University of Kent

Kent Academic Repository

Lu, Yang, Sinnott, Richard O., Verspoor, Karin and Parampalli, Udaya (2018) ***Privacy-Preserving Access Control in Electronic Health Record Linkage.*** **In: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). . pp. 1079-1090. IEEE ISBN 978-1-5386-4389-1. E-ISBN 978-1-5386-4388-4.**

Downloaded from

https://kar.kent.ac.uk/80959/ The University of Kent's Academic Repository KAR

The version of record is available from

https://doi.org/10.1109/TrustCom/BigDataSE.2018.00151

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal* , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering

Privacy-Preserving Access Control in Electronic Health Record Linkage

Yang Lu
School of Computing and
Information System
University of Melbourne
Melbourne, Australia
luy4@student.unimelb.edu.au

Richard O. Sinnott
School of Computing and
Information System
University of Melbourne
Melbourne, Australia
rsinnott@unimelb.edu.au

Kain Verspoor
School of Computing and
Information System
University of Melbourne
Melbourne, Australia
karin.verspoor@unimelb.edu.au

Udaya Parampalli
School of Computing and
Information System
University of Melbourne
Melbourne, Australia
udaya@unimelb.edu.au

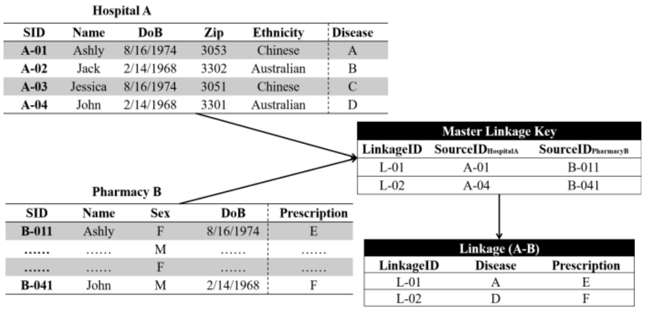
***Abstract*—Sharing aggregated electronic health records (EHRs) for integrated health care and public health studies is increasingly demanded. Patient privacy demands that anonymisation procedures are in place for data sharing. However traditional methods such as *k-anonymity* and its derivations are often over-generalizing resulting in lower data accuracy. To tackle this issue, we present the *Semantic Linkage K-Anonymity* (*SLKA*) approach supporting ongoing record linkages. We show how *SLKA* balances privacy and utility preservation through detecting risky combinations hidden in data releases.**

***Keywords*—record linkage; privacy preservation; k-anonymity; semantic technologies;**

I. INTRODUCTION

Electronic health records (EHRs) offer great opportunities for both healthcare and health research. Based on obtaining consent and ethical approvals, data custodians can share aspects of patient records for secondary use, e.g. in clinical trials and studies. To meet the increasing research demands, record linkage techniques are often adopted to integrate EHRs associated with the same entity (e.g. a patient) where the data originates from potentially different organisations. Linkage can be used for many purposes, e.g. at the population level to correlate obesity and socio-economic status [1] or to predict lung cancer coverage based on mortality statistics [2], amongst many other scenarios. By leveraging linkage techniques, a comprehensive profile of patients and populations can be produced for many in-depth studies [3].

Over the last decade, numerous health data linkage units have been established across Australia [4][5][6]. Through probabilistic matching approaches, records obtained from remote organisations can be structured and standardised based on agreed information [7]. Depending on the kinds of identifying information, various algorithms can be used to compare attributes resulting in a vector of numeric similarities [8][9]. With two thresholds (lower and upper thresholds) chosen for the matching step, record pairs are typically classified into three groups: *Matched*, *Non-Matched* and *Possible Matched*. To support highly accurate linkages, the "possible matched" group requires manual reviews. Finally, the linkage unit needs to keep the mappings of records across different organisations. For instance, Fig. 1 shows a typical linkage between hospital A and pharmacy B. Initially, each organisation provides their raw data, which is identified by pseudo names. Patient registered in both databases have more than one source identifier (SID), such as *A-01* and *B-011* both pointing to Ashly who is uniquely identified as *L-01* using a Master Linkage Key (MLK). To ensure the real-time and sustainable management at the linkage unit, it is essential to generate the MLK in advance.



Hospital A

SID	Name	DoB	Zip	Ethnicity	Disease
A-01	Ashly	8/16/1974	3053	Chinese	A
A-02	Jack	2/14/1968	3302	Australian	B
A-03	Jessica	8/16/1974	3051	Chinese	C
A-04	John	2/14/1968	3301	Australian	D

Pharmacy B

SID	Name	Sex	DoB	Prescription
B-011	Ashly	F	8/16/1974	E
……	……	M	……	……
……	……	F	……	……
B-041	John	M	2/14/1968	F

Master Linkage Key

LinkageID	SourceID$_{HospitalA}$	SourceID$_{PharmacyB}$
L-01	A-01	B-011
L-02	A-04	B-041

Linkage (A-B)

LinkageID	Disease	Prescription
L-01	A	E
L-02	D	F

Figure 1. Linkage$_{A\text{-}B}$ generated between *Hospital A* and *Pharmacy B*.

Linkage generation should not result in re-identification of individual patients. Therefore, a separation principle was proposed and implemented in current linkage centres [11]. By separating patient identifying data from actual health information, it can restrict the access to and use of demographic information by researchers (users). This could well protect patient privacy, however limit the application in public health studies where the personal features are sometimes necessary to learn. Alternatively, using the 'safe harbour' protocol defined in the US Health Insurance Portability and Accountability Act (HIPAA) only a set of specific identifying attributes (social security number, name, driver license ID etc.) need to be removed from health datasets to ensure privacy [12]. However, malicious users (attackers) can still narrow records to specific individuals based on combinations of key attributes. These key attributes are often called *quasi-identifiers* (*QIs*). For instance, it has been shown that 87% of the US population can be re-identified by combining such de-identified data sets [13]. To tackle this problem, a number of statistical disclosure control (SDC) approaches have been proposed to control the leakage chances to safer levels.

Typical SDC approaches including *k-anonymity* are defined to generalise *QI* values against a numerical constraint, i.e. any individual represented in an equivalent group (class) must be indistinguishable from at least *k-1* other individuals appearing in the same group [14]. In addition to identity protection, anonymity models such as *l-diversity* and *t-closeness* have been developed with emphasis on protecting sensitive attributes [15]. [16]. Furthermore, data published in distributed environments is also often protected using *QI* data sets [18] [19]. In addition to one-time request/disclosure scenarios, the composition attacks may arise if adversaries collectively use anonymous data

2324-9013/18/31.00 ©2018 IEEE
DOI 10.1109/TrustCom/BigDataSE.2018.00151

releases to seek out private information, i.e. in an on-going manner [20]. One solution is to keep sensitive attributes unchanged within a timespan [21], e.g. the *τ-safety* scheme introduces the temporal concept 'historical releases' for risk analysis since a disclosure may occur based on changing attribute combinations [22].

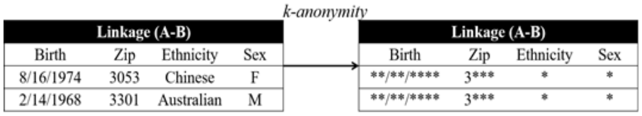
Although linkage jurisdictions achieve success in supporting health studies, there are challenges identified in governing linkage releases through the gradual erosion of privacy over time and the increased chances of potential re-identification. These should be considered and addressed while designing anonymising mechanisms for ongoing linkage release.

II. CHALLENGES OF LINKAGE DATA ANONYMISATION

A. Information Loss

Reducing information loss is critical when designing techniques for anonymisation of health data especially when the records contain both categorical data such as gender, geo-location information and numerical data such as age in years, length of stay-in hospital etc. Dankar and El Eman (2012) discussed the limitations of applying differential privacy methods to de-identify health information due to the addition of Laplace Noise to data that can cause significantly distorted results [23]. Non-perturbation approaches such as *k-anonymity* perform well based on the assumption that "end users may know patients who exist in the data set from all of their attributes". However the optimisation of *k-anonymity* is proven to be an NP–hard problem, hence *weak k-anonymity* was proposed for scenarios where the "presence" of individuals cannot be confirmed [24].

Problem scenario-1. Fig. 2 shows a linkage scenario[1] where the linkage is anonymised by using *k-anonymity* (*k=2*), assuming requestors have viewed full (linkage) content. However, this is not the case in many linkage applications where linkage is created and processed at trusted third party (TTP) sites. To ensure protection of data based on local regulations, linked data should not be directly released to stakeholders. Instead of direct generalisation to all attributes, local knowledge (Hospital A or Pharmacy B) is needed to evaluate the anonymisation of results. As shown in Fig. 3, the linkage released to Hospital A researchers is processed using *weak 2-anonymity*. For tuple *<8/16/1974, 305*, Chinese>*, two individuals (*Ashly* and *Jessica*) are possibly located at Hospital A and thus meet the privacy assumption, i.e. users should not be able to re-identify individuals from known entities if they are *2-anonymised*. With fewer *QI* attributes being generalised, the *weak k-anonymity* model outperforms standard *k-anonymity* in terms of preserving the utility of linked data. As adversaries are unsure of the given presence of the target in a dataset a less-transformed (obfuscated) linkage set can be retained and thus be of greater value for secondary use.



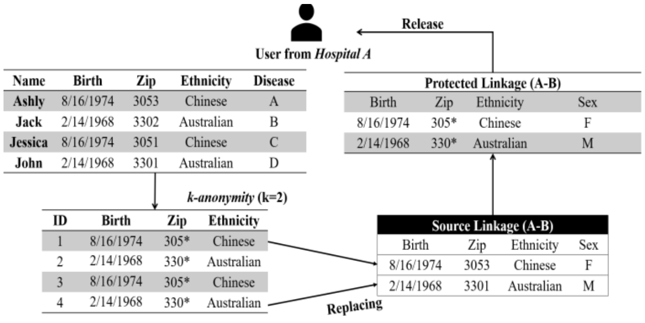
Linkage (A-B)

Birth	Zip	Ethnicity	Sex
8/16/1974	3053	Chinese	F
2/14/1968	3301	Australian	M

Linkage (A-B)

Birth	Zip	Ethnicity	Sex
//****	3***	*	*
//****	3***	*	*

Figure 2. Generalising linkage with *k-anonymity* (*k=2*).



Name	Birth	Zip	Ethnicity	Disease
Ashly	8/16/1974	3053	Chinese	A
Jack	2/14/1968	3302	Australian	B
Jessica	8/16/1974	3051	Chinese	C
John	2/14/1968	3301	Australian	D

ID	Birth	Zip	Ethnicity
1	8/16/1974	305*	Chinese
2	2/14/1968	330*	Australian
3	8/16/1974	305*	Chinese
4	2/14/1968	330*	Australian

Protected Linkage (A-B)

Birth	Zip	Ethnicity	Sex
8/16/1974	305*	Chinese	F
2/14/1968	330*	Australian	M

Source Linkage (A-B)

Birth	Zip	Ethnicity	Sex
8/16/1974	3053	Chinese	F
2/14/1968	3301	Australian	M

Figure 3. Generalising linkage with *weak k-anonymity* (*k=2*).

B. Inference disclosure

Regardless of background knowledge, privacy leakage can still happen due to inferences of the data based on knowledge and associated standards. To tackle these issues, SEMANTIC TECHNOLOGIES have been applied in various contexts where privacy risks need to be considered and minimised. THROUGH knowledge modelled in the web ontology language (OWL) and semantic web rule language (SWRL) rules, policies formalised with semantic meaning can be used to reason about "next-stage" measures based on current conditions [34]. As one example of this, Paci AND ZANNONE (2015) OFFERed an improved access control approach to medical datasets based on use of SNOMED CT - the ontology-structured health terminology [25][35]. Semantically, if access to patients with a value of "*Cancer*" in the *Disease* field is restricted, access to other patients with related (subtype) diseases should also be restricted. Through identifying the associated variables and sensitive values, distribution attacks can BE DETECTED WHILE GENERALIsing micro-data to safe LEVELS USING SDC PRINCIPLES [39][40][41]. For instance, it is suggested to remove the *ancestor-descendant* dependencies from anonymised datasets. Related hEALTH PROBLEMS CAN BE organised with dependencies between individual clinical concepts, E.G. *CERTAIN INFECTIOUS AND PARASITIC DISEASES* [A00-B99] are an ancestor term to *VIRAL Hepatitis* [B15-B19] based on an inclusive relationship [37][38]. To tackle homogeneous attacks on sensitive attributes, Wang *et al.* (2013) proposed the *(k, ε)*-anonymity mechanism to maintain the distinctiveness of sensitive attributes within each equivalent class [36], e.g. {*Diabetes*} can be associated with {*Type-1 Diabetes} or {Type-2 Diabetes}*.

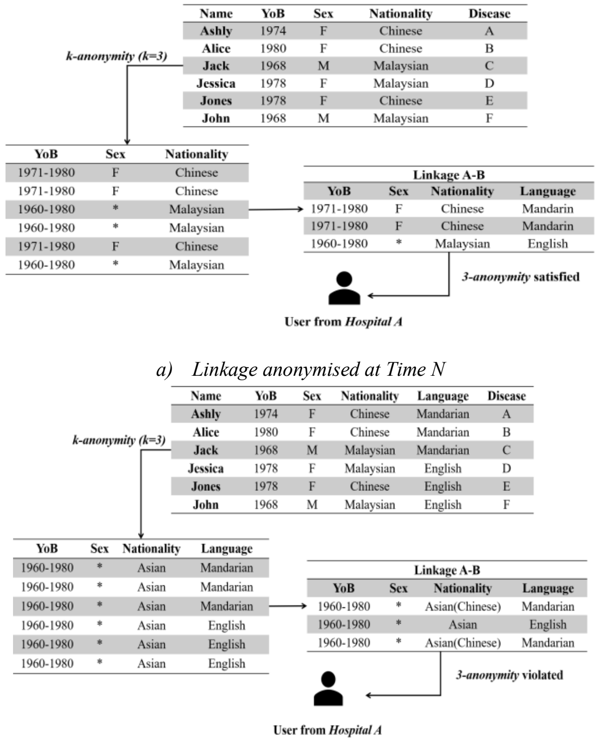
In addition to **explicit associations** provided in given vocabularies, it is also necessary to look for **implicit association** rules from linked datasets. The risk of re-identification increases as researchers collect and combine data with other 'co-

[1] In this and the following examples we assume demographic attributes such as *date of birth*, *sex*, *age*, *ethnicity* and *language* etc. are *QI* attributes defined by local custodians.

relations', e.g. using demographic features such as job, language, ethnicity, living suburbs etc. This may however result in privacy breaches. Through measuring the values shared by the same individuals, association rules can be identified in the linkage. Existing methods on this topic mainly relies on association rule mining, which focuses on transaction records with *0/1* values related to the item appearance [32]. Furthermore, numerical factors support and confidence can be calculated to represent the *item significance* as well as the *association strength* in the current linkage set.

Problem scenario-2. Detecting association leakages from data linkage release is challenging for many real-world linkages since they can be arbitrarily constructed in an *ad hoc* fashion by combining datasets. Furthermore, candidate datasets can be updated periodically. An example scenario of such a situation is demonstrated in Fig. 4. As discussed, *weak k-anonymity* offers a better model to protect datasets against local users and thus the identifier *Language* in Pharmacy B does not initially help users in Hospital A in re-identifying the linked records. As shown in Fig. 4 a), an anonymised linkage set can be generated under the *weak 3-anonymity* model. Based on such a result, it is common to learn associated factors in public health research. For instance, it is not difficult to discover the implicit association *Mandarin* → *Chinese* from anonymous releases between *Ethnicity* and *Language* variables. Theoretically, such associations would not directly breach the privacy protection in a single one-time release however they could weaken the anonymization over time. As shown in Fig. 4 b), suppose the same user (role) wishes to initiate a linkage between A and B after collecting the *Language* information as the fourth *QI* attribute. The Hospital A aggregated values in the column *Ethnicity* can then be utilised by adversaries to breach the pre-agreed scheme through refining the *3-anonymity* item, <*1960-1980*,*, *Asian*, *Mandarin*> in a *2-anonymity* item such as <*1960-1980*,*,*Chinese*, *Mandarin*>. A key challenge in this context is the dynamic nature of such violations. Ideally manual review should be minimised, yet any/all policies from all autonomous stakeholders should be checked for their overarching, integrated anonymity requirements.

As stated, individual privacy and the resultant data quality preservation are critical factors in designing SDC techniques. The privacy of individuals can be preserved at the cost of data utility [26]. Therefore, information loss is a crucial part of optimal anonymity approaches [27]. Previously, we identified the privacy risks by applying optimised SDC methods on record linkage and demonstrated how semantic technology can help mitigating the issue [28]. In this paper, we design the *Semantic Linkage K-Anonymity* (SLKA) method for anonymising record linkage data, balancing the privacy, data utility as well as associated risk analysis. Building on the eXtensible Access Control Markup Language (XACML) framework, we extend the anonymity schemes and obliged components for de-identification and privacy verification (potentially). As personal attributes can be freely integrated or extended locally, semantic implications and skewed distributions in attribute combination may threaten patient privacy. In this work, associations mined from "history publications" are extended and later enforced on the transformed linkage. More importantly, instead of focusing on simply protecting access to and use of sensitive attributes [17], this work effectively limits the chance of re-identification to an acceptable level and shows how semantic reasoning can deliver protection over real-time schemes and verify attributes based on the ever-growing external knowledge of collaborators and indeed potentially adversaries. To the best of our knowledge, no existing solutions have been designed for repeated linkage[2] which can be demanded in the long-term health studies [51]. Considering the sensitive data and the data privacy erosion challenges they give rise to; this work can fill the gap.



Name	YoB	Sex	Nationality	Disease
Ashly	1974	F	Chinese	A
Alice	1980	F	Chinese	B
Jack	1968	M	Malaysian	C
Jessica	1978	F	Malaysian	D
Jones	1978	F	Chinese	E
John	1968	M	Malaysian	F

YoB	Sex	Nationality
1971-1980	F	Chinese
1971-1980	F	Chinese
1960-1980	*	Malaysian
1960-1980	*	Malaysian
1971-1980	F	Chinese
1960-1980	*	Malaysian

Linkage A-B

YoB	Sex	Nationality	Language
1971-1980	F	Chinese	Mandarin
1971-1980	F	Chinese	Mandarin
1960-1980	*	Malaysian	English

a) Linkage anonymised at Time N

Name	YoB	Sex	Nationality	Language	Disease
Ashly	1974	F	Chinese	Mandarian	A
Alice	1980	F	Chinese	Mandarian	B
Jack	1968	M	Malaysian	Mandarian	C
Jessica	1978	F	Malaysian	English	D
Jones	1978	F	Chinese	English	E
John	1968	M	Malaysian	English	F

YoB	Sex	Nationality	Language
1960-1980	*	Asian	Mandarian
1960-1980	*	Asian	Mandarian
1960-1980	*	Asian	Mandarian
1960-1980	*	Asian	English
1960-1980	*	Asian	English
1960-1980	*	Asian	English

Linkage A-B

YoB	Sex	Nationality	Language
1960-1980	*	Asian(Chinese)	Mandarian
1960-1980	*	Asian	English
1960-1980	*	Asian(Chinese)	Mandarian

b) Linkage anonymised at Time N+1

Figure 4. Ongoing release of linkage A-B by *weak 3-anonymize*d.

III. Method

Most statistical disclosure control practices are based on tabular data recognition and categorising sensitive information including *sequence identifiers* (*SID*) and *quasi-identifiers* (*QI*) [42]. To preserve patient privacy while reducing unnecessary transformations, we design a semantic-based linkage k-anonymity (*SLKA*) framework with dynamic risk detection and prevention using semantic reasoning related to the potential risks associated with these identifiers. We introduce the basic mathematical models associated with these concepts.

A. Basic Concepts

DEFINITION-1 (OVERLAPPING POPULATION). In the context of record linkage between two data resources *X* and *Y*, an overlapping population ($OP_{X\text{-}Y}$) refers to a group of common individuals (CI_{X-Y}) in both datasets, where the resources themselves may be maintained by different organisations.

A partial injection relation typically exists between EHR linkage and local datasets. For instance, Fig. 1 shows patients identified by *A01*, *A02*, *A03* and *A04* in Hospital A dataset that may match with another dataset containing *B011*, *B021*, *B031* and *B041* in the Pharmacy B dataset. In this case, only two pairs of records are matched to the same individuals, i.e. (*A01*, *B011*) and (*A04*, *B041*) Therefore, the OP_{A-B} consists of the common individuals (CI_{A-B}) uniquely identified in the linkage set, i.e. *L-01* and *L-02* respectively. In other words, OP_{A-B} refers to the intersection of the independent databases and thus it can be constructed as a subset of the member databases, expressed as $CI_{A-B} = LR_A \cap LR_B$.

In case the presence of an individual can be inferred, data custodians may have the anonymity model switched from *k-anonymity* to *weak k-anonymity* by computing the overlapping rate (*OR*), which is given as $OR_{X-Y} = \frac{|OP_{X-Y}|}{|LR_X|}$ or $OR_{Y-X} = \frac{|OP_{X-Y}|}{|LR_Y|}$. In the example of Fig. 3, a policy can be defined in Hospital A such as *weak k-anonymity* which can be used if OR_{A-B} is within the safe range say (0, 0.7); otherwise *k-anonymity* should be used since there is an increased chance for adversaries to infer the patient presence. On this basis, *weak 2-anonymity* gives a 50% chance of re-identification and can be applied to protect the record linkage.

DEFINITION-2 (ANONYMITY SCHEME). Locally, data privacy is preserved based on an anonymity scheme (AS), including the quasi-identifier set and the numerical requirements given as AS_x= <QI_x, k_x>.

As shown in the Fig. 3, anonymity schemes are defined in both datasets as AS_A = <QI_A = {*Date of Birth, Zip, Ethnicity*}, k_A = *2*> and AS_B = <QI_B = {*Date of Birth, Sex*}, k_B =*2*>. These schemes can be used for privacy protection for either local access (dataset A or dataset B) or remote linkage (*linkage A-B*)

DEFINITION-3 (LOCAL RELEASE). *Local release* (LR) refers to the local records disclosed to users who are authorised to access them. The data subjects in LR_X (or LR_Y) are called *local individuals* (LI_{Xi} or LI_{Yj}) and the QI_X or QI_Y attributes are generalised according to given local anonymity schemes AS_X or AS_Y.

As shown in Fig. 3, the generalised dataset about Hospital A refers to a local release of $linkage_{A-B}$, i.e. the individuals included in LR_A. Based on the defined scheme AS_A, the details of LI_{Ai} are generalised and can be used for replacing linkage values during anonymisation.

DEFINITION-4 (DOMINANT AND SUBORDINATE DATABASE). In the context of sharing $linkage_{X-Y}$, there must be databases from which users use when initiating linkages. These databases are called **dominant databases**; otherwise they are regarded as **subordinate databases** in terms of the linkage.

Returning to the example in Fig. 4, dominant databases are given along with the linkage requests from Hospital A. Therefore, the database in Hospital A is the dominant database and its anonymity scheme AS_A = <QI_A = {*Date of Birth, Zip, Ethnicity*}, k_A=2> is the **dominant anonymity scheme**. Given the hypothesis that any adversary can have background knowledge only from the local release, the anonymity scheme AS_A will instigate the on-going linkage anonymisation and release.

The access rights of **subordinate databases** are not necessarily granted to users at the beginning of a linkage, e.g. when no trust exists, however this may be achieved over time. This requirement must adhere to all lawful regulations on health data confidentiality, e.g. protected health data should only be disclosed to authorised users [42].

DEFINITION-5 (LINKAGE K-ANONYMITY): Record linkage $Linkage_{X-Y}$ satisfies a given anonymity requirement if, for each common individual in OP_{X-Y} there are at least k_{X-Y} matching tuples in the local release LR_X (or LR_Y).

As with access rights in authorisation decisions, anonymity requirements for each linkage applications are produced by composing all independent anonymity policies of the relevant datasets. Suppose the numerical requirements enforced on record linkage should not be less than the associated requirements of any participating dataset resources so as to satisfy all statistical requirements. For instance, the numerical requirement for $Linkage_{A-B}$ can be achieved by selecting the maximal *k*, e.g. k_{A-B} = *max* (k_A, k_B) = *2* (both A and B are required with a minimum of 50% risk).

DEFINITION-6 (LINKAGE QUASI-IDENTIFIER) In the context of linking datasets *X* and *Y* , the linkage quasi-identifiers LQI_{X-Y} are equal to the dominant current anonymity scheme (e.g. QI_X). The non-quasi-identifiers (NQI_{X-Y}) are quasi-identifiers that only belong to the subordinate databases.

Distinguishing the *LQIs* from all candidates can help reduce unnecessary generalisation and preserve linkage utility. For instance, with the dominant database in *Hospital A*, the non-quasi-identifiers NQI_{A-B} refers to the QI attributes of Pharmacy B only, i.e. $NQI_{A-B} = \{qi | qi \in QI_B \backslash QI_A\}$. According to Definition 4, $NQI_{A\text{-B}}$ will not help (Hospital A) staff to re-identify common individuals and therefore they need not be generalised.

DEFINITION-7 (GENERALISATION): In the context of anonymising $linkage_{X-Y}$, the *LQI* attributes need to be generalised as $GV[QI_{X-Y}]$ until all tuples involving common individuals $T[CI_{X-Y}]$ can meet the numerical criteria k_{X-Y}.

Generalised tuples reflect privacy realisations. It is noted that for categorical attributes in the set QI_{A-B}, the generalised values (*GV*) on *n-level* hierarchies can often be established, i.e. $G_nV_m[qi_{A-B}]$. For instance, a subtree $G_5(3000)$ *[Zip]* can be established for the postcode *3000*, comprised of *3000*, *300**, *30***, *3**** and ****. In addition, certain pre-processing is often demanded for numerical attributes, e.g. based on *ad-hoc* integer ranges.

DEFINITION-8 (COMPOSITION ATTACKS): Composition attacks occur when linkage tuples involving common individuals CI_{X-Y} are found to be matching with less than k_X (or k_Y) local individuals LI_X (or LI_Y) in the dominant database.

Linkage publication may lead to privacy leakage when adversaries acquire auxiliary information from previous linkage releases. As shown in Fig. 4 a), the association *Mandarin* → *Chinese* can be learned from previous releases that help local users to refine the generalised items and then potentially breach the privacy requirement (*1/3* → *1/2*). To maintain privacy protection levels, verification is required against previous associations and local knowledge.

DEFINITION-9 (RELATED VALUE): Related values are based on record linkage releases, denoted as $RV_{X\text{-}Y}(V_m, V_n)$ where $\{V_m \rightarrow_r V_n | \rightarrow_r \in R\}$.

In addition to the explicit relations among categorical attributes such as *isAncestor* (↑), *isDescendent* (↓), *equivalentWith* ($\rightarrow_{eq}$) within or across hierarchies, the relation set *R* also includes implicit relations ($\rightarrow_{ir}$) that may exist among attributes as **condition** and **consequence** elements. For instance, related values such as $RV_{A\text{-}B}$*(Mandarin, Chinese)* give rise to privacy issues by specialising the protected value *Asian* as *Chinese* in Fig. 4 b). Therefore, we establish a generic workflow through which associations between *non-quasi-identifiers* (*NQIs*) and protected *QI* attributes can be established. As noted, the *NQIs* are not generalised. Therefore, rule mining starts with their unit values (level n) and aggregated values of *QIs*, e.g. $RV_{A\text{-}B}$*(Mandarin, Female)*, given that only full associations can threaten patient privacy. To protect the release from privacy breaches, such associations will be added into the knowledge base as the basis for privacy verification, i.e. conditional values occurring in future releases need to be generalised until no protected elements can be specialised.

Speciality in the linkage or local datasets is unpredictable. For instance, if the mandarin speakers are the only Asian language speakers in the current linkage set, then it is not enough to verify the release with one associations formed as *"Mandarin → X"*. In addition, the ancestor *"Asian_Language → X"* should be added to the knowledge base, too. In other words, values with one-step generalisation could still narrow the equivalence groups to the "*less-than-k*" form. As a result, it is necessary to continue the mining of conditional items with one-level reduced and then include resultant associations in an accumulative process. As shown in Fig. 5, after mining the anonymised records, all conditional identifiers should be analysed from their highest levels in the association rule set, i.e. for each identifier, more associations can be established by reducing details gradually. This process will terminate at level 0 and then start for another identifier. For instance, *Language* is a *NQI* at *Time N* and the unit values are available (level 4). With the *initial k-anonymised* result, all language values can be processed through 3 other rounds.

DEFINITION-10 (RISKY INDIVIDUAL): Individuals in overlapping populations may suffer from composition attacks if matching records can be found from the local releases where the number is less than $k_{X\text{-}Y}$.

Associations identified in ongoing linkage applications may allow adversaries to obtain specialised information. As a result, the anonymisation for a set of individuals may be eroded and eventually violated over time, even though they were anonymised and met all required (known) statistical data risk disclosure demands at the current time point. For instance, the anonymised *3-anonymised linkage*$_{A\text{-}B}$ in Fig. 4 is refined into *2-anonymised* linkage due to the association mined from the previous release.

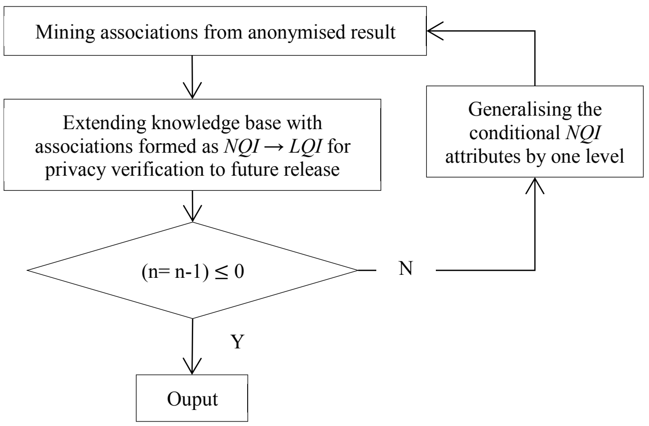

Mining associations from anonymised result
Extending knowledge base with associations formed as NQI → LQI for privacy verification to future release
Generalising the conditional NQI attributes by one level
(n= n-1) ≤ 0
N
Y
Ouput

Figure 5. Accumulative association mining among attributes (*Time N*).

B. Semantic based Policy Specification

1) XACML-based Acccess Frameork

Leveraging an existing solution [17] where XACML policy components are expressed and used for semantic reasoning in supporting access control policy compliance and privacy protection, we formalise anonymity schemes within obligation components so that the semantic-based anonymisation can be enforced when evaluating linkage requests and releases. Fig. 6 shows an extended XACML framework where the access to record linkage can be managed in a distributed environment. This architecture consists of components such as Policy Administration Points (PAPs), Policy Information Points (PIPs), Policy Decision Points (PDPs) and Policy Enforcement Points (PEPs). Initial policies written in PAPs are made available to PDPs. These policies represent the restrictions related to use of certain resources (data and services). In a given XACML policy evaluation, it is possible that PDPs require more attributes (credentials) from PIPs, e.g. related to the user's credentials and trustworthiness. The resultant suggestions are returned to PEPs, which then permit or deny access requests to potentially fulfil obligations. Specific to the data linkage case, the requester sends a linkage request to the central PEP (step 1) and it passes a XML-formatted request to the PDP for specific requirements (step 2). The PDP firstly retrieves policies from the local PAP (step 3), to which the original policies are transferred from the remote PAPs (step 4). Through semantic reasoning of the related information, the linkage anonymity scheme can be produced at the central PAP (step 5). At this point, the PIP may be requested to disclose attributes stored centrally and/or remotely. According to the decisions made by the PDP, the actual values are pulled from local datasets (step 6) and the generated obligations are enforced by PEPs (step 7). This semantically-extended XACML framework forms the foundation for protected linkages using reasoning capabilities.

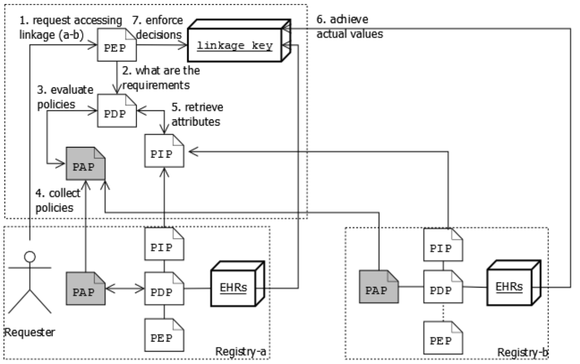


Figure 6. Record linkage access control in *SLKA*-extended framework.

2) *Semantic-based Privacy-preserving XACML*

Table I shows generic concepts of the XACML framework (**Request**, **Obligation** and **Target**) and the extended anonymising method (e.g. **Anonymity**, **RiskAnalysis** etc.). Along with the linkage request, a linkage anonymity scheme needs to be constructed. According to Definition 2, related **Anonymity** schemes include the *QI* attributes and numerical requirement *k* pre-defined to the datasets. From the candidate databases, e. g. DB_A and DB_B, it is possible to determine what anonymity requirements need to be considered. As defined in *Rule 1*, the AS_{A-B} will be enriched with results such as *hasAnonymity*(LS_{A-B}, Ano_A) and *hasAnonymity*(LS_{A-B}, Ano_A).

1. Request(?req), hasResource(?req, ?ls), linkFrom(?ls, ?db), hasAnonymity(?db, ?ano) → hasAnonymity(?ls, ?ano)

TABLE I. SEMANTIC NOTATIONS IN LINKAGE K-ANONYMITY

Purpose	Semantic facts	
	Class	*Associated Properties*
Scheming anonymity policy	**Request**	*hasResource; hasAction; hasSubject*
	Anonymity	*hasQI; enforceAnoReq;*
	LinkageScheme	*hasAnonymity; linkFrom;hasRA*
	Role	*authenticatedWith; hasRA*
	Attribute	*hasAncestor; hasDescendant; sameAs*
Verifying privacy preservation	**Obligation**	*hasTarget*
	Target	*hasResource; hasSubject; hasAction*
	RiskAnalysis	*hasSubRA; hasConditionAttr; hasConsequenceAttr; hasFunction*
	Function	*enforce*
	Tuple	*hasAttribute*

As discussed, linkage anonymisation depends on the local releases and dominant anonymity schemes. Based on Definitions 3-4, databases that users (or roles) are authorized to access will dominate the linkage anonymity. Correspondingly, *Rule 2* is used to find the relation between credentials and local anonymity schemes via *enforceAnoReq*(*Clinician*, Ano_A). Based on the result, *Rule 3* can be used to derive the possible *LQI* attributes in the linkage, e.g. *hasLinkageQI*(LS_{A-B}, *Gender*) from the *QI* attributes in the related databases such as *hasQI*(Ano_A, *Gender*)[3]. According to Definitions *4-5*, *non-QI* attributes are not helpful to re-identify patients from record linkage as long as *linkage k-anonymity* holds. In this case, user knowledge is assumed to be consistent with the home sites where they authenticated.

2. Request(?req), hasSubject(?req, ?role), hasResource(?req, ?ls), hasAnonymity(?db, ?ano), linkFrom(?ls, ?db), authenticatedWith(?role, ?db) → enforceAnoReq(?ano, ?role)

3. Request(?req), hasSubject(?req, ?role), enforceAnoReq(?ano, ?role), hasResource(?req, ?ls), hasQI(?ano, ?qi) →hasLinkageQI(?ls, ?qi)

Numeric requirements on the linkage LS_{A-B} can be selected through reasoning using Rules *4-5*. According to Definition 5, after comparing all related values such as *hasAnoReq*(LS_{A-B}, *2)*, a composite requirement can be achieved and used in the *hasLinkageAnoReq*(LS_{A-B}, *2)* for linkage anonymisation.

4. LinkageScheme(?ls), hasAnonymity(?ls, ?ano), hasAnoReq(?ano, ?n) → hasAnoReq(?ls, ?n)

5. LinkageScheme(?ls), hasAnoReq(?ls, ?n1), hasAnoReq(?ls, ?n2), greaterThan(?n1, ?n2) → hasLinkageAnoReq(?ls, ?n1)

3) *Privacy Verification*

In addition to tracing anonymity schemes, requests are also used to locate the authorisation and obligation rules to be executed. In Fig. 4, $linkage_{A-B}$ demands access privileges exist for DB_A and DB_B. If this can be confirmed by *Rule 2*, the anonymisation will be conducted based on the scheme given previously. As Table I shows, **Obligation** rules can be specified using objects with particular functions. For instance, the obligation (Ob_1) can be defined to enforce the inferred anonymity from target *subject* (e.g. *Clinician*) onto resource (e.g. $linkage_{A-B}$). Semantically, this can be formalised as *hasTarget*(Ob_1, Tar_1) and *hasSubject*(Tar_1, *Clinician*). Given the reasoned anonymity scheme *enforceAnoReq*(*Clinician*, Ano_A), *Rule 6* is reasoned to enforce the scheme for target linkage via *enforceAnonymity*.

6. Obligation(?o), hasTarget(?o, ?tar), hasSubject(?tar, ?role), enforceAnoReq(?ano, ?role) → enforceAnonymity(?o, ?ano)

However, it may not be sufficient to protect the on-going linkage from associated attributes. For this reason, the linkage instances should be attached with all associations mined from their "previous releases" as well as the processing functions. Together these are classified as **RiskAnalysis**, such as the generic instance RA_{AB} representing all associations about the $linkage_{A-B}$ cohort. Specially, concrete value pairs mined from each release should be added to the user (role), such as *hasRA*(*Clinician*, RA_{AB-1}), *hasRA*($linkage_{A-B}$, RA_{AB}) and *isA*(RA_{AB-1}, RA_{AB}) during the Time *N* release. Back to the example in the Figure 4, the association *Mandarin → Chinese* found at Time *N* can be formalised as *RA* instances with condition and consequence items. Formally, RA_{AB-1} can be expressed with *hasConditionAttr*(RA_{AB-1}, *Mandarin)* and *hasConsequenceAttr*(RA_{AB-1}, *Chinese)*, implying the consequent and antecedent attributes of the 2-ary relations. During the accumulative rules mining for each release of certain linkage, associations may establish after generalising the condition item

[3] Attributes related to linkage instances via *hasQI* are not the *QI* attributes to be generalised for the linkage. In fact, the actual *linkage QI* attributes need to be confirmed by reasoning on rules *1-3*.

in the RA instances, such as *Asian_Language* → *Chinese* from *Mandarin* → *Chinese*. In this case, they can be added to the original associations via *hasSubRA*, e.g. *hasSubRA*(RA_{AB-1}, SRA_{AB-1}) while the formalisation of SRA_{AB-1} is identical to RA instances.

To mitigate the risk hiding in **Tuples** (the records of linkage sets) , it is necessary to replace antecedent values with generic forms whenever their generalised consequent values can be refined and subsequently weaken the privacy effect. In other words, contained attribute ought to be checked in the combination form. For this purpose, *Rule 7* is defined to locate the associations that are necessary to be used to protect current release. As stated, knowledge (associations) learned in the previous releases are related to role names. Supposing a clinician at Hospital A viewed the $linkage_{A-B}$ at Time *N,* he/she can be assumed knowing the associations (such as RA_{AB-1}) about the population in $linkage_{A-B}$. As long as the type can be confirmed via *isA*, concrete analysis can be enforced through the **Obligation**, e.g. *enforceRA*(Ob_1, RA_{AB-1}). The aim of filtering out associations from RA_{AB} based on the role name is to mitigate the unnecessary data transformation while preserving privacy.

7. Obligation(?o), hasTarget(?o, ?t), hasSubject(?o, ?r), hasRA(?r, ?ra1), hasResource(?t, ?ls), hasRA(?ls,?ra2), isA(?ra1,?ra2) → enforceRA(?o, ?ra1)

To check and process risky individuals (tuples), *Rule 8*-9 are defined with reasoned associations by conducting the designate **Functions** (e.g. *Generalisation* based on generalising current values by "one step") to the "current release" of the same linkage schemes. Within the knowledge base, hierarchies in categorical attributes can be specified such as *hasAncestor(Mandarin, Asian_Language)*. Besides, anonymised tuples such as *<1960-1980*, * , *Asian*, *Mandarin>* in the Time *N+1* release can be specified with *hasAttribute(Tup1, Mandarin)* and *hasAttribute(Tup1, Chinese)*, as well as the pre-defined function via *hasFunction*(RA_{AB-1}, *Generalisation).* Once detecting the risky attributes in tuples, pre-defined functions can be suggested to the compromised values (e.g. *Mandarin*) via *enforce(Generalisation, Mandarin)*.

After finishing this check, a temporary tuple will be created. Potentially, *Rule 9* is defined to search (potentially) more RA cases along with the predicate *hasSubRA* to test if there is a risk from the ancestor value. According to the definition, if there is one attached to active **RiskAnalysis** instances then the tuple cannot be released. It has to be verified against all related cases e.g. *enforce(Generalisation, Asian_Language)*.

8. Obligation(?o), enforceRA(?o, ?ra), hasFunction(?ra, ?fun), hasConditionAttr(?ra, ?a2), hasConsequenceAttr(?ra, ?a1), hasAncestor(?a1, ?a_1), Tuple(?tup), hasAttribute(?tup, ?a2), hasAttribute(?tup, ?a_1)→ enforce(?fun, ?a2)

9. Obligation(?o), enforceRA(?o, ?ra), hasSubRA(?ra, ?sra), hasFunction(?sra, ?fun), hasConditionAttr(?sra, ?a2), hasConsequenceAttr(?sra, ?a1), hasAncestor(?a1, ?a_1), Tuple(?tup), hasAttribute(?tup, ?a2), hasAttribute(?tup, ?a_1)→ enforce(?fun, ?a2)

IV. EXPERIMENT DESIGN

A. Experiment Design

To analyse the performance of the *SLKA* model, we consider a linkage scenario where data can be linked, anonymised and verified at a linkage centre. In this study, we utilised data sources from a health survey dataset related to residents in Victoria, Australia. To understand the population health, the Department of Health in Victoria periodically undertakes a survey involving 25,000+ Victorians to assess their overall health and lifestyle considering a variety of factors covering work-life balance, discrimination and domestic violence, drinking and smoking habits and basic demographics [43].

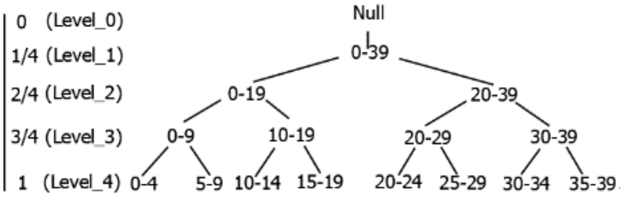
To reflect real-world linkage applications, such survey data can be used to support research by linking with other clinical data. As one example, the Australian Diabetes Data Network (ADDN) includes information on over 13,000 patients with Type-1 and Type-2 diabetes [44]. In this experiment, we consider linkage between ADDN and VicHealth to explore the impact of alcohol consumption on diabetes. To compare the performances of different anonymizations, we sampled 1000 records from the VicHealth survey results while assuming there were overlapping populations in both data resources.

As shown in Table II, attribute details such as the data types, the value distribution as well as the hierarchical structure are listed. Prior to data linkage, repositories submit their data dictionaries to linkage centres leveraging (wherever possible) standardised sources. Similar to the categorical attributes, numeric variables are aggregated based on *ad-hoc* schemes for transformation. Fig. 7 shows a fragment of the data dictionary: the 'Age' values are processed into intervals, such as *0-4*, *5-9* etc. in Fig. 7 a); standard taxonomies of geo-locations (e.g. postcodes, statistical area levels etc.) and demographics (e.g. languages, ethnicities etc.) are organised in Fig. 7 b) and Fig. 7 c) [29][33][45][46]. Based on such a hierarchical structure, the specificity of values can be quantified through attaching "*0, 1/3, 2/3, 1*" to a 4-level classification scheme where "1" stands for the raw units (e.g. *3205-Greek*) while "0" for empty cells [47]. In addition, to serve the multi-level association mining, values in the data dictionary are tagged with the level number, e.g. "Level_0", "Level_4" etc.

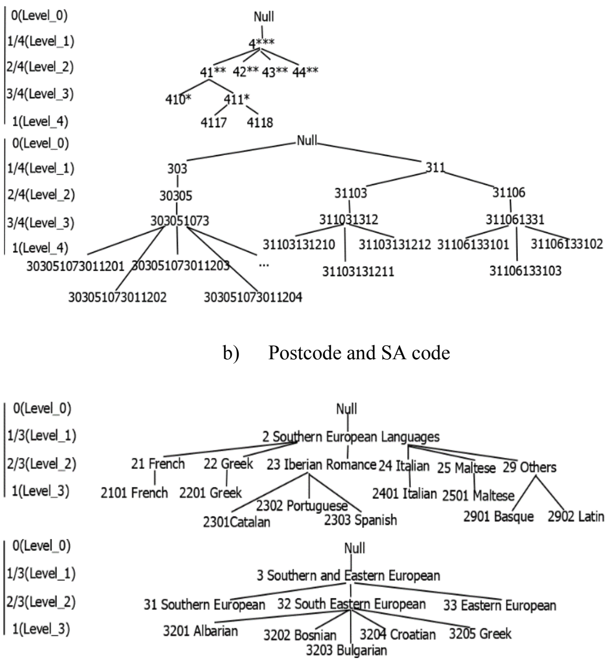
TABLE II. DETAILS OF ATTRIBUTE VALUES

Dataset	**Attributes**			
	Attribute name	***Number of values***	***Number of levels***	***Type***
ADDN	Gender	3	2	Categorical
	Ethnicity	45	4	Categorical
	Postcode	320	5	Categorical
VicHealth	Age	87 (15 groups)	5	Numerical
	Language*	36	4	Categorical
	SA1 code	704	5	Categorical

*{Language} is the dynamic attribute changing over time.



a) 5-year age hierarchy



b) Postcode and SA code

c) Ethnicity and Language

Figure 7. Attribute variable hierarchies.

B. SLKA Implementation

To realise semantic privacy-oriented linkages for repeated linkage, we developed a prototype using Protégé 4.0 to show how semantic reasoning can dynamically inform the *linkage k-anonymity* with reasoning capabilities. Based on the above facts and semantic *Rules 1-9* the working mechanism of *SLKA* can be demonstrated through a linkage scenario. Initially users request to view the linkage from ADDN. On this basis, policy requirements on linkage sets can be generated through semantic reasoning, instead of through manual review. As shown in Fig. 8, the linkage protection requirements are derived from ADDN and VicHealth policies. Once receiving the request for such linkage as *Vic_ADDN*, related patient EHRs are then extracted from both databases (*ADDN* and *VicHealth*) with local anonymity schemes, *ano_1* and *ano_2*. Here both *ano_1* and *ano_2* contain the *quasi-identifiers* and statistical requirements related to the two databases, i.e. QI_{ADDN} = {*Gender, Ethnicity, Postcode*}$_{k=3}$ and $QI_{VicHealth}$ = {*Age, Language, Postcode*}$_{k=2}$ for ADDN and VicHealth respectively. Through reasoning about semantic rules, the linkage anonymity scheme can guide the privacy preservation demands through the use of *hasLinkageQI* as well as *hasLinkageAnoReq*.

In addition to anonymisation, the resultant tuples need to be verified based on associations in case of inference risks. Through analysing tuples (*tuple1*) against related RA instances (*RA1*), Fig. 9 illustrates how the previous linkage releases affect the privacy verification via associations such as *2201-Greek* → *3205-Greek*. Once a risky combination is detected, the strategy *Generalisation* is suggested to enforce the original value and antecedents, e.g. replacing *2201-Greek* with *22-Greek* given the *RA1* learned from previous releases. As discussed, it is necessary to extend associations mined in the accumulative procedure and semantically it can be realised via using *hasSubRA*. For instance, based on *tuple1* and *tuple2* standing for two (linked) patient records after anonymisation, privacy verification is conducted twice: after testing *2201-Greek*, it will continue with *22-Greek* by searching for all subordinate RA cases, e.g. *22-Greek*→*3205-*Greek. Through reasoning on the *Rule 9*, another risky combination (*SubRA1*) can be detected and hence the original *2201-Greek* will be finally replaced with *2-Southern European Language* in this example.

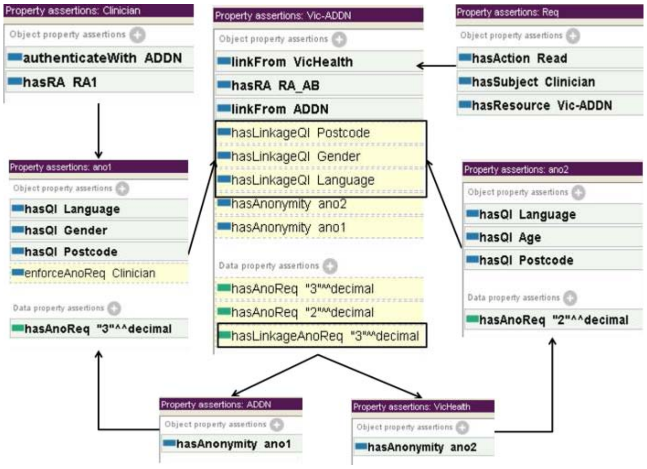


Figure 8. Semantic reasoning using anonymity scheme composition.

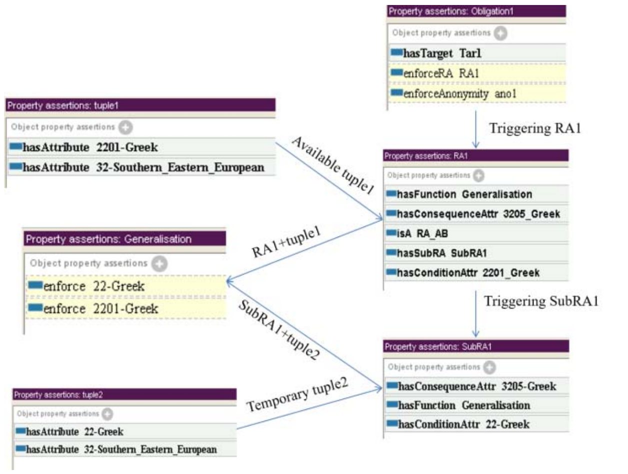

Property assertions: Obligation1
Object property assertions
hasTarget Tar1
enforceRA RA1
enforceAnonymity ano1
Triggering RA1
Property assertions: tuple1
Object property assertions
hasAttribute 2201-Greek
hasAttribute 32-Southern_Eastern_European
Available tuple1
Property assertions: RA1
Object property assertions
hasFunction Generalisation
hasConsequenceAttr 3205_Greek
isA RA_AB
hasSubRA SubRA1
hasConditionAttr 2201_Greek
RA1+tuple1
Property assertions: Generalisation
Object property assertions
enforce 22-Greek
enforce 2201-Greek
Triggering SubRA1
SubRA1+tuple2
Property assertions: SubRA1
Object property assertions
hasConsequenceAttr 3205-Greek
hasFunction Generalisation
hasConditionAttr 22-Greek
Property assertions: tuple2
Object property assertions
hasAttribute 22-Greek
hasAttribute 32-Southern_Eastern_European
Temporary tuple2

Figure 9. Resultant operations for linkage.

C. Result Analysis and Discussion

The experiments were performed on a laptop with Windows 10 operation system (3.20 GHz Intel Core processor and 8GB Memory). Anonymised records were produced using ARX, an open source anonymisation tool [48]. As the underpinning algorithm, Flash was used to explore optimal data utility [49]. After importing datasets and configuring the basic settings including *k-anonymity, l-diversity, t-closeness, δ-Presence*; data dictionaries; *k* values as well as attribute weights[4], the results show how basic privacy criteria are satisfied.

1) *Data Sources*

Continuing with the collaborative scenario where researchers at ADDN apply to link patients with the survey respondents in the VicHealth dataset. To explore the method performance, we simulate ten linkage scenarios involving 10 records, 20 records, …, 100 records by sampling ADDN individuals based on $OR_{ADDN\text{-}VicHealth}$ from 1% to 10%. Selected individuals are assumed to be common individuals included in both ADDN and VicHealth. To reduce the impact of the value distribution, we shuffle all rows and then select the first 10, 20, …, 100 records and construct the linkage sets. Before we consider anonymisation, it is necessary to explore the raw records in the linkage sets. As shown in Fig. 10, given security thresholds of 25%, 33.3% and 50% (corresponding to *4-anonymity*, *3-anonymity* and *2-anonymity* [50]), the proportions of data that fall into safe ranges increases when more common individuals are involved. In each cluster, we can find that the higher *k* value that is set, the fewer individuals satisfy the security constraint.

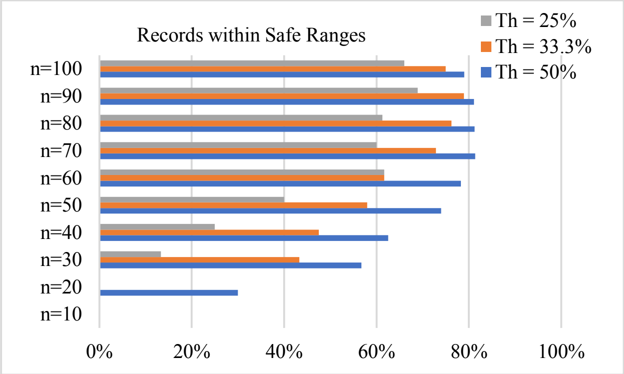
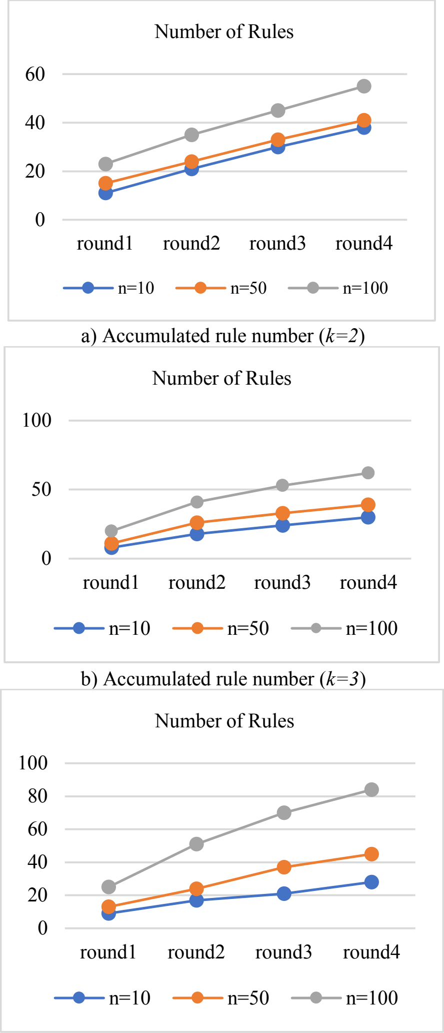


Figure 10. Qualified records in linkage groups.

2) *Attribute Associations*

To prevent privacy leakage from occurring in linkages, the contents of previous releases needs to be parsed and association rules used to check the privacy of current release requests. In this work, association mining was conducted using the *Apriori* algorithm on released linkages [32]. To show the association mining with different conditions, we focus on using full-chance[5], 2-ary associations to protect the same linkage release over time (with 10, 50 and 100 records respectively). As the accumulative mining proceeds *Language* values, Fig. 11 shows how more associations can be discovered when privacy checking. Specifically, *28* more association rules were added to the knowledge base through the multi-round processing on the group (*n=50*, *k=3*). With more sample records involved, the number of associations increases to *21*, *24* and *35* rules mined from groups (*n=10*, *k=2*), (*n=50*, *k=2*) and (*n=100*, *k=2*) in the second round. Such value associations evaluated at different levels can avoid privacy breaches during dynamic linkages.



a) Accumulated rule number (*k=2*)

b) Accumulated rule number (*k=3*)

c) Accumulated rule number (*k=4)*

Figure 11. Association mining with {*Language*}hierarchical values.

In this work, the key is to protect linkage release from privacy breaches. Results may be specialised by association rules and thus lead to a higher re-identification chance - potentially greater than *1/k*. To show the potential risk without semantic privacy protection, we compare the average number of compromised records by assuming the anonymity requirements can never decrease in the repeated releasing, i.e. there are 6 groups with non-decreasing *k* sequences. As shown in Fig.12, for a first-time value such as *k=2*, the higher k the more records need to be processed, e.g. 21 in group (k_1*=2*, k_2*=2*) and 26 with (k_1*=2*, k_2*=3*). This can be explained given the same associations mined from the previous release, hence it is easier to breach the security requirements since a larger equivalent group size is demanded. The requirement gap in the two-time release will also impact the protection efforts e.g. zero-gap groups (k_1*=2*, k_2*=2*), (k_1*=3*, k_2*=3*) and (k_1*=4*, k_2=4) have records at risk in less than one-gap groups (k_1*=2*, k_2*=3*) and (k_1*=3*, k_2*=4*). This reaches a

[4] Attribute weights can reflect the precedence of generalisation, i.e. ARX tries to reduce the modification on attributes with higher weights.

[5] *Full chance associations* refer to associations with the highest confidence (i.e., minimum confidence =1).

maximum (i.e. 31 records) in the group ($k_1=2$, $k_2=4$), which has the biggest gap.

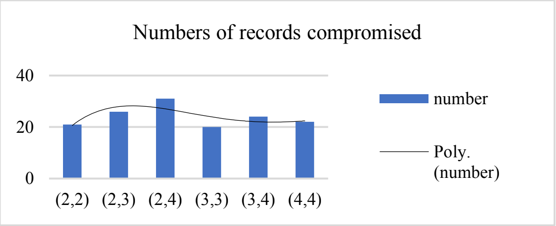


Figure 12. Comparing the numbers of records at risk.

3) *Average utility loss*

For the *useful* privacy preservation, data should have high utility, hence the information loss caused by generalising values needs to be compared among different approaches. In this context, SSE/SST are used to measure how many details remain in the releases [27]. When dealing with categorical attributes, the original and aggregate values can be quantified by subtracting hierarchical levels [29] [30] [31]. More values to be transformed subsequently result in less utility. Specifically, data utility loss can be measured by (1). where x_{ij} is the original value of attribute *j* about individual *i* and x'_{ij} refers to its protected form; *m* is the number of *QI* attributes in the linkage and the function *level_Dis()* denotes the structural distance between categorical values in the respective hierarchies.

$$\frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} \text{level\_Dis}(x_{ij},x_{ij}^{'})^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} \text{level}(x_{ij})^2} \qquad (1)$$

As shown in Fig. 13, after employing *k-anonymity* and *linkage k-anonymity* given different constraints and data sizes, we see how different degrees of utility loss can occur. In the cross-group comparison, as the data (linkage) size increases, there is an overall decline of the data to be changed for privacy-preservation. Within each group (e.g. *n=20*) anonymised with the same algorithm (e.g. *k-anonymity*), a higher *k* value tends to cause higher levels of distortion, e.g. 16% with *k*=2 and *40%* with *k=3*. This can be explained by the relation between the privacy cost and the resultant data utility, i.e. stronger protection can normally cause greater data loss.

Finally, for each scenario the benefit has been shown by using *linkage k-anonymity*. For instance, with the same constraint (e.g. *k=2*) in the 30-record group, the *linkage 2-anonymity* only results in 5.83% of modifications to records for nearly half of the *2-anonymity* (11%). Furthermore, the gap between the two methods becomes more obvious when using a higher constraint, e.g. *5.83%* vs *23%* with *k*=3 and *7.71%* vs *33%* with *k=4*. Such results demonstrate an increased impact on data utility by setting the statistical models accordingly. As discussed, *linkage k-anonymity* has the potential to cause privacy issues through on-going releases. In addition to comparing compromised records, we explore the extra costs in using *SLKA* to avoid such risk disclosures. Different from the comparison between *k-anonymity* and *linkage k-anonymity* in static scenarios, protection here assumes an iterative (repeated) linkage release process, i.e. the same linkage scheme is used whilst the local policies are allowed to change. Continuing with the example where {*Language*} is changed from a *NQI* to a *QI* attribute in the linkage set, after mining the association rules, the resultant values must be processed through reasoning using the Rules *8-9*, i.e. conditionals in associations will be generalised from the tuple until all records satisfy the policy constraint. As shown in Fig. 14, the average information loss shows the extra cost in *SLKA* in dealing with risky individuals. As with the last comparison, *k* values can affect the level of distortion from *11.5%* (*k=2*) vs *12.9%* (*k=3*). Compared with *linkage k-anonymity*, using *SLKA* for each group leads to additional loss after the privacy checks from *13.7%* vs *5.90%* (*k=4*).

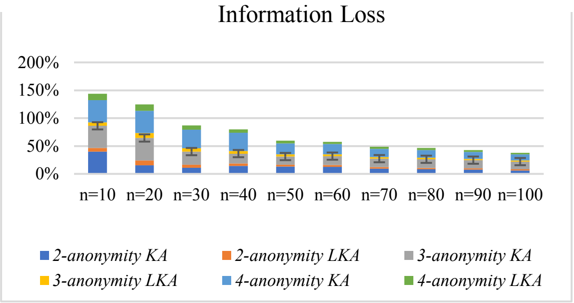


Figure 13. Information loss with *k-anonymity* and *linkage k-anonymity*.

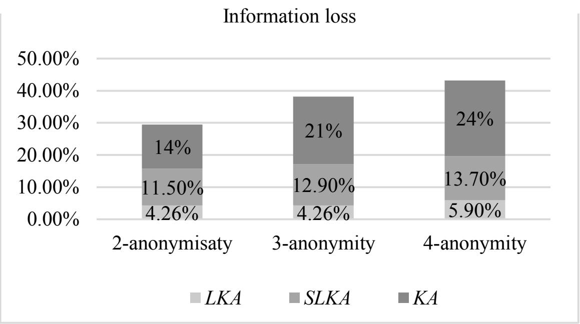


Figure 14. Average information loss with *k-anonymity*, *linkage k-anonymity* and *semantic linkage k-anonymity*.

V. Conclusions

In this paper, we propose a privacy‒preserving linkage framework by extending *weak k-anonymity* algorithms with linkage features and semantic reasoning capabilities. Based on the properties of EHR linkage infrastructures, we provide an example of distributed, data-driven applications with dynamic security demands. In addition to balancing data privacy and utility, adding reasoning capabilities helps to adjust the composite scheme with any changes that occur when performing linkage. The approach is demonstrated using a linkage scenario where researchers apply to link data from an Australia-wide national type-1 diabetes platform with survey results from 25,000+ Victorians based on their health and wellbeing. With the focus of identity protection, three privacy-preserving approaches were analysed using practical metrics. The findings showed the efficiency of linkage units and could help to minimise the need for manual review of linkage results before release, which is still the mainstream when dealing with data linkage requests.

It is noted that we are not proposing to entirely replace *linkage k-anonymity* or *k-anonymity* with *SLKA*, but rather to have the method available when conditions are satisfied. The adoption criteria of such approaches may vary from case to case, depending on factors such as the acceptable overlapping rate, dataset sizes, attribute numbers and value hierarchies. We have shown that semantic verification against implicit associations through *SLKA* maintains aggregate values and thus avoids the excessive data obfuscation that occurs whilst protecting against individual re-identification risks. The *SLKA* framework provides dynamic protection for repeated linkage releases while preserving data utility by avoiding unnecessary generalisation as typified by *k-anonymity*.

Several future research topics are identified based on this work. First, we have identified that the overlapping rate should affect the method selection, therefore there is a need for diverse sample-based tests to justify and refine the metrics, i.e. considering how many attributes and records to use without knowing of the presence of an individual. Second, it would be interesting to consider circumstances where different minimal support and confidence values were applied for association mining. In this work, we only consider full-chance associations, however, other possibilities may also be explored. Third, we explored *SLKA* with two data-sets but different datasets may have distinct demands not represented in this example. For this reason, it is essential to continue the experiments on different types of datasets with a wide range of privacy requirements and data demands.

Even with the potential of semantic-based reasoning which can require extra generalisation and hence information loss, the acceptance of solutions put forward in this paper will take time to be adopted due to the risk aversion of many health organisations.

ACKNOWLEDGMENTS

The research reported in this paper is supported by the ADDN project funded by the Juvenile Diabetes Research Foundation and the VicHealth project funded by the Department of Health and Human Services, Victoria. We gratefully acknowledge their support.

REFERENCES

[1] Biro, S., Williamson, T., Leggett, J. A., Barber, D., Morkem, R., Moore, K. & Janssen, I. (2016). Utility of linking primary care electronic medical records with Canadian census data to study the determinants of chronic disease: an example based on socioeconomic status and obesity. BMC medical informatics and decision making, 16(1), 1.

[2] Winkler, W. E. (2006). Overview of record linkage and current research directions. In Bureau of the Census.

[3] Centre for Health Record Linkage. http://www.cherel.org.au/. Accessed on 20th January 2017.

[4] SA-NT DataLink. https://www.santdatalink.org.au/. Accessed on 20th January 2017.

[5] Data Linkage – Western Australia. http://www.datalinkage-wa.org.au/. Accessed on 20th January 2017.

[6] Kelman, C. W., Bass, A. J., & Holman, C. D. J. (2002). Research use of linked health data—a best practice protocol. Australian and New Zealand journal of public health, 26(3), 251-255.

[7] Baldwin, E., Johnson, K., Berthoud, H., & Dublin, S. (2015). Linking mothers and infants within electronic health records: a comparison of deterministic and probabilistic algorithms. Pharmacoepidemiology and drug safety, 24(1), 45-51.

[8] Alexandros Karakasidis and Vassilios S. Verykios. 2012. Reference table based k-anonymous private blocking. In Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12). ACM, New York, NY, USA, 859-864. DOI=http://dx.doi.org/10.1145/2245276.2245444

[9] W.W. Cohen, P. Ravikumar, and S. Fienberg, "A Comparison of String Distance Metrics for Name-Matching Tasks," Proc. Workshop. Information Integration on the Web (IJCAI '03), 2003.

[10] P. Christen, "Febrl: An Open Source Data Cleaning, Deduplication and Record Linkage Systemwith a Graphical User Interface," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 1065-1068, 2008.

[11] Holman, C. A. J., Bass, A. J., Rouse, I. L., & Hobbs, M. S. (1999). Population-based linkage of health records in Western Australia: development of a health services research linked database. Australian and New Zealand journal of public health, 23(5), 453-459.

[12] Atchinson, B. K., & Fox, D. M. (1997). The politics of the health insurance portability and accountability act. Health Affairs, 16(3), 146.

[13] Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

[14] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[15] Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006, April). l-diversity: Privacy beyond k-anonymity. In Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on (pp. 24-24). IEEE.

[16] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In 23rd International Conference on Data Engineering (pp. 106-115). IEEE.

[17] Lu, Y., & Sinnott, R. O. (2016, August). Semantic-Based Privacy Protection of Electronic Health Records for Collaborative Research. In Trustcom/BigDataSE/ISPA, 2016 IEEE (pp. 519-526). IEEE.

[18] Goryczka, S., Xiong, L., & Fung, B. C. (2014). Privacy for Collaborative Data Publishing. IEEE. Transactions on Knowledge and Data Engineering, 26(10), 2520-2533.

[19] Jurczyk, P., & Xiong, L. (2009). Distributed anonymization: Achieving privacy for both data subjects and data providers. In IFIP Annual Conference on Data and Applications Security and Privacy (pp. 191-207). Springer Berlin Heidelberg.

[20] Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 265-273). ACM.

[21] Xiao, X., & Tao, Y. (2007). M-invariance: towards privacy preserving re-publication of dynamic datasets. In Proceedings of the 2007 ACM SIGMOD international conference on Management of data (pp. 689-700). ACM.

[22] Anjum, A., & Raschia, G. (2013). Anonymizing sequential releases under arbitrary updates. In Proceedings of the Joint EDBT/ICDT 2013 Workshops (pp. 145-154). ACM.

[23] Dankar, F. K., & El Emam, K. (2012). The application of differential privacy to health data. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 158-166). ACM.

[24] Atzori, M. (2006). Weak k-anonymity: A low-distortion model for protecting privacy. In International Conference on Information Security (pp. 60-71). Springer Berlin Heidelber

[25] Paci, F., & Zannone, N. (2015). Preventing Information Inference in Access Control. In Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (pp. 87-97). ACM.

[26] Shlomo, N., & Young, C. (2006, November). Statistical disclosure control methods through a risk-utility framework. In Privacy in Statistical Databases (pp. 68-81).

[27] Domingo -Ferrer J, Martinez-Balleste A, Mateo-sanz JM, Sebé F (2006) Efficient multivariate data-oriented microaggregation. VLDB J 15:355–36

[28] Lu, Y., Sinnott, R. O., & Verspoor, K. (2017). A Semantic-Based K-Anonymity Scheme for Health Record Linkage. Studies in health technology and informatics, 239, 84-90.

[29] Domingo-Ferrer, J., Sánchez, D., & Rufian-Torrell, G. (2013). Anonymization of nominal data based on semantic marginality. Information Sciences, 242, 35-48.

[30] Abril, D., Navarro-Arribas, G., & Torra, V. (2010). Towards semantic microaggregation of categorical data for confidential documents. In International Conference on Modeling Decisions for Artificial Intelligence (pp. 266-276). Springer Berlin Heidelberg.

[31] Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 279-288). ACM.

[32] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM

[33] Australian Statistical Geography Standard (ASGS, 2011). http://www.abs.gov.au/websitedbs/D3310114.nsf/home/Australian+Statistical+Geography+Standard+(ASGS).

[34] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific american, 284(5), 28-37.

[35] SNOMED Clinical Terms technical specifications (2000). College of American Pathologists. http://www.snomed.org. Accessed on 20th January 2017.

[36] Wang, H., Han, J., Wang, J., & Wang, L. (2013). (k, ε)-Anonymity: An anonymity model for thwarting similarity attack. In Granular Computing (GrC), 2013 IEEE International Conference on. pp. 332-337. IEEE.

[37] Landberg, A. H., Rahayu, J. W., & Pardede, E. (2011). n-Dependency: dependency diversity in anatomised microdata tables. Logic Journal of IGPL, 19(5), 679-702.

[38] World Health Organization (1992). International statistical classification of disease and related health problems, Tenth Revision (ICD-10), Geneva.

[39] Wang, K., Fung, B. C., & Yu, P. S. (2005). Template-based privacy preservation in classification problems. In Fifth IEEE International Conference on Data Mining (ICDM'05) (pp. 8-pp). IEEE.

[40] Wang, K., Fung, B. C., & Philip, S. Y. (2007). Handicapping attacker's confidence: an alternative to k-anonymization. Knowledge and Information Systems, 11(3), 345-368.

[41] Mohammed, N., Fung, B., Hung, P. C., & Lee, C. K. (2009). Anonymizing healthcare data: a case study on the blood transfusion service. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. (pp. 1285-1294). ACM.

[42] Miaoulis, W. M. (2010). Access, use, and disclosure: HITECH's impact on the HIPAA touchstones. Journal of AHIMA, 81(3), 38-39.

[43] Department of Health and Human. VicHealth. https://www.vichealth.vic.gov.au/. Accessed on 20th January 2017

[44] Australasian Diabetes Data Network. www.addn.org.au/. Accessed on 20th January 2017

[45] Australian Standard Classification of Languages (ASCL, 2011). http://www.abs.gov.au/ausstats/abs@.nsf/mf/1267.0.

[46] Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG, 2011). http://www.abs.gov.au/ausstats/abs@.nsf/mf/1249.0.

[47] Ayala-Rivera, V., Murphy, L., & Thorpe, C. (2016). Automatic Construction of Generalization Hierarchies for Publishing Anonymised Data. In International Conference on Knowledge Science, Engineering and Management (pp. 262-274). Springer International Publishing.

[48] Prasser, F., & Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The ARX data anonymization tool. In Medical Data Privacy Handbook (pp. 111-148). Springer International Publishing.

[49] Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., & Kuhn, K. A. (2012, September). Flash: efficient, stable and optimal k-anonymity. In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom) (pp. 708-717). IEEE.

[50] El Emam, K., Dankar, F. K., Issa, R., Jonker, E., Amyot, D., Cogo, E., ... & Roffey, T. (2009). A globally optimal k-anonymity method for the de-identification of health data. Journal of the American Medical Informatics Association, 16(5), 670-682.

[51] Hertog, M. G., Kromhout, D., Aravanis, C., Blackburn, H., Buzina, R., Fidanza, F., ... & Pekkarinen, M. (1995). Flavonoid intake and long-term risk of coronary heart disease and cancer in the seven countries study. Archives of internal medicine, 155(4), 381-386.