A Measure of Personal Information in Mobile Data

Ian Oppermann
Dept. Customer Service
NSW Government, Australia
Ian.oppermann@customerservice.nsw.gov.au
University of Technology Sydney
Ian.oppermann@uts.edu.au

Jakub Nabaglo
Triplebyte
San Francisco, USA
jakub@triplebyte.com

Wilko Henecka
CSIRO's DATA61
Sydney, Australia
wilko.henecka@data61.csiro.au

This paper describes fundamental aspects of a framework for privacy-preserving data sharing in a mobile context. The principal technical challenge is measuring the level of personal information (PI) in datasets that are shared for the delivery or enhancement of mobile enabled services. Another challenge is determining the threshold delineating a "reasonable likelihood" of an individual being identifiable from the data. The risk of reidentification defines personally identifiable information (PII). The measure of PI must go beyond simply analysing personal attributes captured in data and consider preference revealed through use of services, temporal and spatial aspects of data, as well as context for use of services. Keywords—data sharing, privacy, mobile services

I. INTRODUCTION

6G is expected to advance connectivity for a myriad of personal devices, sensors, and autonomous applications. This breakthrough will enable smart services for homes, workplaces, cities, and governments. These services in turn will rely on sharing of large volumes of, often personal, data between individuals and organisations, or between individuals and governments. A smart light in your home, which turns on and off as you move around the house, can provide a more efficient use of energy for lighting, but will use deidentified data about when you are home, which rooms you use and when, if there are other people in your home, and where in your home you spend your time. Within this deidentified data, there are insights about you and your relationships, habits, and preferences. In aggregate form, this data can be used by a smart lighting provider to deliver more efficient lighting services to a suburb, by a smart grid to match energy demand to energy supply, or by a smart micro energy service provider to make best use of spot energy prices. The benefit is the ability to locally optimise or individually personalise services based on personal preference, as well as to optimise the wider network for users and providers. This idea of ubiquitous computing covers an ever-increasing range of "smart" devices (phone, TV, scales, toilet, refrigerator, watch) and "smart" environments (home, workplace, city). Increased efficiency, improved effectiveness, and greater personalisation can present enormous benefits. If these datasets are linked, the resulting merged dataset may contain a great deal of PI, possibly enough to reidentify the individuals represented therein. How this data is used, by whom, and for what purposes determines risks and concerns. In many economies, it may force service providers and operators to rethink governance models to support regulatory requirements such as GDPR [1].

With the multidimensional flows of rich data, the challenge is quantifying the amount of PI in a dataset at any point in time and in any given context, and developing threshold tests for when an individual is reasonably identifiable, while considering personal attributes, temporal and spatial aspects of data, and rich contextual environments.

II. PI VERSUS PII

The terms *personal information* (PI) and *personally identifiable information* (PII) are often used interchangeably in legislative frameworks as well as in different technical literature. PI is typically described in a way that covers a very wide field, and this description varies in different parts of the world. For example, in the Australian state New South Wales:

"... personal information means information or an opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is apparent or can reasonably be ascertained from the information or opinion" [2 Section 4].

For example, date of birth is considered PI but not PII as it cannot uniquely identify an individual. Similarly, spatial location or time of service use will not uniquely identify anyone, except in isolated cases. The question becomes, how many features must be linked before PI becomes PII for an individual known to be in a dataset? Context and rarity play important roles in the answer to this question.

The legal tests for PI generally relate to the situation where an individual's identity can *"reasonably be ascertained".* The definition is very broad and in principle covers any information that relates to an identifiable individual, during their lifetime or for decades after their death [2]. A recent paper published in Nature Communications [3] provides a means to "*estimate the likelihood of a specific person to be correctly re-identified, even in a heavily incomplete dataset"*. The paper is part of a long series showing that only a small number of features need to be linked to identify an individual from a population.

The focus of this paper is a quantitative measure and risk framework for "reasonably" in different contexts. The ACS whitepaper [4] on which this work builds explored a *personal information factor* (PIF) which is a measure of the PI contained in a linked, deidentified dataset or in the outputs of its analysis. A PIF above a certain threshold (for example, 1.0) means sufficient PI exists to identify an individual: this reidentification risk makes this PII. A value of 0 means there is no PI. It is important to note that the PIF envisaged is not a technique for anonymisation; rather, it is a heuristic measure of potential risk of reidentification and of the amount of information which would be revealed from reidentification.

The PIF for both data and the outputs of any analysis based on this data are described in [4] using:

- A measure of the information content of the dataset or the output of the analysis of the data
- The smallest unique group in the dataset or output
- Additional information required to identify an individual from data or output ("epsilon").

Figure 1 shows the context for evaluating the degree of PI as part of assessing the PIF in a closed system. The data available in a closed (sealed) environment is finite and a PIF can be described mathematically based on uniqueness of feature combinations describing individuals. Algorithmic functions may contain embedded extrinsic knowledge such as known probabilities of occurrence of certain values or features. This could increase the PIF of outputs beyond the PIF of the dataset analysed. This is ignored for simplicity.

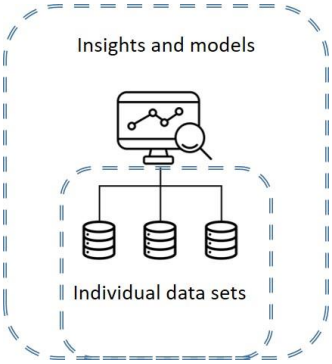

Figure 1. Closed system context for evaluating PIF.

The degree of PI contained in data may be very high (a unique identifier such as a social security number), moderate (surname), low (eye colour) or very low (month of birth). It is expected that the PIF in a linked dataset will generally increase as more datasets are linked. Conceptually shown in Figure 2, as more datasets containing PI are linked, a point may be reached where an individual is personally identifiable (a PIF of 1), or "reasonably" identifiable (a PIF within "epsilon" of 1). The dataset is then considered to have PII. The "epsilon" is an indication of the difference represented by the gap before the "reasonable" threshold is met.

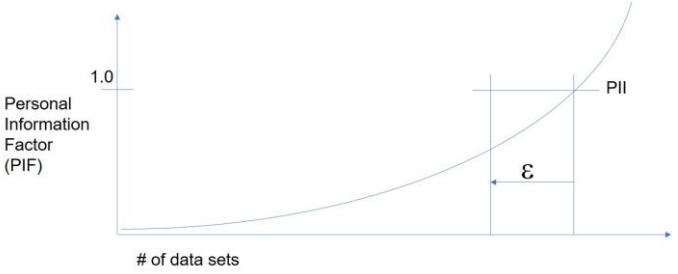

Figure 2. Conceptualisation of a PIF and PII threshold point.

Figure 3 shows the context for evaluating the level of PI when outputs are observed by individuals who have access to data and outputs. The individuals could have access restricted based on relationship to the data, motivation, experience, or expertise. However, each observer has their own knowledge of the world or may have an unexpected connection to the dataset leading to risk of spontaneous reidentification.

When data or analysis outputs are released as "open data", there is no control over who accesses them or which additional datasets they are combined with. Reidentification may require an observer to expend effort or resources to gain additional information beyond what is available from the outputs of analysis. In this environment, there is no way to guarantee the resulting data does not contain PII.

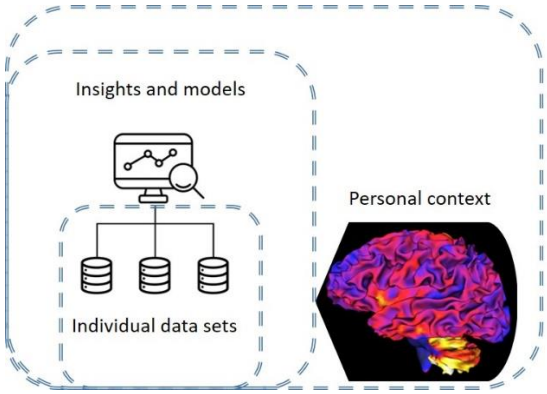

Figure 3. Human context for evaluating PIF.

A. PIF versus Risk of Reidentification

The premise of the PIF is that linking sufficient fields with PI, a threshold will be reached by which an individual can be uniquely (or reasonably) identified (Figure 2).

Starting with the closed system context for evaluating PIF (Figure 1), and representing individuals as rows in a dataset, with columns as features, the ability to identify an individual means that there are sufficient differences in combinations of column values such that there is at least one unique row *and* there is sufficient information in the columns of that unique row to unambiguously (or reasonably) identify that individual. In this context, if there are two or more rows with all the same column values, then there is not sufficient information in the columns to uniquely identify an individual. In this case, the minimum identifiable cohort size (MICS) is greater than one. If there is a unique row, but insufficient information contained in the columns, then the individual can still not be unambiguously (reasonably) identified.

When considering the human context for reidentification (Figure 3), the knowledge or experience of the observer could increase the total information about a unique row to the point where an individual could be reasonably identified even if there is insufficient information in the columns of the dataset. The observer would need to have knowledge of the members of the dataset (at an individual level or a population level) for this to occur. If two or more rows had the same values, then the observer would still not be able to distinguish exactly which individual was referred to by each row.

It is worth emphasising that the discussion on risk of reidentification differs from the risk of gaining information about an individual if they are ***known*** to be represented in a dataset – the "homogeneity attack". For example, if a dataset is created of streaming adult video services, and a particular individual is known to be in the dataset, an observer (Figure 3) will learn that the individual uses streaming adult video services. The observer may learn more if the cohort the individual is in has more features. In the homogeneity attack, the individual is known to be in the dataset, so the risk of reidentification is "certain". What is not certain is the additional information gained from observing the features which were not previously known. The unknown features depend on the particular observer.

When data is released openly, there is no control over which other datasets or other information sources it can be joined with. With open data, there is no absolute protection against reidentification. Relative protections may be applied by limiting the amount of PI released, thus reducing the information gained by someone seeking to identify and individual. Another strategy is to increase the effort required to reidentify by making the minimum identifiable cohort size (the smallest number of rows with the same values) a relatively large number.

III. Formalising Information Gain

A promising approach is to base the PIF on a quantitative measure of the information an attacker might gain about a particular individual by accessing a dataset. The approach is based on information theory and is summarised in Figure 4.

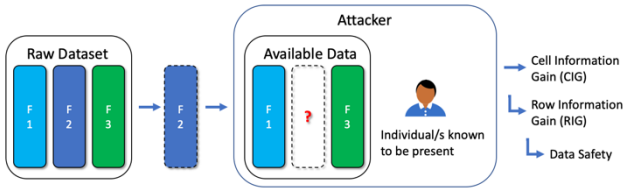

Figure 4. Attacker / information gain approach.

The approach is motivated by the fact that every dataset released into the wider world is available to an "attacker" seeking to reidentify an individual represented in a dataset.

It is important to note that not every reidentification event is equal in terms of the PI revealed. For example, learning that a mobile user's location at city suburb reveals less PI than learning the exact street address. The PIF framework as presented computes the potential information gain for each field in the dataset as well as considering the protection afforded by the size of the MICS.

It is not always necessary to reidentify an individual in order to learn information about them. Consider a scenario where an attacker wishes to learn an individual's birthday but is only able to select three rows of the dataset, any one of which could represent the target. Narrowing the choice of birthday to three options itself constitutes a significant gain in information. Furthermore, if all three possible rows list the same birthday, the attacker has learned the target's birthday with certainty, despite not being able to reidentify them. By focusing on information gain, and not solely reidentification risk, we are able to account for both these scenarios.

For simplicity, the approach assumes that each row of a dataset represents an individual, and each column represents a feature. The framework then allows any holder of data to:

- consider risks on a per-feature (column) basis,
- find individual risk of each individual (row),
- identify comparatively high-risk individuals,
- prioritise anonymisation efforts to focus on the most vulnerable features and individuals, and
- compare the performance of different anonymisation strategies.

A. Using a Threat Model

The approach uses a model from cryptography to formalise the threat from an attacker. An "attacker" is an individual who has access to the dataset and to additional information about an individual they are seeking to reidentify. By locating and reidentifying an individual in the dataset, the attacker seeks to learn more about them. Knowing the information and resources (the strength) of an attacker is difficult as the auxiliary information available to the attacker is unknown. Consequently, the approach assumes different strength of attackers when access to data and results are controlled by technology and process (Figure 3) and a very strong attacker when there is no restriction on access to data or processing resources (open data).

A very strong attacker is characterised as knowing every feature of an individual aside from the one they are attempting to find. Less strong attackers are described as those who know some but not all features, or that they are not fully certain in the information that they have.

B. Quantifying Information Content

In the human context for evaluating information gain (Figure 3), the information gained from a feature value is inversely related to how much you expect it. A highly unexpected value conveys a great deal of information. In the closed system context (Figure 1), we can state more formally that the number of bits of information gained from a feature value is the negative logarithm (base 2) of the probability of that value occurring[1,2]. We can build on this concept by using KL (Kullback–Leibler) divergence calculation to produce a measure referred to as the Cell Information Gain (CIG), the Row Information Gain (RIG) and the Feature Information Gain (FIG).

C. KL Divergence of Probability Distributions

A probability distribution is a set, possibly infinite, of possible values of a feature, along with the probability of occurrence of each value. In the human context (Figure 3), it may represent our belief about a feature that exists but that we do not know for certain. For example, if an attacker has no knowledge about an individual's birthday, they may assign an equal probability (1 in 365) to every day of the year. the attacker then updates their beliefs and their probability distributions with new information. If they learn that the person's birthday is in August, they may update their belief that every day in August has an equal probability (1 in 31) of being the true birthday, and every other day of the year having a probability of zero. Our original probability distribution is the *prior* and the updated one is the *posterior*.

In the closed analytical context, the KL-divergence measures the information gain, in bits, when evaluating the difference between prior and posterior probability distributions. Mathematically given by:

[1] See for example R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, 2014. Available online http://ee.stanford.edu/~gray/it.pdf

[2] Importantly, another fundamental of information theory states that additional processing of data will not create additional information beyond what is already present. This is an important consideration when considering the limits of analytical models.

$$D_{KL}(P \parallel Q) = -\sum_{x \in \chi} p(x) \log q(x) + \sum_{x \in \chi} p(x) \log p(x)$$

where Q is the prior distribution, P is the posterior distribution, X is the set of possible values, x is an element thereof, $q(x)$ is the probability of x according to Q, and $p(x)$ is the probability of x according to P. Learning that the birthday is in August provides 3.6 bits of information gain[3]. If instead the person's exact birthday was learned, the posterior would represent 8.5 bits of information gain[4]. If nothing is learned, then the distribution is not updated, and the KL divergence is zero. Conversely, if information is gained which had previously considered impossible (probability zero), the information gain tends to infinity.

When discussing reidentification risk, we consider the belief an attacker has about the targeted individual's attributes. The prior distribution represents the attacker's belief before obtaining the dataset. The posterior is what the attacker has been able to find by combining existing information about an individual with information in the dataset. Following the example, the prior belief that the target's birthday can be any day of the year, we learned that their birthday is in August, obtaining a posterior belief. Despite not learning the exact date of the target's birthday, information is still gained. The KL divergence can therefore be used to quantify information gain when the attacker does not become fully confident of a feature's value, merely more confident.

D. Cell Information Gain (CIG)

For simplicity, it is assumed that every cell belongs to a row, and every individual is represented by exactly one row. The CIG quantifies the information gain associated with each feature for each individual. Considering a strong attacker scenario, for each cell for which we wish to determine the CIG, we assume the attacker knows every other feature of the target. The CIG value is then defined as the KL-divergence of the attacker's prior and posterior distributions for the true value of that feature. The prior is the attacker's probability distribution for this attribute of the target before they obtain the dataset. Typically, only the attacker has access to this prior. This prior can be approximated from the dataset by tallying the occurrences of every possible value of this feature across the entire dataset.

The posterior can be calculated in a similar manner. Once again, we assume there is a particular individual the attacker is targeting, and the dataset has a row of features for this person. For every person, or row, in the dataset we assign a probability that they are the person the attacker is seeking to reidentify. For example, in the strong attacker model, the attacker knows every feature of the target except the one they wish to find. In this scenario, if the feature values for a row do not match the attacker's knowledge, the probability of that row representing the target is 0; the probabilities of the remaining rows are equal and sum to 1. For every possible value of our cell, we tally the occurrences of the people (rows) who have this value, weighing each row by the probability that it represents the target. The KL divergence of this calculated posterior and the prior gives the CIG in bits.

E. Feature Information Gain (FIG)

By summing the CIG values for each feature, we can determine the FIG for that feature in bits. The FIG can be used to identify the features that are the highest information gain (and so risk of reidentification) in a dataset. The risk of inclusion can be compared to the feature's utility when making the decision to include or exclude it.

F. Row Information Gain (RIG)

In a similar way, the RIG is determined by summing all the CIG values in the row and is a measure of the information gain associated with a particular individual if their information is revealed through reidentification.

G. Calculating Information Gain in Practice

Steps for calculating the information gain values:

- Estimate the prior and posterior distributions for each row and feature
- Calculate CIG values using KL divergence
- Sum the CIG values per row to determine the RIG and per column to determine the FIG values
- Analyse RIG and FIG values to inform protection

High FIG features or high RIG rows may be targeted for suppression, aggregation, or other forms of protection to reduce information gain when data is shared or released. Consider an example Hospital Admissions dataset [5] with CIG values shown in Figure 5 (see example datasets at end). All rows have large information gain for the "POSTCODE" feature making it relatively high-risk to include if released or shared. Row 6 also has relatively large information gain for the "job" feature making this individual relatively high-risk to include if the data is released or shared.

	gender	AGE	POSTCODE	blood_group	eye_color	job
0	0.736966	3.50706	6.44937	3.00353	2.33085	3.50535
1	1.32193	3.52638	7.09423	2.99185	2.31117	3.50535
2	1.32193	3.57562	7.48889	2.97789	2.31117	4.12917
3	1.32193	3.57562	5.98571	2.97789	2.32444	4.38644
4	1.32193	3.54684	9.72346	3.01561	2.33085	4.85394
5	1.32193	3.51905	4.43354	2.99185	2.33236	2.60658
6	0.736966	3.52638	9.6397	2.97789	2.33085	10.0812
7	1.32193	3.56803	4.50407	2.00597	2.33236	2.60658
8	0.736966	3.52638	6.6449	2.99185	2.31098	3.44733
9	1.32193	4.26248	7.38518	2.97996	2.33236	4.79756
10	1.32193	3.52768	6.98528	2.97996	2.33236	3.82016
11	1.32193	3.54684	5.70362	3.01168	2.33085	3.82016
12	1.32193	3.54684	6.96602	2.97789	2.31117	3.44733
13	0.736966	3.57562	9.813	2.99185	2.33236	4.85394
14	1.32193	3.54684	6.35362	2.99185	2.33085	4.38644
15	0.736966	3.56803	3.97857	2.99185	2.31098	2.60658
16	1.32193	3.51518	6.457	2.97996	2.31098	4.74733
17	1.32193	3.50706	7.8409	1.99582	2.33236	3.50535
18	0.736966	3.51905	4.28773	2.97789	2.31098	2.60658
19	0.736966	3.54684	10.2497	3.01561	2.31117	5.157

Figure 5. CIG values for hospital admissions dataset.

[3] The probability changes from 1/365 (all equally likely) to 1/31. The information gain is $\log_2(365/31) = 3.6$ bits.

[4] The probability changes from 1 / 365 to 1, so the information gain is $\log_2(365) = 8.5$ bits.

H. Linking PIF to Reidentification Risk Within a Dataset

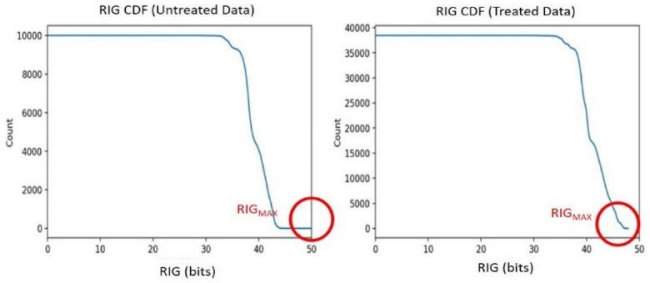

Figure 6. RIG score reverse-order cumulative histogram.

Figure 6 shows the distributions of individuals' RIG levels in two datasets. The vertical axis shows the count of individuals with RIG levels higher than each RIG threshold. The dataset on the left has a lower average for all RIG values than the one on the right. However, the left-hand dataset has a small number of individuals who are at elevated RIG levels shown by the long tail in the bottom right of the plot.

Despite having a lower average of all RIG values, the dataset on the left has a higher absolute risk of reidentification—the risk that any individual will be reidentified—due to a number of high information gain rows (high RIG individuals). Depending on how the data is treated—suppressing, perturbing or aggregating values in cells—a lower risk data product can be created from this dataset without significantly impacting the majority of cells and rows.

By defining the quantity RIG_{95} as the 95th percentile of all of RIG values, it is possible to characterise the reidentification risk of the entire dataset in a single number. Similarly, we can define RIG_{max} as the RIG of the individual about whom the greatest amount of information would be revealed if reidentified. Finally, the PIF for the dataset is driven by both the minimum identifiable cohort size (MICS) and the amount of information which would be revealed if individuals in this cohort were reidentified. Whilst heuristic, the working definition of PIF is

$$PIF = max_{\gamma \in \Gamma}\left(\frac{RIG_{\gamma}}{MICS_{\gamma}}\right)$$

where Γ is the set of RIG values for a dataset, and $MICS_{\gamma}$ is the minimum identifiable cohort size at RIG value γ - the number of rows with all the same column values. If the number of rows at RIG_{max} is 1, then the PIF is equal to RIG_{max}. If the number of rows at RIGmax is 2, and there are no other unique rows in the dataset, then the PIF is RIG_{max} / 2. If there are unique rows at lower RIG, the PIF may be set by this lower RIG value. This measure of PIF reduces the risk of reidentification created by high RIG values by relying on the protections of k-anonymity[5]. Whilst k-anonymity has known weaknesses, it does afford protection if it is also assumed the individuals in the dataset are not known to be in the dataset.

IV. Extending the Information Gain Framework

A. Incorporating Broader Knowledge of Population

If more information is known about the distribution of a particular feature in the entire population rather than just the dataset, it is possible to base KL-divergence measure on these extended priors rather than on the dataset alone. This potentially allows for the data safety of low coverage datasets with unique values to be more appropriately measured. Figure 7 shows the CIG for elements of the hospital admission dataset without (LHS) and with (RHS) knowledge of the distribution of the feature *"icd_code"*.

Similarly, when creating datasets with lower levels of PI, incorporating prior knowledge of how features are distributed across a population allows the approach to take into account broader knowledge about the data and reduce the impact of sampling on safety assessment. The technique for calculating CIG described is agnostic to the kind of anonymisation used.

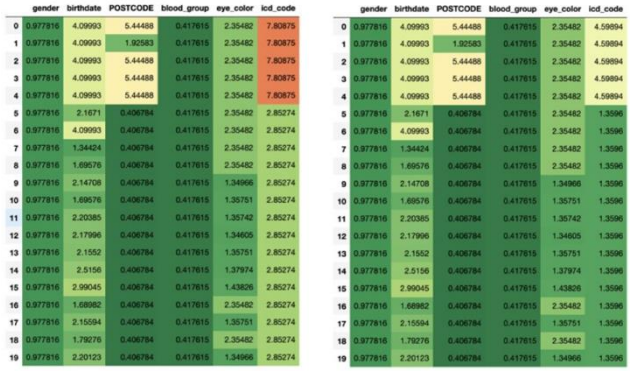

Figure 7. Improved CIG using knowledge of population.

B. Modelling Different Attacker Capabilities

The conservative approach is to assume the strongest possible attackers, that is, they known every feature of the person they are attempting to reidentify except from the one feature they are attempting to find. This approach is relevant for data that is released openly, as, once released, it is impossible to control who accesses the data. Nonetheless, different models for the attacker are also possible. In one model, it is possible to assume the attacker knows *n* features of the individual they are targeting. The feature they are attempting to find is not one of those *n*. Reasonably, an attacker that has less information about the person to begin with has less chance at reidentifying them. This is reflected by lower CIG (and consequently FIG and RIG) scores across the dataset.

Another model assumes that they have some information but are not fully confident that it is correct. The level of confidence is a parameter that forms part of the assumptions in the approach. Intuitively, if we assume that only carefully selected individuals are permitted to view the shared dataset, we may model the attacker as less powerful. This lets our safeguard be reflected in the reidentification risk calculation.

[5] See for example Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"

C. A Major Challenge – Dealing with Trajectories

The discussion of PIF or information gain has a tendency to look at "human" features as being the key to identifiability. This view fails to recognise the impact of *spatial*, *temporal*, and *relationship* features increasing the possibility of reidentification by an attacker. The solution is not always just to aggregate, as time and space have more dimensionality, giving more context that may allow reidentification of individuals, or linkage to other datasets. The trajectory approach considers the combined impact of relationship, spatial and temporal features.

The *trajectory* of an individual is defined as the set of all rows pertaining to this subject (presumably linked by a data-set identifier or study ID), which describes the longitudinal journey of this individual and their interactions with the dataset. In the same way as the processing operations on data may contain extrinsic information which increases the PIF, it is possible that there may be identifiable properties of the trajectory itself which increases the PIF. For simplicity, it is assumed that this is not the case. In a longitudinal dataset, each trajectory may be unique (many variables in space-time), and highly different so each individual is potentially identifiable.

D. Time, Space, Personal and Relationship Features

Mobile services and network interactions potentially capture fine grained location information, timestamp information, patterns of use and relationship information. Data protection such as aggregation, suppression or perturbation can be applied either equally to the entire dataset, or preferentially to temporal, spatial, personal or relationship features with the intention to maintain utility in one or more of these feature domains whilst preferentially protecting features in the other domain (and so reducing Utility of the data in these domains). Developing standard aggregation, suppression or perturbation approaches in each of these domains would assist when analysing data from different sources.

E. Mutual Information (MI) as a measure of Utility

When releasing or sharing a dataset two major, almost certainly competing considerations, in utility and privacy. The MI of two random variables is a measure of the mutual dependence between the two variables. It quantifies the information (in bits) obtained about one random variable through observing the other random variable. The concept of MI is intricately linked to that of entropy, a fundamental notion that quantifies the expected "amount of information" held in a random variable. Not limited to linear dependence like the correlation coefficient, MI is more general and determines how similar the joint distribution of the pair (X,Y) is to the product of the marginal distributions of X and Y. MI is the expected value of the pointwise mutual information (PMI) and is known as information gain (or loss).

The define the relative Utility μ of a generated dataset, the MI can be normalised to values between 0 and 1 by dividing it against the mutual information of the original dataset. A relative Utility of 1 implies no loss compared to the original dataset. A relative Utility of 0 implies complete information loss in the resultant dataset. Figure 8 shows an example of relative Utility declining as a feature "age" is aggregated into 2-year, then 5-year and then 10-year bins in a dataset.

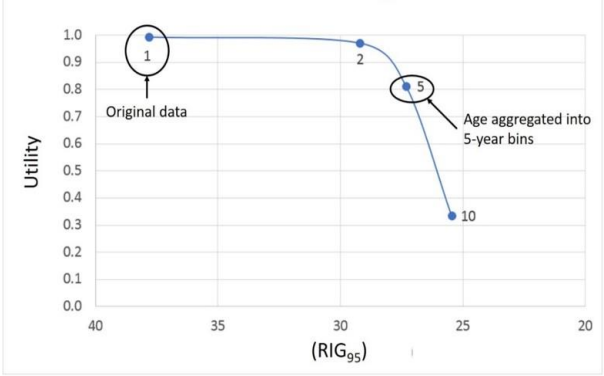

Figure 8. Utility and RIG95 as "age" is aggregated.

V. CONCLUSIONS

This paper has highlighted some of the challenges of creating a measure of PI for linked data used for analytics purposes. Future mobile enabled services will compound the complexity of these simple frameworks as greater fidelity of captured personal attributes, greater resolution in time and space, and richer contextual considerations are harnessed to deliver services or optimise network performance. The paper highlights efforts to explore the fundamental challenge of reidentification risk in deidentified data through a PIF. An understanding of the degree of PI in potentially highly time varying, context rich datasets, will challenge what we mean by deidentified data and the risk of reidentification. Understanding these challenges will provide a basis to anchor principles-based data sharing and governance frameworks to help ensure new mobile enabled services operate within the regulatory frameworks designed to protect us all.

ACKNOWLEDGMENTS

Thanks to Peter Chui and Michael Kam of the NSW Data Analytics Centre (DAC) for trialling a version of the PIF tool.

REFERENCES

1. The General Data Protection Regulation 2016 / EU679. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
2. NSW Privacy and Personal Information Protection Act 1998
3. L. Rocher, J. M. Hendrickx and Y. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models", Nature Communications, Jul. 2019. https://www.nature.com/articles/s41467-019-10933-3
4. I. Oppermann (editor), "Privacy-Preserving Data Sharing Frameworks", Dec. 2019, Australian Computer Society. https://www.acs.org.au/insightsandpublications/reports-publications/privacy-preserving-data-sharing-frameworks.html
5. Hospital Admissions Dataset (Synthetic dataset) see [4]
6. Inmate Admissions Dataset (US Open Dataset)
7. Inmate admissions with race, gender, legal status, top charge.

   https://data.cityofnewyork.us/Public-Safety/Inmate-Admissions/6teu-xtgp