A Model for Metricising Privacy and Legal Compliance

Ian Oliver, Yoan Miche, Wei Ren
Nokia Bell Labs
Espoo, Finland
firstname.lasttime@nokia-bell-labs.com

***Abstract*—In order for a dataset to be legally compliant - in some sense - with privacy laws such as the General Data Protection Regulation (GDPR) various steps must be taken to ensure the removal of data that might compromise or reveal personal data. This can be achieved through a process of removal of information content or semantics; which if done incorrectly can render that dataset in violation of such laws. Machine learning presents a technology based around the analysis of dependencies and correlations of a dataset. This can be used to measure information content within the bounds of the dependencies estimators used. Utilising this we can measure the effects of anonymisation upon a dataset and the efficacy of said anonymisation functions. If we additionally characterise what anonymisation means in terms of information loss and construct classification functions we have a framework in which the decision over whether an anonymisation is sufficient can be made. This can then be extended to an automation scenario where it becomes potentially possible that texts such as as the GDPR can be rendered as said classification functions.**

***Index Terms*—Privacy, Machine Learning, Metrics, Legal, Requirements, Entropy, Information, Data Quality**

I. Introduction

The impact of privacy legislation upon the development of information systems and associated processes has been well documented. Acting upon and demonstrating compliance with said legislation is becoming more complex with the introduction of the EU's GDPR [1] which effectively changes privacy from a compliance activity [2] to a risk management activity [3], [4].

Thus in order to develop said information systems it must be demonstrated that the system in question is processing information in such a way that, depending upon circumstances, a **sufficient amount** of anonymisation has taken place. Herein lies a difficulty, that given a suite of anonymisation techniques, how does one ensure that a dataset has been sufficiently anonymised? Further, how does one assess the effectiveness of an anonymisation algorithm? Utilising the wrong algorithm and not understanding the effects of an algorithm applied to one aspect of the data on another can have potentially catastrophic effects in terms of lack of privacy.

The aim of this paper is primarily to present a *model* for such anonymisation or information content loss functions. By combining an analysis of the effectiveness of an anonymisation algorithm along with a notion of assessing the degree of risk in a dataset we are effectively creating a metric or measurement of the amount of privacy or information content in a dataset [5], [6]. If we can measure the degree of privacy achieved then it becomes possible to set bounds and state if a given dataset is sufficiently anonymised. The implications of this simple statement however would be far reaching: automating privacy compliance and an end to lawyers, perhaps? Simply put, map a dataset to a value and have a function that returns whether the dataset is compliant or not for a given threshold:

$$\mathcal{D} \xrightarrow{m} \mathbb{R}_{[0,1]} \tag{1}$$

$$\text{compliant?}(d : \mathcal{D}, t : \mathbb{R}_{[0,1]}) = \begin{cases} \textbf{Compliant}, & m(d) \leq t \\ \textbf{Not Compliant}, & \text{otherwise} \end{cases} \tag{2}$$

So far however we lack both a useful measure of privacy and a general idea of how such a privacy measurement could be constructed. In this paper we present results based upon utilising mutual information as a metric, analysing the effects of anonymisation (hashing and differential privacy [7]) and constructing classification functions to decide whether a given dataset has been sufficiently anonymised or is effectively *legal*.

We present a case study based upon anonymising telecommunications signalling data to explore properties of anonymisation and said classification functions. We then present a notion of a privacy metric through a pair of classification functions and a discussion of what it means to be legally compliant given the existance of such structures. The authors note that a true privacy metric is elusive, however we aim to show that such a metric and structures over this can be constructed thus admitting a discussion of what such a metric and notions of compliance mean in such a framework.

One particular element to emphasise initially is that while characteristic functions of the form shown above can be constructed, the authors note that firstly that these are not

the only considerations in deciding whether a data set is compliant or not, that the binary nature of the function above is deliberately very abstract and that said functions can not be properly applied in isolation. The model and analysis being constructed here is primary to provide a framework for further analysis of said constructions. We note with some irony however that the final decision on whether a system or data set complies with privacy laws is often ultimately a binary yes or no decision on the part of the privacy lawyer/officer.

II. Case Study

Telecommunications operators necessarily collect vast amounts of customer information as mobile devices interact with network components under stricter laws than "over-the-top" providers [8]. Specifically when a mobile device connects or disconnects from a base station we are left with a record containing user identifiers, device identifiers, timestamp and base station identifier; the latter of which can be transformed to a precise location. Collection of this data over time allows triangulation of coördinates and an effective method of accurately tracking users' movements [9]. The further analysis of this data is obvious in terms of user/group behaviour analysis [10], [11]. We extracted data from a system which processes telecom signalling data through various processes to output datasets which are *sufficiently* anonymised to satisfy the requirements of telecommunications privacy laws. The data flow of this is shown in figure 1.

The process of collecting and supplying the finalised data involves 5 datasets which may include personal data which would be 'illegal' to release: the initial data storage, the file storage, raw data, atomic data and report storage as name in the data flow diagram. We must prove that the data being released after the aggregation process, the data collected in the store called 'report storage' has been sufficiently anonymised with respect to any preceding dataset.

Three conditions therefore must hold over the process (or subprocesses as the case maybe): the amount of information in the output data must be less than the amount of information in the input data, the amount of information in the output data must be less than some legally acceptable threshold, and, the output data must contain enough information to be usable for a given purpose. Let $\mathcal{D}_i$ and $\mathcal{D}_o$ be any pair of datasets where i denotes the input and o the output. The function m is a measure of the information content and t threshold value for some property:

1) $m(\mathcal{D}_i) > m(\mathcal{D}_o)$
2) $m(\mathcal{D}_o) \leq t_{legal}$
3) $m(\mathcal{D}_o) \geq t_{usable}$

The input signalling dataset contains three fields, an identifier, a timestamp and a location. The base dataset is extended by interpolation of the locations and a finer grained timestamp resulting in a processed dataset. This latter dataset is the one to be released through an anonymisation algorithm. We therefore are required to show that the released dataset has been sufficiently anonymised. To the processed dataset in the experiments we apply the following anonymisation functions:

- hashing of the identifier
- differential privacy of the location
- differential privacy of the timestamp
- differential privacy of the location and timestamp

The measurement of the information content is made by calculating the mutual information of the dataset: a measure of the internal consistency. We demonstrate that increasing noise lowers the mutual information; also we demonstrate that hashing of the identifier effectively has no effect on the information content - though we do not specifically rule this out as being a method of anonymisation (see discussions).

III. A Theory of Anonymisation

Anonymisation can be formulated as the application of a function to a data set which reduces its information content [12]. For example, removing possibility to recovering or relinking that data so that a unique individual (or group of individuals) can be re-identified. In order to achieve this, anonymisation function add noise or decrease the information content of a dataset. The goal is to find a dataset that is sufficiently anonymous enough to be legally compliant *and* containing enough information to be useful for whatever purposes it will be put to.

A. Properties of an Anonymising Function

We can list a set of functions that anonymise a dataset. Each function in $\mathcal{A}$ below takes a dataset and set of parameters $\pi_1 \dots \pi_n$ as input and returns a dataset.

$$\mathcal{A} = \{(\epsilon, \delta)\text{-}differentialPrivacy, \kappa\text{-}anon, hash(\dots), \dots\} \tag{3}$$

We may then take a particular function $\alpha \in \mathcal{A}$ with suitable values for its parameters, for example, in one instance α might be a differential privacy function applied to the location fields in some input dataset $\mathcal{D}_i$: diffP($\mathcal{D}_i, field = \text{LOCATION}, \epsilon = 0.1, \delta = 0.9$) and expect a suitably processed set of location fields in the output set $\mathcal{D}_o$:

$$\mathcal{D}_i \xrightarrow{\alpha(\pi_1 \dots \pi_n)} \mathcal{D}_o \tag{4}$$

We expect an anonymising function to reduce the information content present in the input dataset. Given a suitable measure m of information content the following diagram commutes under an anonymising function [13]:

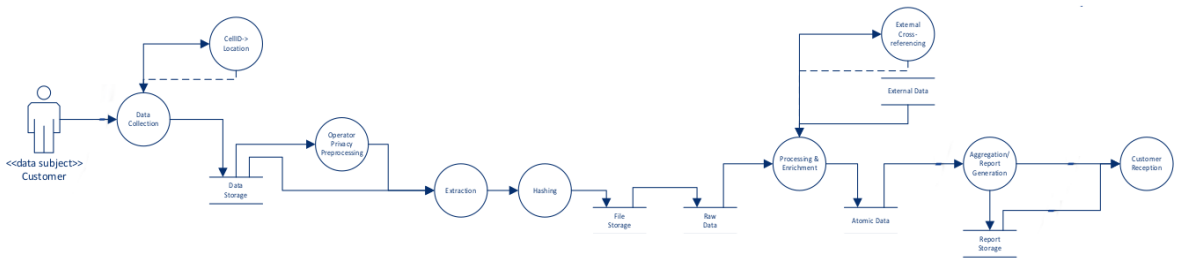


Fig. 1. Case Study Data Flow Diagram

$$\begin{array}{ccc} \mathcal{D}_i & \xrightarrow{\alpha(\pi_1 \dots \pi_n)} & \mathcal{D}_o \\ \downarrow m & & \downarrow m \\ \mathcal{M}_i & \xrightarrow{>} & \mathcal{M}_o \end{array} \tag{5}$$

Subsequent applications of anonymising functions can composed such that the information reduction is monotonic over the subsequent applications, ie:

$$\begin{array}{ccccccccc} \mathcal{D}_i & \xrightarrow{\alpha_1} & \mathcal{D}_1 & \xrightarrow{\alpha_2} & \mathcal{D}_2 & \xrightarrow{\alpha_n} & \dots & \xrightarrow{\alpha_n} & \mathcal{D}_o \\ \downarrow m & & \downarrow m & & \downarrow m & & & & \downarrow m \\ \mathcal{M}_i & \xrightarrow{>} & \mathcal{M}_1 & \xrightarrow{>} & \mathcal{M}_2 & \xrightarrow{>} & \dots & \xrightarrow{>} & \mathcal{M}_o \end{array}$$

B. Desirable Properties

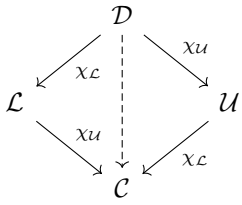
We introduced a naïve characteristic function in (2) earlier and here we extend this notion to other desirable properties of datasets. We restrict ourselves to two such desirable and prominent properties in this discussion:

- whether the dataset is legal
- whether the dataset is useful

Datasets are often presented in terms of whether the dataset is compliant and *sufficiently* anonymised such that it can be used or released. Data scientists would prefer the dataset to be as correct and detailed as possible to increase the accuracy of processing and statistical inference (amongst others). Using the subobject classifier construct defines the properties of such functions. In (6) we out legal datasets $\mathcal{L}$ from $\mathcal{D}$ through a classifying function $\chi_{\mathcal{L}}$.

$$\begin{array}{ccc} \mathcal{L} & \longrightarrow & \mathbf{1} \\ \downarrow & & \downarrow \\ \mathcal{D} & \xrightarrow{\chi_{\mathcal{L}}} & \{0,1\} \end{array} \tag{6}$$

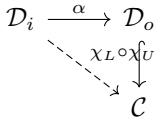
Similarly other classifier functions can be defined for other properties such as statistical usefulness [14]. These are composable, eg: $\chi_{\mathcal{L}} \circ \chi_{\mathcal{U}}$ extracts the legal and useful datasets denoted $\mathcal{C}$ in (7).



(7)

C. Successful Anonymisation

The application of an anonymisation function, or sequence of anonymisation functions as the case may be, occurs when a the output dataset exists in the set of legal and usable datasets; the diagram shown in (8) holds.



(8)

It is of course another matter whether the structure $\mathcal{C}$ is empty, or if not, which particular dataset is of most value with respect to the legal and usable aspects.

IV. Analysing Anonymisation Functions

While many potential information entropy metrics exist, one of the most prevalent at this point in time is Mutual Information [15], which can be used in a data mining/ machine learning context, e.g., to infer and quantify possible relationships between variables (or fields of data structures) for further analysis. In effect, Mutual Information measures the higher order dependencies (higher order meaning here more than simple correlation, including possibly complex non-linear relationships) between variables. Typically, mutual information can be computed between any number of variables, although in practice, measuring it efficiently and correctly between more than two variables is extremely challenging [15], [16], [17].

In the context of this paper, we look specifically at the mutual information values between pairs of variables, and collate these values in a matrix, following the scheme in Eq. 9.

$$\mathbf{X} = \left[\left[\begin{array}{c} \\ \mathbf{C}_1 \\ \\ \end{array} \right] \left[\begin{array}{c} \\ \mathbf{C}_2 \\ \\ \end{array} \right] \left[\begin{array}{c} \\ \mathbf{C}_3 \\ \\ \end{array} \right] \right] \longrightarrow \mathrm{MI}(\mathbf{X}) \quad (9)$$

where $\mathrm{MI}(\mathbf{X})$ is defined:

$$\mathrm{MI}(\mathbf{X}) = \left[\begin{array}{ccc} 1 & \mathrm{MI}(\mathbf{C}_1, \mathbf{C}_2) & \mathrm{MI}(\mathbf{C}_1, \mathbf{C}_3) \\ \mathrm{MI}(\mathbf{C}_2, \mathbf{C}_1) & 1 & \mathrm{MI}(\mathbf{C}_2, \mathbf{C}_3) \\ \mathrm{MI}(\mathbf{C}_3, \mathbf{C}_1) & \mathrm{MI}(\mathbf{C}_3, \mathbf{C}_2) & 1 \end{array} \right]. \quad (10)$$

The matrix $\mathrm{MI}(\mathbf{X})$ contains in effect the pairwise mutual information values between all combinations (pairs) of variables in the original data set (here, three variables/ data fields).

One rationale for using mutual information when measuring the impact of privacy techniques on the usability of data, is that tampering with variables that have dependencies between them, will lower the mutual information between them [18]. This is rather straightforward in the case of added random noise: Suppose that two random variables X and Y are being used for this example, and that X and Y are not purely independent, then $\mathrm{MI}(X, Y) \neq 0$. In the case of privacy techniques that behave like additive random noise, we can very simply see that adding such noise to one (or to both) variable, will decrease the mutual information between them. Given a random noise variable Z, independent from X and Y, we have then

$$\begin{array}{rcl} I(X+Z,Y) - I(X,Y) & = & H(X+Z) - H(X+Z|Y) \\ & - & H(X) + H(X|Y) \\ & \leq & H(X) + H(Z) - H(X|Y) \\ & - & H(Z|Y) - H(X) + H(X|Y) \\ & \leq & H(Z) - H(Z|Y) = 0 \end{array} \quad (11)$$

which means that $I(X+Z,Y) \leq I(X,Y)$ as required by Eq.5.

Thus, the mutual information is a possible means of quantifying the loss of information generated by such privacy techniques as random additive noise (as in the case of some Differential Privacy techniques). For more complicated privacy techniques, the proofs would obviously be more complex to derive, and potentially impossible to express properly analytically (such cases as a change of data field (by hashing, encrypting…), non-independent, non-additive noise…).

A. Analysing Hashing

Hashing is a very common technique for hiding data with cryptographic hashes being the primary mechanism. One of the problems with hashing is that they are mistaken as mechanisms for reducing information content over the dataset whereas their actual function is to hide the contents of the data and not change it to another form. Ostensibly hashing functions are one-way, but in type theoretic terms effectively change, for example, identifiers into identifiers, albeit of a different kind.

If hashing is to be used then case must be taken to change the salting of the hash over a suitable period of time to reduce the possibility of tracking and/or re-identifying structures. Even after this, advanced statistical techniques can extract patterns.

In figure 2 we show the change in mutual information between two datasets where one has the original data and the other where the identifier field has been cryptographically hashed.

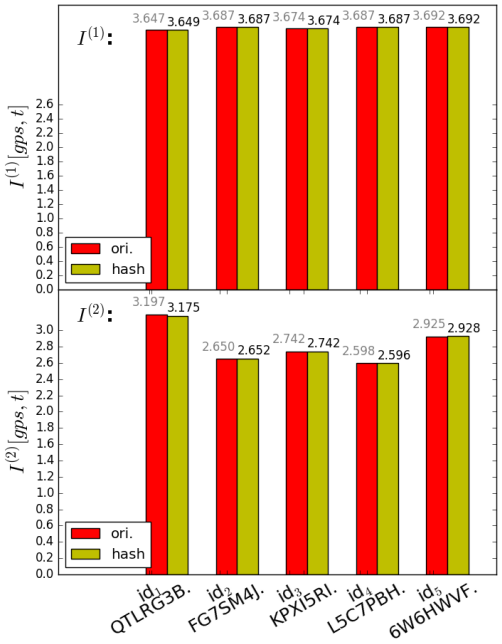


Fig. 2. Example of Lack of Mutual Information Change Under Hashing

The point here is to state clearly that hashing does not imply information loss, and according to our specific definition we do not consider hashing to be an anonymisation function in this context as defined in (5).

B. Analysing Differential Privacy

We take differential privacy as the canonical example of an anonymisation function and one that (should) adheres to the constraints in (5). Differential privacy works by adding noise to a given structure, for example, locations or timestamps. Note: we do not differentiate between different kinds of differential privacy functions but base this on the original work in [7] and use that as a basis for this family of anonymisation functions.

Given our signalling dataset we apply differential privacy with suitable values of ϵ and δ to add enough noise such

that the chosen structures and thus the dataset as a whole is rendered more private. We wish then to investigate and assist in the following:

1) the selection of structures within a dataset on which to apply differential privacy
2) the selection of suitable/acceptable ϵ and δ values
3) the effects upon the dataset as a whole

The first point is relatively simple in that most datasets can be characterised in terms of containing three structures: identities, locations and timestamps. The latter two are suitable for the application of differential privacy. However we are left with the following choices along with the selection as noted in the second earlier point of suitable values of ϵ and δ:

1) diffP applied to location
2) diffP applied to timestamp
3) diffP applied to location and timestamp

We can construct a result space that contains the mutual information matrices for a range of and therefore indexed by (ϵ, δ) values for a given anonymisation function.

$$\begin{array}{ccc} \mathcal{D}_i & \xrightarrow{\alpha_{(\epsilon\in[0,\infty),\delta\in[0,100)}} & \mathcal{D}_o \\ \downarrow m & & \downarrow m \\ \coprod_{\pi\in(\epsilon\in[0,10],\delta\in[0,100])} \mathcal{M}_\pi & \xrightarrow{>} & \coprod_{\pi\in(\epsilon\in[0,10],\delta\in[0,100])} \mathcal{M}_\pi \end{array} \tag{12}$$

For our anonymisation property to hold, some degree of anonymisation must take place. If the value of (ϵ, δ) is set such that no noise is added and/or no elements in $\mathcal{D}_i$ are affected then $\mathcal{D}_i$ is isomorphic to $\mathcal{D}_o$ and the mutual information measurements of both are effectively the same.

A fragment of the result space is:

$$\left\{ \ldots, \left((\epsilon = 0.1, \delta = 90), \begin{array}{c|ccc} & ID & LOC & TS \\ \hline ID & 1 & & \\ LOC & 0.8 & 1 & \\ TS & 0.1 & 0.3 & 1 \end{array} \right), \ldots \right\} \tag{13}$$

From which we can extract or project individual, normalised correlation values from each matrix for each (ϵ, δ) value and plot this to obtain graph of the changing mutual information values for any pair of structures, eg: location against timestamp as shown in figures 3 and 4. The former of which shows the decrease in mutual information as $(\epsilon, 0)$-differential privacy being applied to location, timestamp and then both in equal measure.

In the second case in figure 4 we extend the plot to (ϵ, δ)-differential privacy being applied only to location.

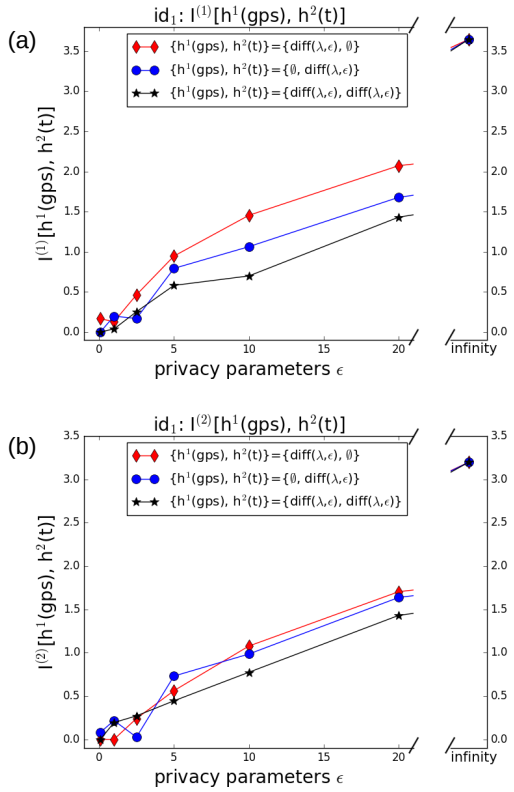


Fig. 3. Example Plot of MI for $DiffP_{(\epsilon,0)} \in [0, 20]$ Applied to the Location, Timestamp and Location-Timestamp Structures

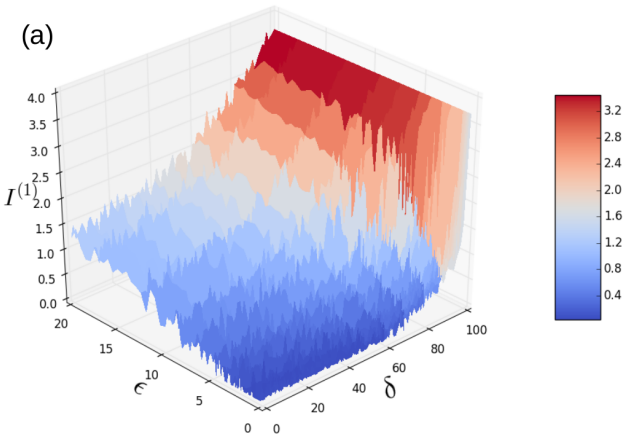


Fig. 4. Example Plot of MI for $DiffP_{(\epsilon,\delta)\in[0,20],[0,100]}$ Applied to the Location Structure

From this graph we can see that if no noise is added then the mutual information remains at its maximum. As noise is added for $(\epsilon, 0)$ differential privacy, ie: all rows in the dataset are affected then the mutual information decreases slowly but always remains low even with low levels of noise (high values for ϵ).

The implication of δ is interesting in that when only a

small set of row, say 10% ($\delta = 90$) are affected then mutual information between location and timestamp does not reduce until very high levels of noise are added. Such a result has implications here that much of the data is actually correct despite noise is being added.

This however is still just one projection of the results and a fuller knowledge of the affects of applying an anonymisation function to one structure in the dataset can only be gained if we project into three (in this case) separate visualisations.

The set of protections is calculated by taking the set of pair-wise correlations of the lower matrix triangle structures in the source dataset. In our case this gives three projections: identity by location, identity by timestamp and location by timestamp (as shown in figure 4). This gives us another structure $\mathcal{V}$ which is the structure of projections by pair-wise correlations:

$$\coprod \mathcal{M}_\pi \longrightarrow \coprod_{s \in structPairs} \mathcal{V}_s \tag{14}$$

Given the result set and projections we can now properly construct and visualise the characteristic functions χ_L and χ_U. These can be constructed by taking the limit over each characteristic function applied to each element $M\mathcal{M}_p i$ in the above (14). Another way of visualising this is that χ_L (or any other classifier) slices through each manifold (as in 4) generated for each structure. For example in figure 5 we show a possible slice separating this particular space into 'legally' and 'non-legally' compliant based on the decrease in information content.

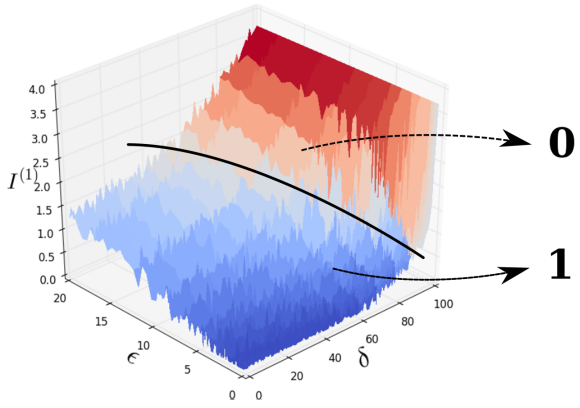


Fig. 5. Demonstration of Classifying

The classifying function for χ_U, ie: usability, is similarly constructed and the composition of these functions corresponds to the intersection of the slices over the manifolds. If it is the case that the intersection does not exist then either we are anonymising too much or are too strict on the required usability characteristics, possibly.

C. Causal vs Non-Causal Fields

We have developed a hypothesis that states that if anonymisation is enacted upon causal data, ie: the data that controls the ordering of events within a dataset, eg: a timestamp field, then anonymisation has a potentially greater effect upon the information loss. In figure 6 we show three examples where differential privacy of the same type was applied with the same (ϵ, δ) values to location, timestamp and both location and timestamp fields.

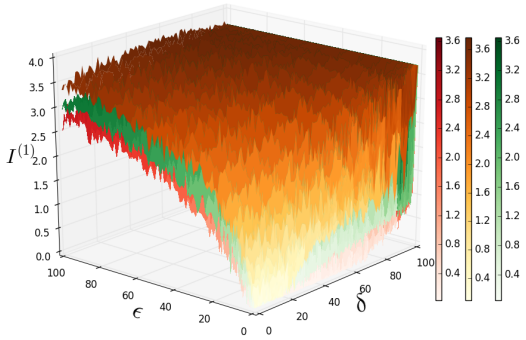


Fig. 6. Information Loss Applied to Temporal and Non-Temporal Fields

The uppermost manifold corresponds to differential privacy being applied to only the location field, the middle to the temporal field and the lower to both. In all cases the measure property of mutual information is preserved as expected. The decrease in mutual information when applied to the timestamp is attributed to the inability of the mutual information estimators to find correlations between the location and the identifier, that is the relationship of the tracks to a person as the ordering of the fields is disrupted.

More work is required here to set the bounds on the differential privacy with respect to the granularity of the events but all early indications show that as soon as event ordering is disrupted then information loss is greater when causal fields are anonymised.

Care should be noted that applying an anonymisation function to a single field may not render the dataset immune to deanonymisation!

D. Properties of Anonymisation Functions

As already noted and shown visually in figure 6 and defined in Eq.(11) applications of anonymisation functions observe the rules of metrics or distance spaces. This means that on a manifold such as that in figure 7 we could effectively plot the optimal trajectories and thus optimise over a set of

anonymisation functions in order to reach a particular point representing a certain value of mutual information and thus information loss.

In figure 7 we show four anonymisation functions' trajectories in terms of the starting and ending values of the mutual information metric on that particular manifold. We show seven anonymisation functions a_0 to a_6 (variations on differential privacy most likely with one exception).

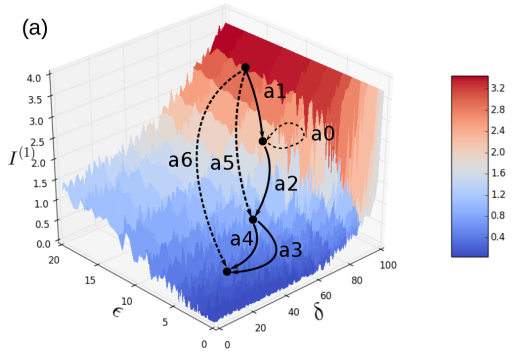


Fig. 7. Plotting Anonymisation Function Trajectories

The first thing to note is that some anonymisation functions are compositions, or *behave* similarly to compositions of others, eg: $a_5 \approx a_2 \circ a_1$. Some anonymisation functions are behaviourally equivalent such as $a_3 \approx a_4$, while some do not fit our definition of anonymisation, ie: a_0 which may be some hashing function in this context. If we have choices of anonymisation functions available and knowledge of their properties such as computational complexity or other efficiencies then it becomes possible to optimise their application.

From a legal perspective knowing the existence of intermediate datasets that may fall outside of any compliant area defined by χ_L provides us information about which compositions of anonymisation functions should be combined into indivisible functions. For example, we might choose to compost a_1 and a_2 as a_5 to avoid the accidental release of an intermediate data set. Consider how this compares with the intermediate datasets shown earlier in figure 1.

Furthermore we can also identify anonymisation functions which do not work or are not being applied correctly. For example, we stated that a_0 might be a hashing function which does not fit our particular definition of anonymisation here. This is easily seen as a function which effectively behaves as an identity function, ie $a_0 \approx \mathbf{Id}_D$.

V. Discussion

We have now demonstrated a mechanism for measuring the mutual information content of a given dataset and a structure for deciding whether that dataset is legally compliant and/or useful. In terms of privacy requirements, automation and the mathematical formalisation of a legal text such as the GDPR is profound. We discuss the following within the context of this work here.

We assert that a mathematical formulation of privacy in terms of information theory, type theory and model theory *is* possible, albeit 'hard' [19]. Contributing to this of course is the fact that many techniques, such as the measurement techniques presented here, a general theory of privacy and a detailed understanding of some of the finer semantic effects, eg: causal vs non-causal data, does not exist in a full nor coherent form at this time [20].

What can be shown is that a structure can be created that allows aspects of privacy to be properly formulated. The act of stating that an anonymisation function is one that reduces information content now gives us a starting point for a proper classification of anonymisation functions and a possibility of extracting aspects of the assumed context in which they work. One of the major problems seem with anonymisation function is their use is often flawed in that their application is made just because they are denoted anonymisation functions [21], [22]. The correct usage as we have presented depends upon the internal structures of a dataset and how these are related. Incorrect application to a structure or part of a structure might still render a dataset unanonymised [23]. The AOL Search Data Leak from August 2006 is a classic example of not understanding how anonymisation works and its incorrect application.

Further to this the amount of data required for an accurate estimation and later further learning is extensive. For the dataset we are using extracted from the signalling data, which itself can reach potentially millions of records per day, smaller extracts were taken. However, if the extract is too small then the error rates from the machine learning estimation functions will be too great. This is somewhat obvious from the results when visualised as in figure 8 where the variation in accuracy of the estimators either becomes too variable or degenerate.

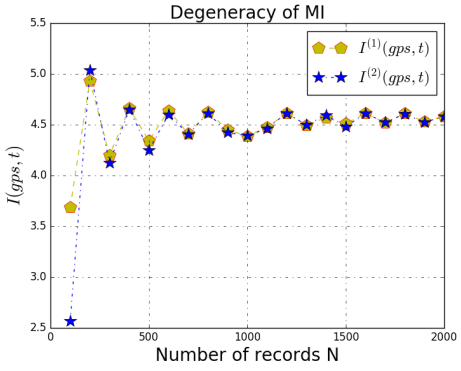


Fig. 8. Estimator Degeneracy Visualisation

Machine learning as a tool for increased automation [24] is well discussed along with the implications for its use in a legal sphere, eg: contract analysis [25], [26]. The application of this to the decision whether a given dataset is anonymised enough depends upon successful construction of the above mentioned structures.

This then leads us to how the decision of whether a dataset has been anonymised and how the decision or classification function generated. We have defined the existence and properties of such as function only in the classifier function χ_L. Anonymisation must monotonically decrease the information content of a dataset, given this we can simply render the classification function χ_L as a composition (the categorical product - or limit even) of all slices through the manifolds created by the variance in parameters of that anonymisation function over the pairs of structures that exist in the dataset being anonymised.

Meaning, in the case study example here we have three structures: identifier, locations and timestamps. This gives us 3 structures as can be seen in the matrix in Eq. (13)[1], ie: $\{ID \times LOC, ID \times TS, TS \times LOC\}$. For each of these we create a classifier function χ_L and thus compliance with a given legal characteristic is defined that all three possible correlations must have mutual information below a certain threshold as defined by the composition of those functions.

We can then characterise a legal text, such as the GDPR as:

$$\text{GDPR} \longrightarrow \prod_{s \in structPairs(\mathcal{D})} \chi_\ell \tag{15}$$

The challenge is therefore not the fact that such a mapping can or can not be made, but rather actually picking out the parameters that decide whether the mutual information of a given structural pair is below a certain limit. If this is achieved, and by trial and error it is possible in simple cases, then we can reduce the choice of anonymisation algorithm and parameters to those algorithms to an optimisation problem. A simple choice here might be to pick the datasets that correspond to the average mutual information values over the characteristic functions χ_L, χ_U etc that are used to pick compliant and useful datasets.

ACKNOWLEDGEMENTS

This work was partially funded by the Scott Project - Secure Connected Trustable Things - under EU Grant Agreement 737442.

[1]3 is because machine learning estimators become increasingly difficult to utilise when there 3 or more parameters to the correlation

VI. CONCLUSIONS FUTURE WORK

The example and structure already show demonstrate that a metric for privacy, a structure for anonymisation and classifier functions generated of particular privacy aspects can be constructed. Having these presents the first step in the process of metricising and automating the decision of whether a given dataset is sufficiently anonymous. Deciding whether a dataset is useful is relatively trivial; deciding whether it is legal is exceptionally difficult in that we still do not have either a good notion of context of a dataset, ie: the situations in which it is to be used and how re-identification/deanonymisation might take place, and how one might even construct a function such as χ_L from a text such as the GDPR. Here much work remains regarding the characterisation of what 'personal data' is and a proper ontologisation of such a concept along with the contexts in which that term is being used [27].

Our experiments have concentrated on utilising mutual information within a single dataset and in the direction of the anonymisation function. There are still a number of restrictions that need to be addressed such as properly characterising the nature of the ordering between result spaces: currently we can only do this on a per value basis. Work has been made on the use of eigenvectors and the magnitude of matrices in areas such as ecology and biology.

Another important restriction we currently face is that the structures of the input and output datasets are the same. This means that suppression function can not be adequately analysed at this time. This then has implications towards measuring the amount similarity between datasets which is important for constructing a measure of the degree of re-identification possible.

The structure of the boundary values above is relatively simple, in that they are normalised typically to $\mathbb{R}_{[0,1]}$ and therefore between individual values comparison is simple. However simply denoting the comparison of two matrices of the forms above as a comparison of values within those matrices is naïve and does not fully capture the true nature of the decreasing nature of the mutual information. We are investigating the use of the magnitude of mutual information correlation matrices, their eigenvectors and more advanced probabilistic metrics [28], [29].

Further work continues on the overall semantics with particular emphasis being made on the nature of anonymisation function as visualised in figure 7. There is a relationship between how these trajectories generalise over many manifolds and how characterisation functions act. We are currently exploring ideas from homotopy type theory [30] in order to better understand the relationship between, say, legal compliance and the behaviour of such functions.

In summary, we have shown that metricisation is possible

with at least the metrics we have shown here and that encoding notions of usability and legality over the result metric spaces is feasible if complex. This complexity may prove to be the human's advantage when it comes to properly understanding privacy and compliance of complex information sets under automation.

ACKNOWLEDGEMENTS

This work has been partially funded by EU ECSEL Project SECREDAS (Grant Number: 783119) and EU Horizon 2020 Project SCOTT (Grant Number: 737422).

REFERENCES

[1] E. Commission, "European Commission's press release announcing the proposed comprehensive reform of data protection rules," 25 January 2012.

[2] T. D. Breaux and A. I. Anton, "Analyzing regulatory rules for privacy and security requirements," *IEEE Transactions on Software Engineering*, vol. 34, no. 1, pp. 5–20, 2008.

[3] R. Rodrigues, D. Barnard-Wills, D. Wright, P. De Hert, V. Papakonstantinou, L. Beslay, E. JRC-IPSC, N. Dubois, and E. JUST, "Eu privacy seals project," *Publications Office of the European Union*, p. 19, 2013.

[4] "Nist privacy engineering objectives and risk model discussion draft," http://www.nist.gov/itl /csd/upload/nist_ privacy_engr_objectives_risk_model_discussion_ draft.pdf, April 2014.

[5] D. Rebollo-Monedero, J. Parra-Arnau, C. Diaz, and J. Forné, "On the measurement of privacy as an attackers estimation error," *International journal of information security*, vol. 12, no. 2, pp. 129–149, 2013.

[6] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Generation Computer Systems*, vol. 33, pp. 53–63, 2014.

[7] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[8] P. M. Schwartz, "German and us telecommunications privacy law: Legal regulation of domestic law enforcement surveillance," *Hastings LJ*, vol. 54, p. 751, 2002.

[9] J. Krumm, "Inference attacks on location tracks," *Pervasive computing*, pp. 127–143, 2007.

[10] K. Zheng, Z. Yang, K. Zhang, P. Chatzimisios, K. Yang, and W. Xiang, "Big data-driven optimization for mobile networks toward 5g," *IEEE Network*, vol. 30, no. 1, pp. 44–51, 2016.

[11] I. Oliver and S. Holtmanns, "Aligning the conflicting needs of privacy, malware detection and network protection," in *2015 IEEE TrustCom/BigDataSE/ISPA, Helsinki, Finland, August 20-22, 2015, Volume 1*. IEEE, 2015, pp. 547–554. [Online]. Available: http://dx.doi.org/10.1109/Trustcom.2015.418

[12] I. Oliver and Y. Miche, "On the development of a metric for quality of information content over anonymised data-sets," in *Proceedings of Quatic 2016*, 2016.

[13] J. Baez, T. Fritz, and T. Leinster, "A characterization of entropy in terms of information loss," *Entropy*, vol. 13, no. 11, pp. 1945–1957, 11 2011.

[14] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM Sigmod Record*, vol. 29, no. 2. ACM, 2000, pp. 439–450.

[15] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun 2004. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevE.69.066138

[16] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1849–1857.

[17] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of R\'enyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *ArXiv e-prints*, Mar. 2010.

[18] Y. Miche, I. Oliver, S. Holtmanns, A. Kalliola, A. Akusok, A. Lendasse, and K.-M. Björk, "Data anonymization as a vector quantization problem: Control over privacy for health data," in *International Conference on Availability, Reliability, and Security*. Springer, 2016, pp. 193–203.

[19] B. Schneier, "Architecture of privacy," *IEEE Security & Privacy*, vol. 7, no. 1, p. 88, 2009.

[20] J. Brickell and V. Shmatikov, "The cost of privacy: destruction of data-mining utility in anonymized data publishing," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 70–78.

[21] *Privacy Enhancing Technologies (PETs)*, Memo/07/159 ed., European Commission, Brussels, May 2007.

[22] *Anonymisation: Managing Data Protection Risk Code of Practice*, The Information Commissioner's Office (UK), November 2012.

[23] H. Zang and J. Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 2011, pp. 145–156.

[24] J. M. Wing, "Computational thinking," *Communications of the ACM*, vol. 49, no. 3, pp. 33–35, 2006.

[25] G. Lame, "Using nlp techniques to identify legal ontology components: concepts and relations," in *Law and the Semantic Web*. Springer, 2005, pp. 169–184.

[26] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[27] I. Oliver, *Privacy Engineering: A Data Flow and Ontological Approach*. CreateSpace Independent Publishing, July 2014, 978-1497569713.

[28] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 03 1951. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177729694

[29] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Les Mémoires de Mathématique et de Physique, Année 1781*, 1781.

[30] T. Univalent Foundations Program, *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study: https://homotopytypetheory.org/book, 2013.