Efficient probabilistic estimation of quasi-identifier uniqueness

Koot, M.R.; Mandjes, M.; van 't Noordende, G.; de Laat, C.

Publication date
2011
Document Version
Final published version
Published in
Proceedings of ICT.OPEN 2011: 14-15 November 2011, Veldhoven, The Netherlands

Efficient Probabilistic Estimation of Quasi-Identifier Uniqueness

Matthijs R. Koot
Informatics Institute
Science Park 904
Amsterdam, Netherlands
koot@uva.nl

Michel Mandjes
Korteweg-de Vries Institute for Mathematics
Science Park 904
Amsterdam, Netherlands
m.r.h.mandjes@uva.nl

Guido van 't Noordende
Informatics Institute
Science Park 904
Amsterdam, Netherlands
noordende@uva.nl

Cees de Laat
Informatics Institute
Science Park 904
Amsterdam, Netherlands
delaat@uva.nl

ABSTRACT

This paper proposes a method to quantify anonymity. Anonymity can be quantified as the probability that each member of a group can be uniquely identified using a *quasi-identifier*: a combination of variables which combined can be used to identify an individual within a group. 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. In this paper, we present an approach to estimate anonymity for the more realistic case where the variables composing a quasi-identifier follow a non-uniform distribution. We present an efficient and accurate approximation of the uniqueness probability using a measure of heterogeneity called the Kullback-Leibler distance and the group size. The approach is thoroughly validated by comparing the approximation with results from a simulation using real demographic information from the Netherlands.

Categories and Subject Descriptors

K.4 [**Computer and Society**]: Public Policy Issues Privacy

General Terms

re-identification, data anonymity, probability, birthday problem, Kullback-Leibler distance

1. INTRODUCTION

Large amounts of personal data is collected and stored nowadays. Some of it is intended for policy research on, for example, finance, health and public administration. In that scenario it is common, for reasons of privacy protection, to de-identify the data before disclosing it to the researchers. The de-identified data often still contains personal attributes such as age, location and gender. Combinations of those attributes might sometimes allow re-identification of the anonymized records. A decade ago, Latanya Sweeney established that 87% of the US population was uniquely identifiable by a quasi-identifier (QID) composed of three demographic variables [20, 21]: date of birth, gender and 5-digit ZIP code. To improve privacy she proposed the k-anonymity model, where a mandatory rule is applied to a table before it is disclosed [22]. A table is said to be k-anonymous if each of the QID values in that table occur at least k times. If the table does not satisfy the rule, the attributes comprising the QID are generalized or eliminated from the table, up until the rule is satisfied. This effectively ensures unlinkability of records and individuals, by ensuring each record can be associated with at least k different individuals.

In his short paper revisiting Sweeney's work, Philippe Golle mentions a lack of available details about the data collection and analysis involved Sweeney's work as a reason for being unable to explain the big difference between the outcome between both studies: in Golle's study of the 2000 U.S. Census data, only ∼63% of U.S. citizens turned out to be uniquely identifiable, as opposed to ∼87% that Sweeney determined by studying the 1990 U.S. Census data [7]. This may be attributed to inaccuracies in the source data. By using registry office data we are confident that our results (for the Dutch population) are likely to be highly accurate.

In an earlier study we analyzed quasi-identifiers in two data sets containing information about hospital intakes and welfare fraud [11]. The quasi-identifier in the hospital intake data set consisted of 4-digit postal code, gender, month of birth and year of birth, and in the welfare fraud data set it contained the municipality rather than the 4-digit postal code. The objective of the study was to assess the level of anonymity enjoyed by persons present in the data sets. The results were roughly comparable to the results obtained by Sweeney in the U.S. For example, 67.0% of the sampled population turned out identifiable by date of birth and four-

digit postal code alone, and 99.4% by date of birth, full postal code and gender.

One of the common challenges in k-anonymity and its developments is the recognition of quasi-identifiers. The method we develop in this paper provides a new way of efficiently estimating the likeliness that a given set of attributes will function as a perfect quasi-identifier, i.e,. that each value of a quasi-identifiers unambiguously identifies an individual. That quantification may be useful in privacy impact assessments and policy research.

Usually, QIDs are addressed -after- data has been collected, and each data collection deals with QIDs for itself. In our scenario, a data collector (perhaps Statistics Netherlands) collects data and publishes a single number representing the heterogeneity of the QID distribution over the records in his table. That number, the *Kullback-Leibler distance* that will be introduced shortly, represents the distribution skew in the prior data collections. Using that number, our method allows future data collectors to predict properties of QIDs - before- collecting data – and possibly use that information to decide on what (not) to collect and possibly to decide what the impact of combining earlier-collected data may have on privacy.

For QIDs consisting of attributes that don't chance too frequently, such as ZIP code and date of birth, the method introduced in this paper provides an efficient approximation of the probability that every (QID) value in a group of people unambiguously identifies an individual. An entity such as Statistics Netherlands, who have access to enormous amounts of data, may publish precomputed tables that data collectors may use to include or exclude specific pieces of information in their planned data collection.

As a follow up to [11] and related papers of a fully empirical nature, the primary question the current paper addresses is: *'Can we develop a methodology to determine the probability that all persons in a group can be uniquely identified by quasi-identifier X?* This can be used as a measure of anonymity.The main contribution of our work is that we provide a sound technique to accurately approximate this probability. The idea is to translate our question in terms of a birthday problem, and then to rely on probabilistic techniques.

The main problem is that, unlike in the classical birthday problem [17], the probability distribution for many variables and thus for many QID's is non-uniform, i.e., not all possible values occur with equal frequency. This heterogeneity is dealt with by adjusting the outcome of the homogeneous birthday problem (in which all outcomes *are* equally likely) by a measure of heterogeneity, the *Kullback-Leibler distance* [12]. As mentioned, the techniques used are of a probabilistic nature; more specifically, we borrow elements from *large-deviations theory* [5, 16].

It is emphasized that the stated question is of interest both to adversary ('which quasi-identifiers should I want?') and the anonymous subject ('which quasi-identifiers should I avoid?'). Our method will be demonstrated using demographic data from the Netherlands, but the approach can be applied to *any* population.

This paper is organized as follows. In Section 2 we formally describe the problem in terms of a birthday problem with unequal probabilities. Section 3 presents an approximation for the uniqueness probability under heterogeneity, where the deviation from the uniform situation is captured by the Kullback-Leibler distance. In Section 4 we validate the approximation, and use the approximation to perform a number of experiments. In Section 5 we describe related work - not any more. Then the paper is concluded, in Section 6, by a discussion and outlook.

2. PROBLEM

The problems we come across in this paper can be regarded as generalized birthday problems. In the 'classical' birthday problem [6, 23] there are k individuals, each of whom is assigned (uniformly, independently) a value from the set $\{1, \ldots, N\}$. It is a straightforward exercise in probability theory to check that the probability that all values ('birthdays') are unique is given by

$$\pi_{\mathrm{u}}(k, N) = \frac{N}{N}\frac{N-1}{N}\cdots\frac{N-k+1}{N} = \frac{N!}{(N-k)!N^k}.$$

However, things complicate tremendously in case the outcomes $\{1, \ldots, N\}$ are *not* equally likely. To study this situation, suppose that F_i outcomes have probability α_i/N, for $i = 1, \ldots, d$ (that is, there are d groups, within which the probabilities are again uniform). Here it is assumed that $F_1 + \ldots F_d = N$ (each outcome is a member of one group) and $F_1\alpha_1 + \ldots F_d\alpha_d = N$ (the total probability is 1). For this generalized birthday problem, it is not possible to write down a clean expression for the uniqueness probability (although it can be evaluated numerically in quite an efficient way [10]). However, as we will show in this paper, we succeeded in developing an accurate approximation. This approximation is based on the Kullback-Leibler distance, which is a measure for heterogeneity within the population. It turns out that the more heterogeneous the population is, the lower the uniqueness probability. In addition, it is shown that assuming all outcomes are equally likely (so that the above explicit formula can be applied) leads to quite substantial estimation errors.

To simplify the exposition, we use a very simple quasi-identifier in our examples: age. We experimentally assessed the quality of our approximation using real data about the Dutch population: the distribution of age in all Dutch municipalities, which vary in size (1k–750k citizens). In contrast to a previous study [11], the data we use here is publicly available from Statistics Netherlands for others to reproduce our results[1].

3. METHODOLOGICAL FRAMEWORK: BIRTHDAY PROBLEMS

As mentioned above, the uniqueness probability can be calculated straightforward in case all outcomes are equally likely. In this section we present an approximation for the situation where this is *not* the case, that is, the situation in which probabilities of the outcomes $1, \ldots, N$ differ from $1/N$.

[1]Statistics Netherlands, StatLine: http://statline.cbs.nl

3.1 Approximations for the general birthday problem

In this subsection we describe a way to find an approximation for the uniqueness probability in the non-uniform scenario. The approximation relies heavily on the idea of 'Poissonization'.

Approximations for the uniform case. We briefly describe a classical approximation for the uniform case (i.e., $d = 1$), and show that this approximation is exact in a particular asymptotic regime. To this end, observe that

$$\begin{aligned}\pi_{\mathrm{u}}(k,N) &= \exp\left(\sum_{i=0}^{k-1}\log\left(1-\frac{i}{N}\right)\right)\\ &\approx \exp\left(-\frac{1}{N}\sum_{i=0}^{k-1} i\right) \approx \exp\left(-\frac{k^2}{2N}\right). \quad (1)\end{aligned}$$

This approximation can be formally justified if k scales like $\sqrt{N}$: applying 'Stirling',

$$\begin{aligned}\pi_{\mathrm{u}}(a\sqrt{N},N) &= \frac{N!}{(N-k)!N^k}\\ &\sim e^{-a\sqrt{N}}\left(1-\frac{a}{\sqrt{N}}\right)^{N-a\sqrt{N}} \to e^{-\frac{a^2}{2}}, \quad (2)\end{aligned}$$

where the convergence is due to Lemma 1.(i) (see appendix A). Plugging in $a := k/\sqrt{N}$ indeed gives approximation (1).

Poissonization for the uniform case. We show that assuming that k is not given but drawn from a Poisson distribution with mean k yields, remarkably enough, the same asymptotic (2).

To this end, suppose that the sample size is Poisson distributed with mean k. An elementary conditioning argument yields that this gives the uniqueness probability

$$\pi_{\mathrm{Pois,\,u}}(k,N) = \sum_{i=0}^{N} e^{-k}\frac{k^i}{i!}\frac{N!}{(N-i)!N^i} = e^{-k}\left(1+\frac{k}{N}\right)^N.$$

As before an approximation of the type $\exp(-k^2/(2N))$ can be justified, because

$$\pi_{\mathrm{Pois,\,u}}(a\sqrt{N},N) = e^{-a\sqrt{N}}\left(1+\frac{a}{\sqrt{N}}\right)^N \to e^{-\frac{a^2}{2}},$$

applying Lemma 1.(ii). In other words, even though we randomize the number of samples, we obtain the same approximation.

The non-uniform case. We now consider the situation where F_i (for $i = 1,\ldots,d$) of the outcomes have probability α_i/N, with $F_1 + \ldots F_d = N$ and $F_1\alpha_1 + \ldots F_d\alpha_d = N$. As argued earlier, if the α_i are not uniform, then computing the uniqueness probability $\pi(k,N)$ is not straightforward. The idea of Poissonization does ease this task considerably, though, as we will show.

It is first observed that, when sampling k times according to the mechanism described above, the number of these samples that are from group i (with $i = 1,\ldots,d$) has a multinomial distribution with parameters k and (probability vector) $(\alpha_1F_1/N,\ldots,\alpha_dF_d/N)'$. Suppose instead the number of samples from group i is Poisson distributed with mean $(\alpha_iF_i/N)\cdot k$ (rather than the described multinomial distribution). Then the uniqueness probability essentially reduces to the product of the uniqueness probabilities *within each of the d groups* (use independence!). Therefore, in self-evident notation,

$$\begin{aligned}\pi_{\mathrm{Pois}}(k,N) &= \prod_{i=1}^{d}\pi_{\mathrm{Pois,\,u}}\left(\alpha_iF_i\cdot\frac{k}{N},F_i\right)\\ &\approx \exp\left(-\frac{k^2}{2N^2}\sum_{i=1}^{d}\alpha_i^2F_i\right), \quad (3)\end{aligned}$$

and then the idea is to approximate $\pi(k,N)$ by $\pi_{\mathrm{Pois}}(k,N)$, as we did in the uniform case. In [2, Thm. 4] this approximation was made precise, in the sense that, with $f_i := F_i/N$ being the fraction of all individuals that is of type i, as $N\to\infty$,

$$\pi(a\sqrt{N},N) \to \exp\left(-\frac{a^2}{2}\sum_{i=1}^{d}\alpha_i^2f_i\right).$$

3.2 Impact of non-uniformity

A, perhaps naïve, idea could be to ignore the heterogeneity and to simply use the 'homogeneous formula' (1). In this subsection we show that such an approach could lead to highly inaccurate estimates — evidently, the more heterogeneous the population is, the less accurate such an approximation. To study this effect, we further asses the impact non-uniformity has on the uniqueness probability.

Uniform distribution maximizes uniqueness probability. The approximation of the uniqueness probability for the non-uniform case is majorized by the approximation for the uniform case. This can be see as follows. First observe that we need to prove that $\sum_{i=1}^d\alpha_i^2f_i \geq 1$, given that $\sum_{i=1}^d f_i = \sum_{i=1}^d\alpha_if_i = 1$ (where it is noted that the minimum value 1 is attained when all α_i coincide). Let the random variable A have the value α_i with probability f_i. As variances are non-negative, we evidently have

$$\sum_{i=1}^{d}\alpha_i^2f_i = \mathbb{E}A^2 \geq (\mathbb{E}A)^2 = 1,$$

which proves our claim. The fact that the uniform distribution actually *maximizes* the uniqueness probability has been observed before, cf. [9, 19]. In more concrete terms, it means that all perturbations from the uniform distribution *reduce the uniqueness probability.*

Distances between distributions. Observing that

$$\frac{\exp(-\frac{a^2}{2})}{\exp(-\frac{a^2}{2}\sum_{i=1}^d\alpha_i^2f_i)} = \exp\left(\frac{a^2}{2}\sum_{i=1}^{d}(\alpha_i^2f_i-1)\right),$$

we conclude that

$$\frac{1}{2}\sum_{i=1}^{d}(\alpha_i^2 f_i - 1)$$

is a measure for discrepancy between the uniform distribution and the non-uniform distribution under consideration. There are several distance measures between distributions, the most prominent perhaps being the Kullback-Leibler distance [12]. Below we argue that, for small perturbations at least, our discrepancy metric essentially reduces to the Kullback-Leibler distance.

Indeed, if α_i is not too different from 1, the Kullback-Leibler distance with respect to the uniform distribution, say κ, can be evaluated as follows. First observe that

$$\kappa = \sum_{i=1}^{d}\left(Nf_i\frac{\alpha_i}{N}\right)\log\left(\frac{Nf_i\frac{\alpha_i}{N}}{Nf_i\frac{1}{N}}\right) = \sum_{i=1}^{d}\alpha_i f_i \log\alpha_i.$$

Now let α_i equal $1+\beta_i\varepsilon$ for ε small; $\sum_{i=1}^{d}\alpha_i f_i = 1$ then entails that $\sum_{i=1}^{d}\beta_i f_i = 0$. Using the Taylor expansion $\log(1+x) = x - x^2/2 + O(x^3)$, it follows that

$$\begin{aligned}\kappa &= \sum_{i=1}^{d}(1+\beta_i\varepsilon)f_i\log(1+\beta_i\varepsilon)\\ &= \sum_{i=1}^{d}(1+\beta_i\varepsilon)f_i\left(\beta_i\varepsilon - \frac{1}{2}\beta_i^2\varepsilon^2\right) + O(\varepsilon^3)\\ &= \frac{1}{2}\sum_{i=1}^{d}f_i\beta_i^2\varepsilon^2 + O(\varepsilon^3).\end{aligned}$$

Now replacing $\beta_i\varepsilon$ by $\alpha_i - 1$, and using $\sum_{i=1}^{d}\alpha_i f_i = 1$, we arrive at the approximation, for ε small:

$$\kappa \approx \frac{1}{2}\sum_{i=1}^{d}(\alpha_i^2 f_i - 1).$$

In other words,

$$\frac{\pi_{\mathrm{u}}(k,N)}{\pi(k,N)} \approx \frac{\exp\left(-k^2/2N\right)}{\exp\left(-k^2/2N\cdot\sum_{i=1}^{d}\alpha_i^2 f_i\right)} \approx \exp\left(\frac{k^2}{N}\cdot\kappa\right).$$

As a consequence, we obtain the following elegant approximation for the uniqueness probability in the heterogeneous case:

$$\pi(k,N) \approx \pi_{\mathrm{u}}(k,N)\cdot e^{-k^2/N\cdot\kappa} \approx e^{-(\frac{1}{2}+\kappa)k^2/N}.$$

In other words: to approximate the uniqueness probability for the non-uniform case, we have to take the uniqueness probability for the uniform case, and raise it to the power κ. This κ, the Kullback-Leibler distance, measures the discrepancy of the distribution relative to the uniform distribution. More specifically: the larger κ, the more heterogeneous the distribution is, the smaller the uniqueness probability. It is noticed that the approximation formula is consistent with the one for the uniform case; then $\kappa = 0$.

4. EXPERIMENTS WITH DEMOGRAPHIC DATA

In this section we perform two sets of experiments: (i) experiments in which we validate our approximation formula, as was deduced in the previous section; (ii) experiments in which we assess the impact of heterogeneity, where all computations are based on our approximation formula.

4.1 Validation of the approximation formula

In our validation experiment we have considered the following setup, focusing on the level of anonymity one has after revealing her or his age. Supposing that a group of k individuals is considered, our objective is to determine the probability that each of them has a unique age.

Now the key observation is that the distribution of age is in general *not* uniform: some ages evidently have a higher frequency within the population than others (obviously the higher ages do not occur so frequently). It means that we are in the heterogeneous setup of the previous section.

Our experiments are based on the age distribution of all 428 Dutch municipalities. For each of them we computed the Kullback-Leibler distance κ; let κ_j be the Kullback-Leibler distance of municipality j. More specifically, with φ_{ij} the fraction of the population with age i (for i ranging between 0 and the maximum age, say M) in municipality j (where obviously $\sum_{i=0}^{M}\varphi_{ij} = 1$ for all j), we have

$$\kappa_j = \sum_{i=0}^{M}\varphi_{ij}\log\frac{\varphi_{ij}}{1/(M+1)};$$

the $1/(M+1)$ is the uniform density on $\{0,\ldots,M\}$. In our experiments we took $M = 94$ (thus neglecting a tiny fraction of the population).

In our experiments we took $k = 29$, such that under uniformity we would have a uniqueness probability $\pi_{\mathrm{u}}(29,95) = 0.84\%$. The approximation of the uniqueness probability p_j for municipality j is therefore $0.84\cdot 10^{-2}\cdot e^{-k^2/N\cdot\kappa_j}$. The accuracy of this approximation for municipality j can be validated by sampling (independently) n_+ groups of size k from the age distribution $(\varphi_{0j},\ldots,\varphi_{Mj})$, and to check for each of these samples whether all individuals included are unique (if yes, then increase counter n). Then the uniqueness probability of municipality j can be estimated by $\hat{p} := n/n_+$. To guarantee that this estimate is sufficiently reliable, we should have that the ratio of confidence interval's half-width and the estimate (known as the *relative efficiency*) is below some predefined number r, say, 10%, which means that

$$\frac{t_\alpha\sigma(\hat{p})}{\hat{p}} < r,$$

where $\sigma(\hat{p})$ is the standard error of the estimate, which roughly equals

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_+}} \approx \sqrt{\frac{\hat{p}}{n_+}},$$

and t_α is the t-value corresponding to confidence α (1.96 for $\alpha = 0.95$). An easy computation shows that the number n_+ of experiments needed to make sure that the relative efficiency is below r, is $t_\alpha^2/(r^2\hat{p})$. In the setting of this experiment, with $r = 0.1$ and a uniqueness probability of roughly one percent, and choosing $\alpha = 0.95$, it turns out that we

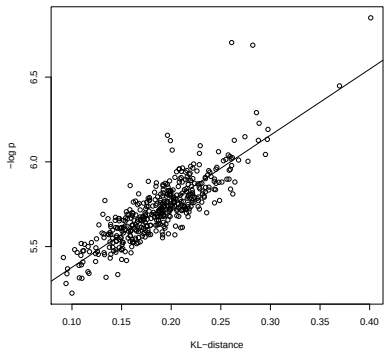

Figure 1: For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age.

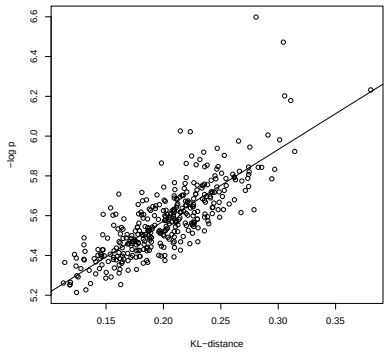

Figure 2: For all Dutch municipalities: the Kullback-Leibler distance and the estimated uniqueness probability, when revealing age and gender.

have to sample until the number of 'unique samples' (that is, the n_+) is about 400. This procedure gives us reliable estimates for the uniqueness probabilities of all municipalities; we call these $\hat{p}_1$ up to $\hat{p}_{428}$.

The question is to what extent the approximation

$$p_j = 0.84 \cdot 10^{-2} \cdot e^{-k^2/N \cdot \kappa_j}$$

is valid, and to this end we can now compare the $0.84 \cdot 10^{-2} \cdot e^{-k^2/N \cdot \kappa_j}$ with the $\hat{p}_j$, for $j = 1$ up to 428. If these numbers would exactly match, then we would have that $\log(0.84 \cdot 10^{-2}) - k^2/N \cdot \kappa_j = \log p_j$, or, in other words, that the logarithm of the uniqueness probability depends linearly on the Kullback-Leibler distance. To study the validity of this relation, we plotted in Figure 1 the value of κ_j against $\log \hat{p}_j$; each dot represents one municipality j.

The main conclusion from Figure 1 is that there is a remarkably good fit, in that the cloud resembles a straight line quite well. The line drawn represents the *least squares fitting*. The percentage of variance that can be explained by the estimator, usually denoted by R^2, provides a measure of the quality of the fit; we obtained $R^2 \approx 0.72$ (popularly: the estimator explained 72% of the variance). We performed the same experiment but then for target probabilities in the order of 10^{-3} and 10^{-4} (rather than the 0.83% of the above experiment); these yield values of the R^2 of even 0.79 and 0.82, respectively.

Another general conclusion is that the use of $\pi_{\mathrm{u}}(k, N)$ without correction by $e^{-\kappa}$ would lead to substantially overestimating the uniqueness probability. Noting that $e^{-5.8} = 3.0 \cdot 10^{-3}$ (where -5.8 is a typical value for $\log p_j$, as seen in Figure 1) indicates that the naïve estimate $\pi_{\mathrm{u}}(29, 95) = 8.4 \cdot 10^{-3}$ is usually off by a factor of about 3, due to the heterogeneity that was not taken into account.

We performed the same experiments for the combination age and gender (that is, $M = 95 \times 2 = 190$). We took $k = 41$, where it is noted that $\pi_{\mathrm{u}}(41, 190) = 0.95\%$. Figure 2 shows that the same effects apply as in the situation in which just age was considered.

4.2 Additional experiments

In this section we report the outcomes of a number of additional experiments; in the numerics we rely on the approximation formula that was developed in Section 3.1, and validated in Section 4.1.

In a first experiment we study the effect of the group size k; we return to our example of Section 4.1, in which the individuals reveal their ages. For clarity of exposition, we chose two municipalities (Laren and Urk) that differ substantially in Kullback-Leibler distance κ (Laren has a κ of 0.0914, Urk has 0.4011). This difference is reflected clearly in the uniqueness probability, as displayed in Figure 3. We approximately have

$$\pi(k, N) \approx \exp\left(-\left(\frac{1}{2} + \kappa\right)\frac{k^2}{N}\right).$$

If we would have assumed uniformity, we have to insert $\kappa = 0$; the resulting graph has been displayed as well.

Our next experiment is inspired by the fact that quite often the data available is relatively coarse-grained and aggregated. For example, in the context of Figure 2 we had information on the number of individuals that were of any given (age, gender)-pair (there were $95 \times 2 = 190$ such pairs). Suppose, however, that we have less information: we only know the number of males and females, and per age the number of individuals (that is, just 97 numbers, where of course the sum over all ages should match with the sum of the male and female). For this situation the same questions can be posed; notice that the machinery developed in this paper does not immediately apply.

Figure 4 provides an indication of the effect that aggregated statistics of age have on the Kullback-Leibler distance for age. The figure shows the Kullback-Leibler at the level of individual ages (i.e., not grouped), at the level of age groups of 2 ('age 0-1', 'age 2-3', 'age 4-5', etc.) and age groups of 5 ('age 0-4', 'age 5-9', 'age 10-14', etc.). The x-axis is a meaningless index of the municipalities, which for clarity of exposition were ordered by Kullback-Leibler distances for the non-grouped scenario.

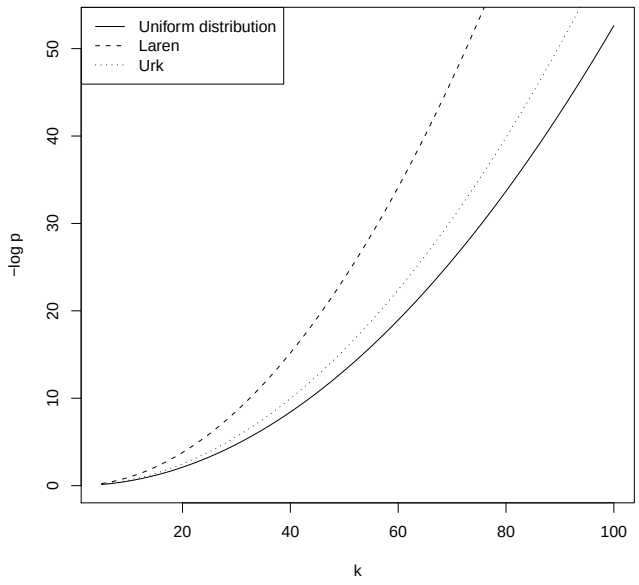

Figure 3: For two Dutch municipalities: the uniqueness probability as a function of the group size k; also the curve under uniformity has been added.

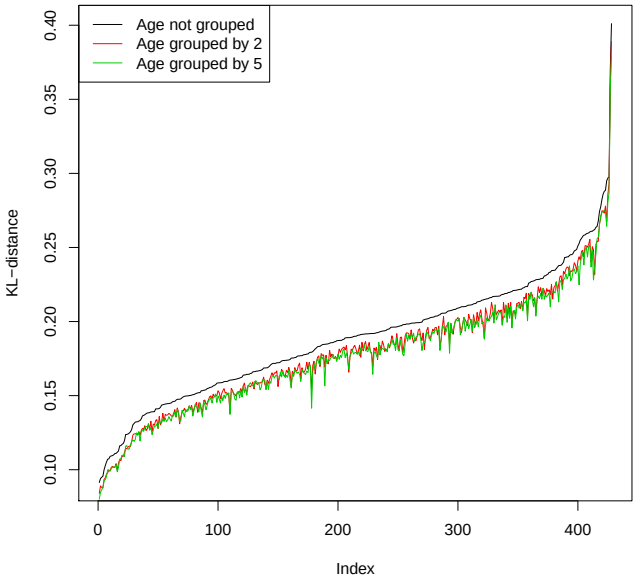

Figure 4: For all Dutch municipalities: the effect of aggregated (age) statistics on the KL-distance.

5. RELATED WORK

In this section we refer to related work. The concept of quasi-identifiers was introduced in [4]; k-anonymity was introduced in [22]. Considering a de-identified data set containing sensitive attributes and quasi-identifiers, the data set is said to be k-anonymous if each quasi-identifier value occurs zero or at least k times within that data set. The concept is intuitive, but it remains unclear how to determine the right k for practical situations considering the disadvantages of information loss involved in the perturbations (generalization and suppression) needed to obtain k-anonymity. k-Anonymity protects against the 'oblivious' adversary targeting *anyone* (re-identifying anything he can, hoping to get lucky) as well as the adversary targeting a *specific individual*. One of the limitations of the original k-anonymity model is that it does not take into account the situation where the sensitive attribute has the same value for all k rows and is revealed anyway. l-Diversity was introduced to address this by requiring that, for each group of k-anonymous records in the data set, at least l different values occur for the sensitive column [15]. Further developments included t-closeness, m-invariance, δ-presence and p-sensitivity [13, 26, 18, 3]. Applications of k-anonymity to communication protocols have been explored in [25, 24].

[14] provides a probabilistic notion of k-anonymity: a dataset is said to be probabilistically $(1-\beta, k)$-anonymous along a quasi-identifier set Q, if each row matches with at least k rows in the universal table U along Q with probability greater than $(1-\beta)$. The authors also found a relation between whether a set of columns forms a quasi-identifier and the number of distinct values assumed by the combination of the columns. $(1-\beta, k)$-anonymity is obtained by solving 1-dimensional k-anonymity problems, avoiding the 'curse of dimensionality' associated with k-anonymity [1]. $(1-\beta, k)$-Anonymity protects against the oblivious adversary, but is insufficient against the adversary targeting a specific individual.

[8] reflects on k-anonymity by introducing the M-score measure, or 'misuseability weight', representing the sensitivity level of the data of each table an individual is exposed to — and, by extension, the harm that misuse of that data can cause to an organization if leaked by employees, subcontractors and partners.

One of the common challenges in k-anonymity and its developments is the recognition of quasi-identifiers. The method we develop in this paper provides a new way of efficiently estimating the likeliness that given set of attributes will function as a perfect quasi-identifier, i.e. that each value of a quasi-identifier unambiguously identifies an individual. That quantification may be useful in privacy impact assessments and policy research.

6. DISCUSSION AND FUTURE WORK

The main contribution of this paper is an approximation for the uniqueness probability when sampling k objects from a population of N, for the situation where the N outcomes are *not* equally likely. The deviation with respect to the uniform distribution is captured by the Kullback-Leibler distance. The approximation clearly shows how the heterogeneity affects the anonymity: the more heterogeneous the population is, the lower the uniqueness probability. In terms of k-anonymity: the more heterogeneous the population is, the

lower the probability that every record in a table will unambiguously identify an individual through the approximated QID.

We emphasize that the anonymity metric used in this paper (that is, the uniqueness probability) does not unambiguously reflect the effect for an individual. For instance, if the individual has an age that is relatively rare within the population (the person is relatively old, for instance), then of course he or she is more likely to reveal his or her identity.

While the approximation formula is of great practical use –allowing data collectors and privacy watchdogs to make predictions about future data collection, allowing individuals to predict what information (not) to disclose at the end of a survey– there are still a number of challenging open questions. For example, age and gender (as in Figure 2) are roughly independent of each other, which makes all computations easier, but quite often when considering multiple quasi-identifiers such a property does not hold. Consider for instance age and marital status: in the Netherlands there will be no married people below 18, while being a widow at a young age is highly unlikely. The question is how these dependencies should be dealt with.

Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project (http://www.vl-e.nl). Part of this project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 901–909, 2005.

[2] M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability*, 5:1–18, 2000.

[3] A. Campan, T. M. Truta, and N. Cooper. p-Sensitive k-Anonymity with generalization constraints. *Trans. Data Privacy*, 3:65–89, August 2010.

[4] T. Dalenius. Finding a needle in a haystack-or identifying anonymous census record. 1986.

[5] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications 2nd ed.,*. Springer Verlag, New York, 1998.

[6] W. Feller. *An Introduction to Probability Theory and its Applications, 3rd Edition.* Wiley, New York, NY, United States, 1968.

[7] P. Golle. Revisiting the uniqueness of simple demographics in the us population. In *WPES'06*, pages 77–80, 2006.

[8] A. Harel, A. Shabtai, L. Rokach, and Y. Elovici. m-score: estimating the potential damage of data leakage incident by assigning misuseability weight. In *Proceedings of the 2010 ACM workshop on Insider threats*, Insider Threats '10, pages 13–20, 2010.

[9] K. Joag-Dev and F. Proschan. The birthday problem with unlike probabilities. *American Mathematical Monthly*, 99:10–12, 1992.

[10] J. Klotz. The birthday problem with unequal probabilities. *Technical Report No. 59, Department of Statistics, University of Wisconsin*, 1979.

[11] M. Koot, G. van 't Noordende, and C. de Laat. A study on the re-identifiability of Dutch citizens. In *Electronic Proceedings of HotPETS 2010*, July 2010.

[12] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[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, Apr. 2007.

[14] S. Lodha and D. Thomas. Probabilistic anonymity. In *Proceedings of the 1st ACM SIGKDD international conference on Privacy, security, and trust in KDD*, PinKDD'07, pages 56–79, 2008.

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

[16] M. Mandjes. *Large Deviations for Gaussian Queues.* Wiley, Chichester, 2007.

[17] F. Mosteller. Understanding the birthday problem. In S. Fienberg and D. Hoaglin, editors, *Selected Papers of Frederick Mosteller*, Springer Series in Statistics, pages 349–353. Springer New York, 2006.

[18] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 665–676, 2007.

[19] P. Rust. The effect of leap years and seasonal trends on the birthday problem. *The American Statistician*, 30:197–198, 1976.

[20] L. Sweeney. Uniqueness of simple demographics in the US population, 2000.

[21] L. Sweeney. *Computational disclosure control: a primer on data privacy protection.* PhD thesis, Massachusetts Institute of Technology, 2001. Supervisor: Abelson, Hal.

[22] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 5:557–570, 2002.

[23] R. von Mises. Über Aufteilungs- und Besetzungswahrscheinlichkeiten. *Revue de la Faculté des Sciences de L'Université d'Istanbul*, 4:145–163, 1938.

[24] P. Wang, P. Ning, and D. S. Reeves. A k-anonymous communication protocol for overlay networks. In *Proceedings of the 2nd ACM symposium on Information, computer and communications security*, ASIACCS '07, pages 45–56, 2007.

[25] X. Wu and E. Bertino. Achieving k-anonymity in mobile ad hoc networks. In *Proceedings of the First international conference on Secure network protocols*, NPSEC'05, pages 37–42, 2005.

[26] X. Xiao and Y. Tao. m-invariance: towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 689–700, 2007.

APPENDIX

A. APPENDIX: A USEFUL LEMMA

Below we present a lemma that is of use when deriving our approximation for the uniqueness probability. Its proof is a matter of elementary algebra.

LEMMA 1. *(i) As* $t \to \infty$,

$$-at - (t^2 - at)\log\left(1 - \frac{a}{t}\right) \to -\frac{a^2}{2}.$$

(ii) As $t \to \infty$,

$$-at + t^2 \log\left(1 + \frac{a}{t}\right) \to -\frac{a^2}{2}.$$