Measuring and predicting anonymity
Koot, M.R.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Practical applications

In this Chapter we will share preliminary ideas on applying in real life the techniques developed in this thesis. We provide a conceptual framework for the application of the distribution-informed prediction of anonymity properties via Kullback-Leibler distances (KL-distances) as developed in Chapter 4 and Chapter 5. The KL-based predictions can be applied to quasi-identifiers consisting of any combination of numerical variables (e.g. \{ age + height \}) and categorical variables (e.g. \{ gender \}).

In addition, we discuss application of the techniques developed in Chapter 5 and Chapter 6, which only apply to numerical variables, such as the analysis of the effect of interval-widths on identifiability (see Chapter 6). The latter enables pollsters, for example, to protect the anonymity of respondents by deciding beforehand, based on quantifications, whether to ask respondents to reveal, say, their exact age or rather the age group to which they belong — instead of collecting exact ages and having respondents trust their unknown pollster that she will make the data less precise afterwards. Quantifications remove some of the pollster’s uncertainty that may otherwise have led the pollster to choose an overly wide interval (while it may be beneficial for the analysis to have more specific information), or perhaps to simply ignore the issue and ask for exact data that puts the respondent at risk (or at least, leave them with a feeling of unease).

The remainder of this Chapter is organized as follows: Section 7.1 will in-
introduce our preliminary model; Section 7.2 will discuss ‘non-functional’ aspects crucial to real-life application of the model; Section 7.3 will describe steps to take toward implementing a real-life application; Sections 7.4 and 7.5 will discuss various practical aspects that need to be taken into account, including the limitations of our work; and Section 7.6 will conclude this Chapter. For further inspiration we refer to the example analysis of anonymity in Appendix B, that considers a questionnaire observed in real life.

Remark 7.1 Measuring unidentifiability is measuring identifiability. Our techniques are intended for privacy protection but can be used directly for purposes of identifiability as well, such as in marketing and forensics. Our perspective, however, is that of privacy protection.

7.1 Preliminary model

We now introduce a preliminary conceptual framework for applying distribution-informed prediction techniques. Figure 7.1 shows the framework, distinguishing a repository (that stores Kullback-Leibler distances), policy (decisions about what data (not) to disclose, collect and share), data holder(s) (anyone with access to personal data) and policy maker(s) (anyone deciding about the processing of personal data, notably including the subjects themselves). We distinguish four tasks, chronologically ordered: publish, query, analyze and decide. These will be explained below.
7.1. PRELIMINARY MODEL

Figure 7.1: Preliminary model for applying distribution-informed privacy predictions as part of privacy policy making.
7.1.1 Data holder

The data holder collects and stores personal information about individuals. Although entities that are legally assigned the role data controller or data processor can act as data holder, not only they can. For example, an individual can be data holder of his/her own personal data. To prevent confusion with the legal domain we use the label “data holder”, consistent with Solove [74] (see Section 1.1). In our model, anyone having access to a collection of personal data can act as data holder.

The disclosure of information by a person to a data holder establishes a context conform Nissenbaum [61], including context-relative informational norms, compliancy to which constitutes contextual integrity; i.e., that the disclosed information does not end up in a situation where presence of it constitutes a privacy violation (as perceived by data subjects or wider society, but not necessarily made explicit in laws; also, note that translating implicit, subjective and changing contextual roles into a disclosure policy is non-trivial). The publication of a statistic about a population of which that person is part is unlikely to violate privacy law. In the Netherlands, for example, privacy law only applies to the processing of data that can be traced to individuals without considerable effort. The practical meaning of ‘considerable effort’ remains unclear to us; from a privacy perspective we hope it means no less than ‘disproportionate to potential gain’. Such publication might, however, violate a context-relative informational norm, for example when the person does not agree with their data being part of a openly published statistic (such as in our model). Additional work is needed to assess the moral and legal risk in openly publishing statistics computed from existing collections of personal data.

Task:

- publish: submit to repository one or more Kullback-Leibler distances, accompanied by specification of the QID and population.

7.1.2 Policy maker

Policy maker assesses privacy risk and decides what data (not) to collect and what data (not) to share. The decision is influenced by legal norms and, if data holder and policy maker are the same entity, the context-relative informational norms between data holder and the persons about whom the data holder stores data. As a special case, policy maker can be a self-assessing individual that wants to decide what (combined) information not to disclose during, for example, an anonymous questionnaire.
7.1. PRELIMINARY MODEL

Task:

- **query**: request from repository the Kullback-Leibler distance \((KL-distance)\), given a specification of the QID and population;

- **analyze**: apply our methodology to analyze QIDs;

- **decide**: decide what data (not) to collect and what data (not) to share.

Presumably, these tasks will be part of a more comprehensive privacy risk management process that takes into account existing information collection and sharing that might influence the privacy risk involved in the per-instance context to which our methodology is most relevant.

7.1.3 Repository

The repository is a publication facility that allows the data holder to submit KL-distances, and the policy maker to query KL-distances. The repository stores three-tuples consisting of a description of the QID, a description of the population and the KL-distance. While KL-distance is a number, the description of QID and population will be less trivial. For clarity of exposition, we will assume that policy maker and data holder use the same ontologies and data structures; i.e., they have a shared vocabulary. Under that assumption, the QID and population can be defined in terms of that vocabulary. Consider a shared ontological concept and data structure \(Citizen\) that contains, among others, the attributes \(PostalCode\), \(Gender\), and \(BirthYear\); second, that population can be specified in terms of \(City\); and third, that data holder stores this data for all citizens of Amsterdam. To publish the KL-distance that applies to citizens of Amsterdam regarding \(QID = \{PostalCode, Gender, BirthYear\}\), data holder first computes the KL-distance and sends to the repository the following message:

\[
\begin{align*}
\text{QID} &= \{PostalCode, Gender, BirthYear\} \\
\text{population} &= \{City=Amsterdam\} \\
\text{KL-distance} &= \ldots
\end{align*}
\]

Data holder may also publish KL-distances for less specific QIDs. For example, leaving out \(Gender\):

\[
\begin{align*}
\text{QID} &= \{PostalCode, BirthYear\} \\
\text{population} &= \{City=Amsterdam\} \\
\text{KL-distance} &= \ldots
\end{align*}
\]

In addition, data holder may also publish KL-distances for subpopulations. For example, only including persons that own a car:
QID = {PostalCode, BirthYear}  
population = {City=Amsterdam, CarOwner=yes}  
KL-distance = ...  

The use and applicability of various QIDs and various subpopulations will depend on the (existence of) information collection and sharing to which the involved persons are exposed.

7.2 Issues

7.2.1 Assess privacy risk of KL-repository itself

The purpose of our model is privacy protection, but possibly, the model poses privacy risk itself. At this point, we do not know if or how a public repository of KL-distances might be abused. As we noted, measuring unidentifiability is equivalent to measuring identifiability, and our techniques might be applied for purposes of identifiability rather than unidentifiability. We expect that smallness of populations, as determined by the amount of information in the QID, will be a key issue in deciding what KL-distances (not) to publish. A tradeoff exists between accuracy of prediction (more specific population = more accurate prediction) and protecting against use for identifiability (less specific population = less usable for identifiability).

7.2.2 Disputes

In a perfect world, there are no errors in the data from which KL-distances are computed, and the data covers complete populations. In real life, errors do occur and coverage is often incomplete. The mileage will vary between different data holders. The Dutch municipal registry offices, for example, will tend to cover the complete population within their municipality, while a corporate dataset might only cover the consumers within that population; and not only within the municipality where they are located, but consumers from any location.

In the occasion of multiple data holders submitting different KL-distances for the same QID and population, a decision must be made which information to use. We consider this beyond the scope of our thesis.

7.2.3 Incentives

For policy makers, the legal obligation to comply with privacy law may be incentive to publish KL-distances: e.g., if the policy maker wants to legally avoid collecting personal data, then, under Dutch privacy law, the data collected must not be traceable to individuals without effort that is disproportionate to the risk associated with such disclosure. The policy maker may be motivated
7.3. WHAT STEPS TO TAKE NEXT

7.3.1 Make an inventory of data holders and their data

An inventory is needed of data holders and their data. Specifically, for each relevant data set, a list of columns and description of the population about which data is present need to be established. In the Netherlands, a potential starting point is the Dutch Data Protection Agency (CBP), that maintains a registry of data protection officers and a registry of (reported) processing of personal data. Other pointers for the Dutch can be found in the report ‘Onze digitale schaduw’ (2009) [70] commissioned by the Dutch Data Protection Agency and in the report ‘iOverheid’ (2011) established by the Dutch Scientific Council for Government Policy (WRR) [27]. For the UK, pointers can be found in the report ‘Database State’ (2009) [3] commissioned by the Joseph Rowntree Reform Trust.

Then, a second inventory is needed: a list of the (combined) information that pollsters ask for during anonymous questionnaires (whether online or offline), and the information that is shared in contexts of science and policy research. In the Netherlands, both Statistics Netherlands and the KNAW/NWO Data Archiving and Networked Services (DANS) institute\(^1\) may provide pointers. This second inventory can be jump-started through a simple brainstorm process.

By matching both inventories, and taking into account privacy risks of the KL-repository itself (see Section 7.2.1) and the desired scope of the repository, it needs to be decided which QIDs and populations to include/exclude. Theoretically, the scope of the repository could be unlimited: one could attempt to establish a single nation-wide or even global repository that contains KL-distances for every possible QID and every possible population. Practically, the scope is limited to data holders and policy makers that (are able to) share a data vocabulary (see Section 7.1.3) and are also willing to participate. It is

\(^1\)Website: http://www.dans.knaw.nl/
probably sensible to limit a first attempt at a repository to common QIDs and common populations; both of which can be established via the inventories and common sense.

7.3.2 Build software tools

Information technology will need to be built. First, a repository needs to be set up. Second, software for KL-analysis and publication to the repository needs to be engineered and distributed to data holders (tools for computing KL-distances from data stored in MySQL, MSSQL, etc.). Third, software for performing distribution-informed analysis is needed (possibly local, possibly remote). Fourth, the system needs to be maintained, and policy maker and data holder should be able to get support when needed.

7.4 Other aspects

Apart from above considerations, several other aspects need to be taken into account. Two factors that play an important role in quasi-identifier analysis are the granularity, or interval width, of variables in a quasi-identifier; and the correlation between variables in a quasi-identifier. Regarding the former, obviously the more fine-grained the data is, the more identifying a quasi-identifier will tend to be. The methodology of Section 6.3 can be applied to quantify this effect. Regarding the latter, the technique developed in Section 6.4 can be used to predict identifiability in case there is substantial correlation between multiple non-categorical, numerical variables within the quasi-identifier. In principle the same KL-based technique can be used as in the single-variate case, as long as the correlation between the variables is taken care of adequately as demonstrated in Section 6.4.

Lastly, the repository needs to be protected from misinformation (e.g. unintentionally incorrect KL-distances being submitted) and disinformation (intentionally incorrect KL-distances, e.g. to make privacy risk appear less than it really is). We consider this to be outside the scope of our thesis, but emphasize that it must be addressed when working toward a real-life implementation.

7.5 Limitations and future work

Our distribution-informed prediction techniques require that Kullback-Leibler distances can be computed between the Uniform distribution and the actual distribution. To know the actual distribution, access is required to (personal) data. That data must be representative, in terms of the quasi-identifier variables under consideration, for the population for which the quasi-identifier analysis is performed. As mentioned in Section 7.2.2, the data ideally provides full
7.6. CONCLUSION

coverage of that population, and has few errors. If it is not possible to get access to that data, or the data contains too many errors in the variables present in the quasi-identifier, the distribution-informed techniques cannot be applied. Also, note that while an individual may use these techniques to get an on-the-average estimate of identifiability of members of the population to which he/she belongs, that individual may self have outlier values and be more identifiable than the estimate suggests. In a population where nearly everyone has blue eyes or brown eyes, disclosing that one has green eyes is obviously more revealing than on the average.

Furthermore, in the analysis of singletons outlined in Chapter 5, notably Figure 5.1, it is observed that our approximations become inaccurate in presence of strong outliers. In our example of age distributions, our approximations showed accurate for the range 0-79; we were unable to obtain accurate results when including ages above 79. Clearly, this implies that our methodology is insufficient by itself for performing exhaustive privacy analysis. We propose that other methods and techniques that do sufficiently take outliers into account are applied together with ours.

In Chapter 6, note that the $O(\delta)$ approximation techniques for determining the effects of interval width, as Figure 6.3 shows for height, width and birthday, become less accurate when predicting outcomes for large interval widths. In addition, the techniques we proposed for taking into account the effects of correlation between variables have only been examined for the setting where the variables have a bivariate normal distribution. These aspects must be considered when applying these techniques in practice.

7.6 Conclusion

Although distribution-informed prediction is not the only method we developed throughout Chapter 4, 5 and 6, it is our most innovative result. In this Chapter, we primarily focused on that result and share preliminary ideas about how to apply it in practice. Additional work is needed: first, the privacy risk of a public repository of Kullback-Leibler distances computed from sets of personal data needs to be assessed. Second, inventory needs to be made of (candidate) data holders, and of the QIDs and populations that are most relevant to be subjected to privacy-analysis. We provided pointers to information sources that we believe are useful during these activities.