) Inference without record i

Figure 2: δ-differentiator model: posterior distributions on D and D_i^- differ with at most δ

5 Conclusion

Our proposal is related to the empirical privacy model introduced in [3]. In fact, the empirical privacy model also proposes to test whether the sensitive attribute(s) can be predicted from the released dataset. However, it does not make the distinction between acceptable and unacceptable inferences. It considers all types of inferences as privacy breaches. More precisely, the scheme tries to predict, for each entry, the sensitive attribute(s) using, for example, machine learning techniques, and checks whether the obtained predicted value is actually "similar" to the actual value(s) of the original entry. If the predicted value is close to the actual value (what they call "empirical utility"), the anonymisation scheme does not pass the test. Note that the scheme does not consider whether the prediction was obtained from the records of the tested user (that was somehow poorly anonymized) or from the records of other users (that happen to be correlated with the tested user). In contract, our proposal does not consider data utility (i.e. does not check whether the inferences are correct), but instead propose to modify this testing procedure to make the distinction between private and public inferences. Our goal is to define a testing procedure that only prevents private predictions and, as a result, provides better data utility.

Note that our scheme is also related to the differential privacy model [5]. In fact, our scheme guarantees that, similarly to differential privacy, the predicted sensitive attributes are similar whether that user was part of the anonymised dataset or not. However, our differentiation test provides strictly weaker guarantee than differential privacy. First, we compare the posterior inferences on datasets which can differ only in records included in the original dataset. That is, we do not provide any privacy guarantee to individuals who are *not* in the dataset. For example, one might learn from the sanitized dataset that a user was not part of the dataset. Second, we do not guarantee an upper bound on the differentiation of *all* possible inferences unlike differential privacy. Finally, differential privacy is a property of the sanitization scheme, while our differentiation property holds for a particular inference model, dataset and sensitive attribute. Still, we believe that our differentiation test can be more practical when micro-data is released and differential privacy provides weak utility. In addition, if a differentiator inference model is easily interpretable (such as decision trees), more insights may be provided about why privacy can be violated in practice. That is, our approach may provide a more accessible assessment of privacy than ε in differential privacy.

Our work is still very preliminary and some (important) issues are still open. For example, the choice of the ML algorithm(s) is important. In future work, we will address what should be its properties. Furthermore, the distribution similarity function plays an important role in our scheme. It should be selected and analyzed carefully. We do not consider that there is a privacy breach as long as the group/population size is large enough (i.e. the probability of the inferences is not overwhelmingly large) but this size issue is not fully studied in this paper. This will be addressed in future work.

References

[1] Lunchtime for Data Privacy, Accessed 30-10-2016. `https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md`.

[2] G. Cormode. Personal privacy vs population privacy: Learning to attack anonymization. In *Proceedings of the KDD*, 2010.

[3] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical privacy and empirical utility of anonymized data. In *Workshops Proceedings of the ICDE, 2013*, pages 77–82, 2013.

[4] S. J. De and D. Le Métayer. *Privacy Risk Analysis*. Morgan&Claypool, 2016.

[5] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.

[6] K. E. Emam and C. Alvarez. A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques. 5(1):73–87, 2015.

[7] A. Machanavajjhala and D. Kifer. Designing statistical privacy for your data. *Commun. ACM*, 58(3):58–67, Feb. 2015.

[8] lfar Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 21st ACM Conference on Computer and Communications Security*, Scottsdale, Arizona, 2014.