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.

Download date: 19 Dec 2018
Abstract (English)

In our increasingly computer-networked world, more and more personal data is collected, linked and shared. This raises questions about privacy — i.e. about the feeling and reality of enjoying a private life in terms of being able to exercise control over the disclosure of information about oneself. In attempt to provide privacy, databases containing personal data are sometimes de-identified, meaning that obvious identifiers such as Social Security Numbers, names, addresses and phone numbers are removed. In microdata, where each record maps to a single individual, de-identification might however leave columns that, combined, can be used to re-identify the de-identified data. Such combinations of columns are commonly referred to as Quasi-IDentifiers (QIDs).

Sweeney’s model of $k$-anonymity addresses this problem by requiring that each QID value, i.e., a combination of values of multiple columns, present in a data set must occur at least $k$ times in that data set, asserting that each record in that set maps to at least $k$ individuals, hence making records and individuals unlinkable. Many extensions have been proposed to $k$-anonymity, but always address the situation in which data has already been collected and must be de-identified afterwards. The question remains: can we predict what information will turn out to be identifiable, so that we may decide what (not) to collect beforehand?

To build a case we first inquired into the (re-)identifiability of hospital intake data and welfare fraud data about Dutch citizens, using large amounts of data collected from municipal registry offices. We show the large differences in (empirical) privacy, depending on where a person lives. Next, we develop
a range of novel techniques to predict aspects of anonymity, building on proba-
ibilistic theory, and specifically birthday-problem theory and large-deviations
theory.

Anonymity can be quantified as the probability that each member of a group
can be uniquely identified using a QID. Estimating this uniqueness probability
is straightforward when all possible values of a quasi-identifier are equally likely,
i.e., when the underlying variable distribution is homogenous. We present an
approach to estimate anonymity for the more realistic case where the variables
composing a QID follow a non-uniform distribution. We present an efficient
and accurate approximation of the uniqueness probability using the group size
and a measure of heterogeneity called the Kullback-Leibler distance. The ap-
proach is thoroughly validated by comparing the approximation with results
from a simulation using the real demographic information we collected in the
Netherlands.

We further describe novel techniques for characterizing the number of sin-
gletons, i.e., the number of persons have 1-anonymity and are unambiguously
(re-)identifiable, in the setting of the generalized birthday problem. That is,
the birthday problem in which the birthdays are non-uniformly distributed
over the year. Approximations for the mean and variance are presented that
explicitly indicate the impact of the heterogeneity, expressed in terms of the
Kullback-Leibler distance with respect to the homogeneous distribution. An
iterative scheme is presented for determining the distribution of the number of
singletons. Here, our formulas are experimentally validated using demographic
data that is publicly available (allowing our results to be replicated/reproduced
by others).

Next, we study in detail three specific issues in singletons analysis. First, we
assess the effect on identifiability of non-uniformity of the possible outcomes.
Suppose one has the ages of the members of the group; what is the effect on
the identifiability that some ages occur more frequently than others? Again, it
turns out that the non-uniformity can be captured well by a single number, the
Kullback-Leibler distance, and that the formulas we propose for approxima-
tion produce accurate results. Second, we analyze the effect of the granularity
chosen in a series of experiments. Clearly, revealing age in months rather than
years will result in a higher identifiability. We present a technique to quantify
this effect, explicitly in terms of interval. Third, we study the effect of corre-
lation between the quantities revealed by the individuals; the leading example
is height and weight, which are positively correlated. For the approximation
of the identifiability level we present an explicit formula, that incorporates the
correlation coefficient. We experimentally validate our formulae using publicly
available data and, in one case, using the non-public data we collected in the
early phase of our study.

Lastly, we give preliminary ideas for applying our techniques in real life.
We hope these are suitable and useful input to the privacy debate; practical
application will depend on competence and willingness of data holders and policy makers to correctly identify quasi-identifiers. In the end, it remains a matter of policy what value of $k$ can be considered *sufficiently strong* anonymity for particular personal information.