Practical Estimation of Mutual Information on Non-Euclidean Spaces

Yoan Miche, Ian Oliver, Wei Ren, Silke Holtmanns, Anton Akusok, Amaury Lendasse

Practical Estimation of Mutual Information on Non-Euclidean Spaces

Yoan Miche[1], Ian Oliver[1], Wei Ren[1], Silke Holtmanns[1], Anton Akusok[3], and Amaury Lendasse[2]

[1] Nokia Bell Labs
Karakaari 13, FI-02760 Espoo, Finland
[2] Department of Mechanical and Industrial Engineering and
the Iowa Informatics Initiative, The University of Iowa, Iowa City, USA
[3] Arcada University of Applied Sciences, Helsinki, Finland

Abstract. We propose, in this paper, to address the issue of measuring the impact of privacy and anonymization techniques, by measuring the data loss between "before" and "after". The proposed approach focuses therefore on data usability, more than in ensuring that the data is sufficiently anonymized. We use Mutual Information as the measure criterion for this approach, and detail how we propose to measure Mutual Information over non-Euclidean data, in practice, using two possible existing estimators. We test this approach using toy data to illustrate the effects of some well known anonymization techniques on the proposed measure.

1 Introduction

Legal and ethical data sharing and monetization is becoming a major topic and concern, for data holders. There is indeed a strong need to make use of all the Big Data accumulated by the hordes of devices that are becoming part of the Internet of Things (IoT) scene. One major difficulty holding back (some of) the parties in this data monetization and sharing scheme, is the ethical and legal problem related to the privacy of this collected data. Indeed, IoT devices are becoming more and more personal (even though smartphones are already holding on to very personal data), with wearables, medical-oriented devices, health and performance measuring devices…And while users often agree to the use of their collected data for further analysis by the service provider, data sharing to a third party is another type of problem. In this sense, data anonymisation in the broad sense is a rather hot topic, and of the utmost concern for such data sharing scenarios.

There exist many ways to obfuscate the data before data sharing, with the most extreme ones consisting in basically modifying the data so randomly and so much, that the end result becomes unusable. Encryption [10] (when properly carried out) would be one example of such data alteration. And while the promises of Homomorphic Encryption [8], for example, are appealing, the problem of the usability of the data by a third party remains the same: the data has

already been so utterly modified by the encryption scheme, that the internal data structures are too altered to be used for even basic data mining.

Such approaches that obfuscate totally the data have several practical use cases; for example when storage is to be carried out by an untrusted third party. In this work, we focus on another type of use case: that of the need for usability of the data (in the eyes of a third party) while still carrying out some anonymization. The idea here, is to try and measure how much the data has been altered, in terms of its information content (and not in terms of the actual exact values contained in the data). We are thus looking for a measure that would allow for comparing usability to anonymization/privacy.

In this paper, we do not focus on the means of achieving privacy, or what tools can be used for anonymization, but on how to quantify objectively the information loss created by such techniques. Many techniques have already been proposed to alter the data so as to improve the anonymity levels in it: k-anonymity [9], l-diversity [4], differential privacy [2], as well as working towards ways to perform analysis on such modified or perturbed data [5,1]…We give a brief overview of some of these approaches in the next section 2. One of the issues that we attempt to address in this paper, is the fact that they lack an objective criterion to establish how much the data has actually changed, after using such anonymization techniques. In section 3, we introduce some of the notations for the following section 4 about mutual information as a possible criterion for measuring the data loss. In this section, we detail our approach to estimate mutual information over any data set (including those with non-Euclidean data), and the computational details of how we propose to do it. We present the results of this approach over toy data sets in section 5.

2 A short primer on Anonymization Techniques

We first propose in this section to illustrate the effect of some of the most common anonymization techniques, on a limited, traditional data set, depicted in Table 1. The presented anonymization techniques in the following are by no means an exhaustive account of all the possibilities for data anonymization, but probably represent some of the most widely used techniques, in practice.

	Non-Sensitive			Sensitive
ID	Zip Code	Age	Nationality	Condition
01	13053	28	Russian	Heart Disease
02	13068	29	American	Heart Disease
03	13068	21	Japanese	Viral Infection
04	13053	23	American	Viral Infection
05	14853	50	Indian	Cancer

Table 1. Example of Health Data records from a medical institution.

The example data in Table 1 depicts some medical records for a set of patients, possibly from a health care provider. The classification of the data attributes in "Sensitive" and "Non-Sensitive" categories is somewhat arbitrary in this case. The records from Table 1 show no obvious easily identifiable information when considering single fields. Nevertheless, relationships between the non-sensitive fields in this data can probably make it relatively easy to identify some individuals: within a zip code, the nationality and the age allow someone to restrict the set of possible individuals dramatically. The last individual in the table is even more striking as her age, nationality and zip code surely make her stand out of the rest.

2.1 *k*-Anonymity

The term k-anonymity designates in general both the set of properties a data set has to satisfy to be k-anonymous, and the various techniques that can be used to achieve this property. In practice, a data set is said to be k-anonymous if the information for each individual record in the data set cannot be distinguished from at least $k-1$ other records from the very same data set. Two examples of techniques used to achieve k-anonymity are Suppression and Generalisation, and are described in the next two subsections.

Suppression Suppression is obviously the crudest of the possible data alterations, as the data gets simply removed, either for a specific set of records in the data, or for a whole field of data. In the following example Table 2, the whole field of the Age of the records has been removed. This approach can obviously

	Non-Sensitive			Sensitive
ID	Zip Code	Age	Nationality	Condition
01	13053	*	Russian	Heart Disease
02	13068	*	American	Heart Disease
03	13068	*	Japanese	Viral Infection
04	13053	*	American	Viral Infection
05	14853	*	Indian	Cancer

Table 2. Effect of Suppression (K-Anonymity).

lead to strong data alteration, and thus disturb whatever process using the data afterwards. A more subtle solution is provided by Generalization, as follows.

Generalization The idea behind generalisation is to abstract the values in a certain field to higher level (more general) categories. In the example of the data from Table 1, this could mean replacing the last two digits from the Zip Code by zeros, for example, or abstracting the Nationality to "Asian, Caucasian,…"

	Non-Sensitive			Sensitive
ID	Zip Code	Age	Nationality	Condition
01	13053	[20-30]	Russian	Heart Disease
02	13068	[20-30]	American	Heart Disease
03	13068	[20-30]	Japanese	Viral Infection
04	13053	[20-30]	American	Viral Infection
05	14853	[50-60]	Indian	Cancer

Table 3. Effect of Generalization (K-Anonymity).

instead of country level specifics. In the following example Table 3, we generalised the age of the records to 10 years age ranges.

This approach asks the question of what is satisfying in terms of "granularity" of the abstraction? How much information is actually lost in generalising the data fields, and what is the best way to ensure k-anonymity: generalising several fields a little, or one field a lot?

2.2 Differential Privacy

Differential Privacy [2] aims at preserving higher level data statistical properties, typically by introducing controlled noise in the data fields. Without going into the details and the various versions of Differential Privacy [2], we focus in this work on the specific case of ε-differential privacy, in which the ε parameter basically acts as a control parameter for the trade-off between privacy and usability. More specifically, in the rest of the paper (and for the experiments section), we will use Laplace noise added to the data fields, with the ε parameter being the inverse of the Laplace distribution parameter λ.

In the following section 3, we depart a little from the usual notations used in the data privacy literature, to present the mutual information estimators that we propose to use to measure the information loss created by the use of these anonymization techniques.

3 Notations

Let $\mathscr{X}_i = (\mathbb{X}_i, d_i)$ be a metric space on the set $\mathbb{X}_i$ with the distance function $d_i : \mathbb{X}_i \times \mathbb{X}_i \longrightarrow \mathbb{R}_+$. The $\mathscr{X}_i$ need not be Euclidean spaces, and in the cases discussed in the following sections, are not.

Let us then define by $\mathbf{X} = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$ a $N \times n$ matrix, with each column vector $\mathbf{x}^i \in \mathbb{X}_i^{N \times 1}$. The $\mathbf{x}_i$ are thus discrete random variables representing a set of samples over the set of all the possible samples from the attribute represented here by $\mathbb{X}_i$. And $\mathbf{X}$ is a table over these attributes.

The fact that the $\mathscr{X}_i$ are not necessarily Euclidean spaces in this work poses the problem of the definition of the distance function associated, d_i. Indeed, most data mining and machine learning tools rely on the Euclidean distance and its properties; and even if the learning of the model does not require the use

of Euclidean distances directly, the evaluation criterion typically relies on it, for example as a Mean Square Error for regression problems.

Similarly, as described in section 4, information theory metrics estimators such as mutual information estimators typically rely on the construction of the set of nearest neighbours, and therefore also typically (although not necessarily) on the Euclidean distance.

3.1 Distances over non-Euclidean spaces

The argument for considering the use of distances over non-Euclidean spaces in this work, is that it is possible to tweak and modify such non-Euclidean distances so that their distribution and properties will be "close enough" to that of the original Euclidean distance.

More formally, let us assume that we have two metric spaces $\mathscr{X}_i = (\mathbb{X}_i, d_i)$ and $\mathscr{X}_j = (\mathbb{X}_j, d_j)$, with $\mathscr{X}_i$ the canonical d-dimensional Euclidean space (i.e. $\mathbb{X}_i = \mathbb{R}^d$ and d_i the Euclidean norm over it) and $\mathscr{X}_j$ a non-Euclidean metric space endowed with a non-Euclidean metric. Drawing uniformly samples from the set $\mathbb{X}_j$, we form $\mathbf{X}_j = \left[\mathbf{x}_j^1, \ldots, \mathbf{x}_j^n\right]$, a set of random variables, with $\mathbf{x}_j^l$ having values over $\mathbb{X}_j$. Denoting then by f_{d_j} the distribution of pairwise distances over all the samples in $\mathbf{X}_j$, we assume that it is possible to modify the non-Euclidean metric d_j such that

$$\lim_{n\to\infty} f_{d_j} = f_{d_i}, \tag{1}$$

where f_{d_i} is the distribution of the Euclidean distances d_i over the Euclidean space $\mathscr{X}_i$. The limit here is over n as the distribution f_{d_j} is considered to be estimated using a limited number n of random variables, and we are interested in the limit case where we can "afford" to draw as many random variables as possible to be as close to the Euclidean metric as possible. That is, that we can make sure that the non-Euclidean metric behaves over its non-Euclidean space, as would a Euclidean metric over a Euclidean space.

This assumption is "theoretically reasonable", as it comes down to being able to transform a distribution into another, given both. And while this may not be simple nor possible using linear transformation tools, most Machine Learning techniques are able to fit a continuous input to another different continuous output.

4 Mutual Information for Usability Quantification

4.1 Estimating Mutual Information

Using previous notations from section 3, we use the definition of mutual information $I(\mathbf{x}_i, \mathbf{x}_j)$ between two discrete random variables $\mathbf{x}_i, \mathbf{x}_j$ as

$$I(\mathbf{x}_i, \mathbf{x}_j) = \sum_{x_i \in \mathbf{x}_i} \sum_{x_j \in \mathbf{x}_j} p(x_i, x_j) \log \left(\frac{p(x_i, x_j)}{p(x_i)p(x_j)} \right). \tag{2}$$

In practice, the marginals $p(x_i)$ and $p(x_j)$ as well as the joint $p(x_i, x_j)$ are often unknown, and we can then use estimators of the mutual information.

Most of the mutual information estimators (and most famously Kraskov's [3] and Pal's [7,6]) use the canonical distance defined in the metric space in which lies the data. Typically, this is defined and computable for a Euclidean space, with the traditional Euclidean distance used as the distance function.

In the following, we detail shortly the two mutual information estimators that are (arguably) the most used in practice. The goal of this description being to illustrate their dependency on the metric space's underlying distance functions. This is mainly to make the point that mutual information can thus be estimated using non-Euclidean distances over non-Euclidean spaces, given some precautions, as mentioned in the previous section 3.1.

Kraskov's Estimator In [3], Kraskov *et al.* propose a mutual information estimator (more precisely, two of them) relying on counts of nearest neighbours, as follows.

Kraskov's First Estimator The initial mutual information estimator $I^{(1)}$ between two random variables $\mathbf{x}_j^l$ and $\mathbf{x}_j^m$ is defined as

$$I^{(1)}(\mathbf{x}_j^l, \mathbf{x}_j^m) = \Psi(k) - < \Psi(\mathbf{n}_{\mathbf{x}_j^l} + 1) + \Psi(\mathbf{n}_{\mathbf{x}_j^m} + 1) > + \Psi(N), \tag{3}$$

where Ψ is the digamma function, and the notation $< \cdot >$ denotes the average of the quantity between the brackets. In addition, the quantity $\mathbf{n}_{\mathbf{x}_j^l}$ (and defined in the same way, $\mathbf{n}_{\mathbf{x}_j^m}$) denotes the vector $\mathbf{n}_{\mathbf{x}_j^l} = \left[n_{\mathbf{x}_j^l}(1), \ldots, n_{\mathbf{x}_j^l}(N-1)\right]$ holding the counts of neighbours $n_{\mathbf{x}_j^l}(i)$ defined as

$$n_{\mathbf{x}_j^l}(i) = \text{Card}\left(\left\{x_i \in \mathbf{x}_j^l : d_j(x_j - x_i) \leq \varepsilon(i)/2\right\}\right) \tag{4}$$

where $\varepsilon(i)/2 = ||z_i - z_{k\text{NN}(i)}||_{\max}$ is the distance between sample z_i and its k-th nearest neighbour in the joint space $\mathbf{z} = (\mathbf{x}_j^l, \mathbf{x}_j^m)$, and the distance $||\cdot||_{\max}$ defined as $||z_q - z_r||_{\max} = \max\left\{||x_j^l(q) - x_j^l(r)||, ||x_j^m(q) - x_j^m(r)||\right\}$, where $x_j^l(q)$ clunkily denotes the q-th sample from the random variable $\mathbf{x}_j^l$.

Kraskov's Second Estimator The second mutual information estimator $I^{(2)}$ between two random variables $\mathbf{x}_j^l$ and $\mathbf{x}_j^m$ is defined as

$$I^{(2)}(\mathbf{x}_j^l, \mathbf{x}_j^m) = \Psi(k) - 1/k - < \Psi(\mathbf{n}_{\mathbf{x}_j^l}) + \Psi(\mathbf{n}_{\mathbf{x}_j^m}) > + \Psi(N), \tag{5}$$

with Ψ the digamma function, k the number of neighbours to use (to be decided by the user), and this time, $\mathbf{n}_{\mathbf{x}_j^l} = \left[n_{\mathbf{x}_j^l}(1), \ldots, n_{\mathbf{x}_j^l}(N-1)\right]$ is the vector holding counts of neighbours $n_{\mathbf{x}_j^l}(i)$ defined as

$$n_{\mathbf{x}_j^l}(i) = \text{Card}\left(\left\{x_i \in \mathbf{x}_j^l : d_j(x_j - x_i) \leq \varepsilon_{\mathbf{x}_j^l}(i)/2\right\}\right) \tag{6}$$

where $\varepsilon_{\mathbf{x}_j^l}(i)/2$ is the distance between sample z_i and its k-th nearest neighbour $z_{kNN(i)}$, both projected on the $\mathbf{x}_j^l$ space.

Basically, the calculation requires calculating the nearest neighbours of points in a joint space, and counting how many lie in a certain ball.

Note that while we have adapted the notations to our needs, here, the original article relies on the Euclidean distance, and not on arbitrary distances on non-Euclidean distances.

In the following, we illustrate the calculations of the mutual information by these two estimators, over simple non-Euclidean data, namely GPS traces of people.

5 Experimental results

We take in the following experiments, a toy (synthetic) data set that has the same structure as internal data (which cannot be released), namely timestamped GPS locations. We generate five synthetic GPS traces for 5 individuals, as can be seen on Fig. 1. It is worth noting that some of the traces have similar routes, with identical start and end points, while others are totally different.

5.1 GPS routes (timestamped data)

Assume we have a dataset $\boldsymbol{X} = [\boldsymbol{x_1}, ..., \boldsymbol{x_N}]^T$ to depict the trajectory of one specific person, where the attributes of each record $\boldsymbol{x_i}$ explain the location at the corresponding time $\boldsymbol{t_i}$ for this specific person. The locations are represented in GPS coordinates ($\boldsymbol{gps}$) with the form of latitudes ($\boldsymbol{lat}$) and longitudes ($\boldsymbol{lon}$). Each record $\boldsymbol{x_i}$ can then be described by: $\boldsymbol{x_i} = (\boldsymbol{gps_i}, \boldsymbol{t_i}) = ((\boldsymbol{lat_i}, \boldsymbol{lon_i}), \boldsymbol{t_i})$. Hence, the mutual information of the dataset $\boldsymbol{I}(\boldsymbol{X})$ is in a $d \times d$ matrix (in this case $d = 2$: the number of attributes) with the elements holding the mutual information values of the pairwise attributes, illustrated by:

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{X}) &= \{I(\boldsymbol{x}^i, \boldsymbol{x}^j)\}_{1 \leq i,j \leq d} \\ &= \begin{bmatrix} I(\boldsymbol{gps}, \boldsymbol{gps}) & I(\boldsymbol{gps}, \boldsymbol{t}) \\ I(\boldsymbol{t}, \boldsymbol{gps}) & I(\boldsymbol{t}, \boldsymbol{t}) \end{bmatrix}, \end{aligned} \tag{7}$$

Note that the metric space of the GPS coordinates $\mathscr{X}^{(\boldsymbol{gps})} = (\mathbb{X}^{(\boldsymbol{gps})}, d^{(\boldsymbol{gps})})$ is a non-Euclidean space, because the distance of two GPS coordinates $(\boldsymbol{lat}, \boldsymbol{lon})$ is the shortest route between the two points on the Earth's surface, namely, a segment of a great circle. It is obviously not a Euclidean distance. Meanwhile, the metric space of time $\mathscr{X}^{(\boldsymbol{t})} = (\mathbb{X}^{(\boldsymbol{t})}, d^{(\boldsymbol{t})})$ is a Euclidean space with a typical Euclidean distance function.

We illustrate the mutual information matrices by introducing five experimental datasets, with each dataset recording the trajectory for one person. For each person, 100 timestamps and the corresponding $\boldsymbol{gps}$ locations are recorded,

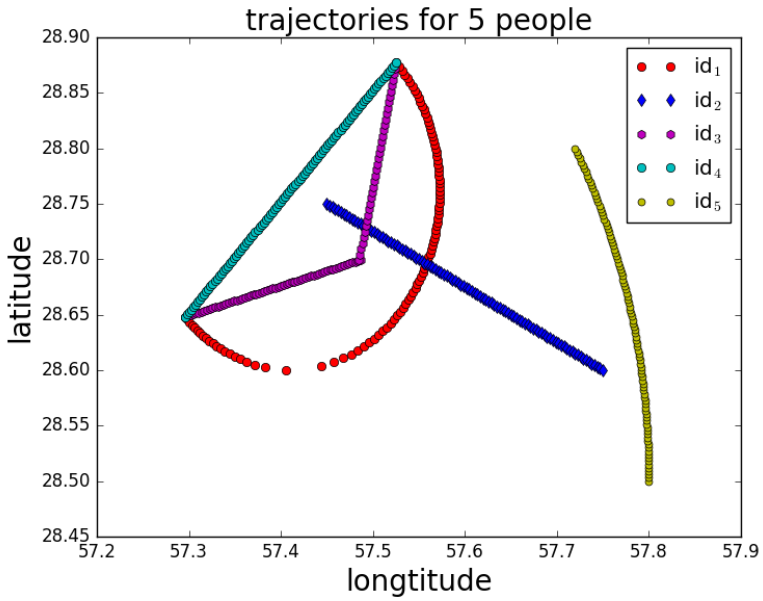

Fig. 1. Five trajectories from the five experimental datasets, respectively.

Table 4. Mutual information (MI) matrices of the five experimental datasets. $I^{(1)}$ and $I^{(2)}$ represent the MI calculated by the first and second *Kraskov* estimators.

	$\begin{bmatrix} I^{(1)}(\boldsymbol{gps},\boldsymbol{gps}) & I^{(1)}(\boldsymbol{gps},\boldsymbol{t}) \\ I^{(1)}(\boldsymbol{t},\boldsymbol{gps}) & I^{(1)}(\boldsymbol{t},\boldsymbol{t}) \end{bmatrix}$	$\begin{bmatrix} I^{(2)}(\boldsymbol{gps},\boldsymbol{gps}) & I^{(2)}(\boldsymbol{gps},\boldsymbol{t}) \\ I^{(2)}(\boldsymbol{t},\boldsymbol{gps}) & I^{(2)}(\boldsymbol{t},\boldsymbol{t}) \end{bmatrix}$
id_1	$\begin{bmatrix} 5.18 & 3.65 \\ 3.65 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 4.18 & 3.20 \\ 3.20 & 4.18 \end{bmatrix}$
id_2	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 4.18 & 2.65 \\ 2.65 & 4.18 \end{bmatrix}$
id_3	$\begin{bmatrix} 5.18 & 3.67 \\ 3.67 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 4.18 & 2.74 \\ 2.74 & 4.18 \end{bmatrix}$
id_4	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 4.18 & 2.60 \\ 2.60 & 4.18 \end{bmatrix}$
id_5	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 4.18 & 2.93 \\ 2.93 & 4.18 \end{bmatrix}$

where the locations are measured at uniform sampling intervals. The trajectories in the datasets are shown in Fig. 1.

Table 4 shows the mutual information (MI) matrices of the five experimental *ids*, respectively. Here we use $I^{(1)}$ and $I^{(2)}$ to represent the values of MI calcu-

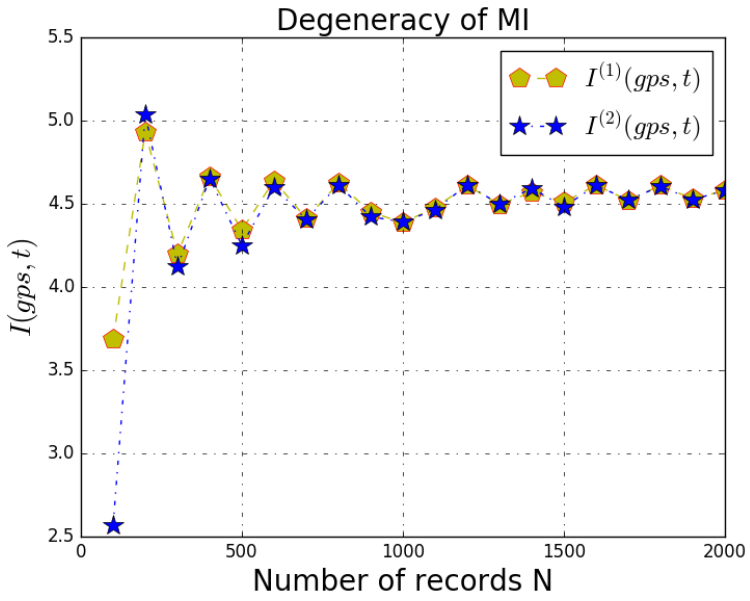

Fig. 2. Mutual information (MI) dependence on the number of samples in the dataset.

lated from the first and second *Kraskov* estimators, respectively. We can see that the MI element value of two identical attributes stays constant, regardless of the variables of the attribute itself: $I^{(1)}(\boldsymbol{gps}, \boldsymbol{gps}) = I^{(1)}(\boldsymbol{t}, \boldsymbol{t}) = 5.18$, and $I^{(2)}(\boldsymbol{gps}, \boldsymbol{gps}) = I^{(2)}(\boldsymbol{t}, \boldsymbol{t}) = 4.18$. Meanwhile, the MI matrices are symmetric with $I(\boldsymbol{gps}, \boldsymbol{t}) = I(\boldsymbol{t}, \boldsymbol{gps})$ for both estimators, of course. The MI values of non-identical pairwise attributes (e.g., $I(\boldsymbol{gps}, \boldsymbol{t})$) are found to be relatively smaller than those values of two identical attributes (e.g., $I(\boldsymbol{gps}, \boldsymbol{gps})$), with the obvious reason that the two identical sets of variables are more mutually dependent than two different variables sets.

The values of $I(\boldsymbol{gps}, \boldsymbol{t})$ are calculated to be in the ranges of $3.65 - 3.69$ and $2.60 - 3.20$ for $I^{(1)}$ and $I^{(2)}$, respectively, compared with the $I(\boldsymbol{gps}, \boldsymbol{gps})$ values of 5.18 and 4.18 for the two estimators. We can see that $I^{(2)}$ is more sensitive than $I^{(1)}$ for *ids* with different trajectories, by giving disparate $I^{(2)}(\boldsymbol{gps}, \boldsymbol{t})$ values. For example, the $I^{(2)}(\boldsymbol{gps}, \boldsymbol{t})$ of id_1 with the value of 3.20 is larger than those for id_2, id_3, and id_4, with values around 2.7. This is mainly due to the relatively more peculiar trajectory of id_1.

5.2 Convergence of the MI Estimators

It is obvious that all the MI values calculated from $I^{(1)}$ are relatively larger than those from $I^{(2)}$. In principle, both estimators should give very similar results. The

difference here is because the number of records with $N = 100$ in each dataset is so small that in the estimators $n_x(i)$ and $n_y(i)$ tend to be also very small with considerably large relative fluctuations. This will cause large statistical errors. We discuss here about the MI convergence with increasing numbers of records.

We take the trajectory of id_4 for example to explain the MI convergence. In the original dataset, there are 100 uniform timestamps and the corresponding 100 uniform locations. We increase the number of records N to 200, 300, 400, ..., 2000, by interpolating uniformly denser timestamps and locations into the trajectory. $I^{(1)}(\boldsymbol{gps}, \boldsymbol{t})$ and $I^{(2)}(\boldsymbol{gps}, \boldsymbol{t})$ is then calculated with the ratio of k/N kept to be 0.01 in the estimators.

The dependence of $I(\boldsymbol{gps}, \boldsymbol{t})$ values over number of record N is illustrated in Fig. 2. It can be seen that the discrepancy of $I^{(1)}$ and $I^{(2)}$ values is getting smaller with increasing N. When N is larger than 800, $I^{(1)}$ and $I^{(2)}$ converge to the values around 4.6.

5.3 *k*-anonymity Effects on the Trajectory Datasets

We have here used the Generalization approach from k-anonymity to modify the data set, and explore the influence of such changes on the mutual information values.

In the following Table 5, k-anonymity applied to the GPS field means that we have in practice rounded the GPS coordinates (lat and lon) by 2 digits, compared to the original precision; when applied to the time field, we have also rounded the time to 10 minutes intervals (instead of second precision).

Table 5. Effects of k-anonymity on Mutual information (MI) matrices for the five experimental datasets.

	$\begin{bmatrix} I^{(1)}(\boldsymbol{gps}, \boldsymbol{gps}) & I^{(1)}(\boldsymbol{gps}, \boldsymbol{t}) \\ I^{(1)}(\boldsymbol{t}, \boldsymbol{gps}) & I^{(1)}(\boldsymbol{t}, \boldsymbol{t}) \end{bmatrix}$			
k-Anon:	none (original)	GPS only	Time only	GPS and Time
id1	$\begin{bmatrix} 5.18 & 3.65 \\ 3.65 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.70 \\ 3.70 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 2.25 \\ 2.25 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.35 \\ 3.35 & 5.18 \end{bmatrix}$
id2	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.73 \\ 3.73 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 2.25 \\ 2.25 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.58 \\ 3.58 & 5.18 \end{bmatrix}$
id3	$\begin{bmatrix} 5.18 & 3.67 \\ 3.67 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.79 \\ 3.79 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 2.30 \\ 2.30 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.58 \\ 3.58 & 5.18 \end{bmatrix}$
id4	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.42 \\ 3.42 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 2.23 \\ 2.23 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.07 \\ 3.07 & 5.18 \end{bmatrix}$
id5	$\begin{bmatrix} 5.18 & 3.69 \\ 3.69 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.64 \\ 3.64 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 2.24 \\ 2.24 & 5.18 \end{bmatrix}$	$\begin{bmatrix} 5.18 & 3.48 \\ 3.48 & 5.18 \end{bmatrix}$

It should be noted that we only report the values for the first estimator, here. In practice, the changes in mutual information incurred by the chosen k-

anonymity values on the GPS are relatively minimal, as can be seen in Table 5. It is interesting to note that the changes on the time cause much more distortion in the data (in terms of the mutual information), possibly because the granularity of the generalization is higher for the time, given the "rounding" chosen. The most interesting feature is that by altering both GPS and time at the same time, the mutual information is higher than when time alone is affected. We explain this by the fact that when these two fields are changed in the same fashion at the same time, the disturbance to the relationship between them is less than when only changing the time. This change to both fields "preserves" some of the relationship better, it seems.

5.4 Differential Privacy Effects on the Trajectory Datasets

We have used ε-differential privacy to obfuscate the trajectory datasets by the Laplace mechanism. We define the privacy function to be a family set of $h = \{h^{(\boldsymbol{gps})}, h^{(\boldsymbol{t})}\}$, where $h^{(\boldsymbol{gps})}$, $h^{(\boldsymbol{t})}$ are the obfuscating functions to perturb the GPS field and time field, respectively.

Differential privacy was applied by adding controllable noise to the corresponding attribute in the dataset, which satisfies the Laplace distribution with mean μ and standard deviation b: $h^{(i)} = \texttt{diff}^{(i)}(\mu, b)$. Let ε be the differential privacy parameter, the standard deviation b of the Laplace noise can be then obtained by:

$$b = \frac{\Delta f}{\varepsilon}, \tag{8}$$

where Δf is the sensitivity of the attribute field.

In the following discussion, we used three family sets of privacy functions, which are:

$$\begin{aligned} h_1 &= \{\texttt{diff}^{(\boldsymbol{gps})}(\mu, b), & \emptyset^{(\boldsymbol{t})} & \}; \\ h_2 &= \{\emptyset^{(\boldsymbol{gps})}, & \texttt{diff}^{(\boldsymbol{t})}(\mu, b)&\}; \\ h_3 &= \{\texttt{diff}^{(\boldsymbol{gps})}(\mu, b), & \texttt{diff}^{(\boldsymbol{t})}(\mu, b)&\}, \end{aligned} \tag{9}$$

where $\emptyset^{(\boldsymbol{i})}$ stands for taking no action to the attribute $\boldsymbol{i}$. For example, h_1 means adding Laplace noise only to the GPS attribute, while the timestamps stay the same; h_2 means adding Laplace noise only to the timestamps attribute; h_3 means adding Laplace noises to both GPS and timestamps attributes.

Fig. 3 shows the obtained pairwise MI values of $\boldsymbol{I}(h^{(\boldsymbol{gps})}, h^{(\boldsymbol{t})})$, where the privacy function sets are applied to the GPS field and time fields with various privacy parameters ε from 0 to 20. We can see that $\boldsymbol{I}(h^{(\boldsymbol{gps})}, h^{(\boldsymbol{t})})$ is monotonically decreasing when the privacy parameter ε decreases. When ε turns to close enough, but not equal, to 0, the MI values collapse at 0, where the fluctuations are the statistic errors caused by small number of sample N in the datasets. It can be well explained by the fact that with smaller values of ε, the amplitudes of the Laplace noise (calculated by Eq. 8) become larger, which distort

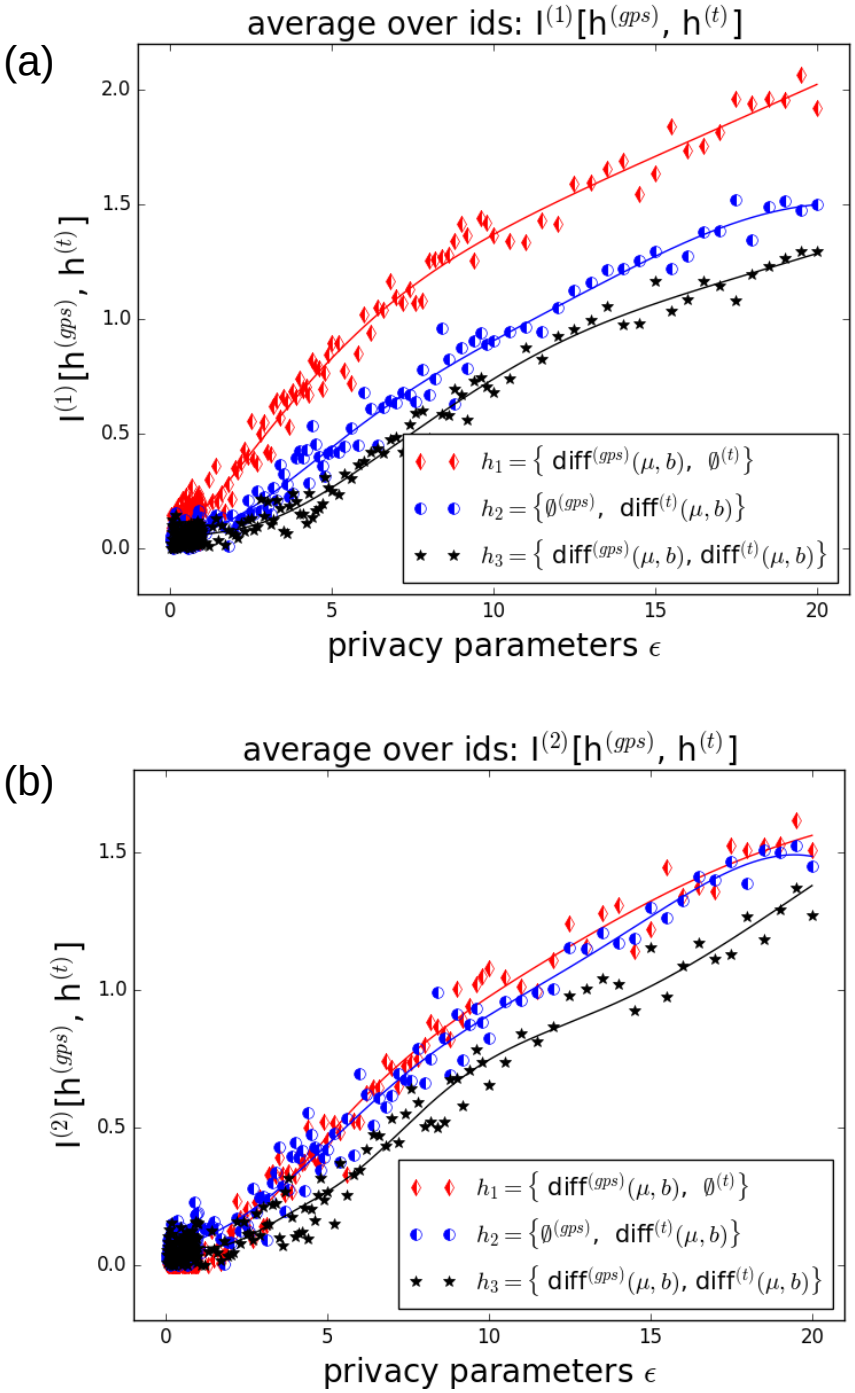

Fig. 3. Mutual information (MI) values of $\boldsymbol{I}(h^{(\boldsymbol{gps})}, h^{(\boldsymbol{t})})$, with the GPS field and time field obfuscated by differential privacy technique with the privacy functions of $h^{(\boldsymbol{gps})}$ and $h^{(\boldsymbol{t})}$, respectively. The MI values are obtained by taking the average of $\boldsymbol{I}(h^{(\boldsymbol{gps})}, h^{(\boldsymbol{t})})$ values from the 5 trajectories. The discrete markers are the obtained averaged MI values, while the corresponding solid lines are the fitted functions with machine learning technique. (a) and (b) are the MI values calculated by the first and second *Kraskov* estimators, respectively.

the metric space or topology of the original datasets more extensively to higher levels with increasing privacy. In another word, we can say that small ε in differential privacy creates greatly anonymised datasets, and effectively alters the metric space with big distortion in terms of the mutual information between the data fields (GPS, time), while the information contents extracted from the anonymised datasets will reduce as a trade-off of increasing privacy. The linkability between the attributes is thus weakened to prevent re-identification of the individuals. Hence the pairwise MI values are decreased.

The efficiencies of altering the MI values by the privacy functions h_1, h_2, and h_3 can be compared in Fig. 3. Both estimators indicate that when applying differential privacy technique on GPS field (h_1) and time field (h_2) separately at the same privacy parameters ε, the time field is more sensitive to reduce the MI values, compared to the GPS field. Moreover, differential privacy applied on both GPS and time (h_3) fields at the same time is the most efficient data anonymization function (in terms of affecting the data relationships regarding mutual information).

As we have discussed before, small MI values stand for high distorsions of the data anonymization, at the possible cost of unusable data, while large MI values imply small alteration of the dataset topology, with a potentially high re-identifiability risk. Therefore, we want to find an acceptable range of MI values, where the dataset is sufficiently anomonyzed to ensure as low as possible risk of re-identification, while the amount of information in the distorted data is still sufficiently usable for future data analysis (in terms of relationships between data fields). Our goal is to control and quantify this distortion, by restricting the privacy parameters in the anonymization functions to specific, acceptable ranges, or by conveying restrictions over the obfuscation functions, also in a controllable manner.

6 Conclusion

In this paper, we have proposed an applied information theoretic approach to measure the impact of privacy techniques such as k-anonymity and differential privacy, for example. We examine, by this approach, the disturbances in the relationships between the different columns ("fields") of the data, thus focusing on a data usability aspect, rather than actually measuring privacy. We propose to do this on any data that can be taken over a metric space, i.e. for which a distance between elements is the sole practical need. We develop an approach to estimate mutual information between such data types, using two well known estimators, and demonstrate their behaviour over simple experimental tests. We finally investigate the effects of k-anonymity (specifically, generalisation) and differential privacy over timestamped GPS traces, and illustrate the effects of these widely used privacy techniques over the information content and the relationships contained in the data. In effect, the results obtained are as expected, except possibly for the case where the generalisation in k-anonymity is performed over both fields at the same time, and leads to some preservation of the data

structure and relationships. Future work will include other data types and other mutual information estimators to verify the results observed in this work.

References

1. Josep Domingo-Ferrer and David Rebollo-Monedero. Measuring risk and utility of anonymized data using information theory. In *Proceedings of the 2009 EDBT/ICDT Workshops*, EDBT/ICDT '09, pages 126–130, New York, NY, USA, 2009. ACM.
2. C. Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, volume 4978 of *LNCS*, pages 1–19. Springer, April 2008.
3. Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004.
4. Ninghui Li and Tiancheng Li. t-closeness: Privacy beyond κ-anonymity and ℓ-diversity. In *In Proc. of IEEE 23rd International Conference on Data Engineering (ICDE'07*, 2007.
5. Bernd Malle, Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. The right to be forgotten: Towards machine learning on perturbed knowledge bases. In Francesco Buccafurri, Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Availability, Reliability, and Security in Information Systems: IFIP WG 8.4, 8.9, TC 5 International Cross-Domain Conference, CD-ARES 2016, and Workshop on Privacy Aware Machine Learning for Health Data Science, PAML 2016, Salzburg, Austria, August 31 - September 2, 2016, Proceedings*, pages 251–266, Cham, 2016. Springer International Publishing.
6. D. Pál, B. Póczos, and C. Szepesvári. Estimation of R\'enyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs. *ArXiv e-prints*, March 2010.
7. Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In J.D. Lafferty, C.K.I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1849–1857. Curran Associates, Inc., 2010.
8. Bharath K. Samanthula, Gerry Howser, Yousef Elmehdwi, and Sanjay Madria. An efficient and secure data sharing framework using homomorphic encryption in the cloud. In *Proceedings of the 1st International Workshop on Cloud Intelligence*, Cloud-I '12, pages 8:1–8:8, New York, NY, USA, 2012. ACM.
9. Latanya Sweeney. κ-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.
10. J. Wei, W. Liu, and X. Hu. Secure data sharing in cloud computing using revocable-storage identity-based encryption. *IEEE Transactions on Cloud Computing*, PP(99):1–1, 2016.