Integrating and Connecting Care
A. Ryan et al. (Eds.)

doi:10.3233/978-1-61499-783-2-84

A Semantic-Based K-Anonymity Scheme for Health Record Linkage

Yang LU[1], Richard O. SINNOTT and Karin VERSPOOR
Department of Computing and Information System,
The University of Melbourne, Melbourne, Australia

Abstract. Record linkage is a technique for integrating data from sources or providers where direct access to the data is not possible due to security and privacy considerations. This is a very common scenario for medical data, as patient privacy is a significant concern. To avoid privacy leakage, researchers have adopted *k-anonymity* to protect raw data from re-identification however they cannot avoid associated information loss, e.g. due to generalisation. Given that individual-level data is often not disclosed in the linkage cases, but yet remains potentially re-discoverable, we propose *semantic-based linkage k-anonymity* to de-identify record linkage with fewer generalisations and eliminate inference disclosure through semantic reasoning.

Keywords. Medical record linkage, de-identification, k-anonymity, semantic reasoning

Introduction

In the biomedical field, record linkage has been recognised as a key approach used to support in-depth research on areas including public health and individual well-being. Different from two-party protocols where only two database owners participate in linkage process, a trusted third party is often adopted where records are sent from distributed sources and used for healthcare and medical research [1]. For instance, the Centre for Health Record Linkage (CHeReL, http://www.cherel.org.au/) uses probabilistic matching on demographic data to create linked health records across the New South Wales and Australian Capital Territory. Using the "Master Linkage Key" (MLK) generated from the matching process, record linkage is forged according to the attributes requested by users. Due to the sensitivities of health information, record linkage typically needs to be de-identified before being released to applicants. However existing methods are often vulnerable to re-identification caused by skewed distributions and data dependencies (e.g. equivalent, inclusive relations) among attributes. To tackle this issue, we propose the linkage anonymity scheme with semantic verification that ensures that latent privacy leakage can be detected and prevented from occurring. This is the focus of this paper.

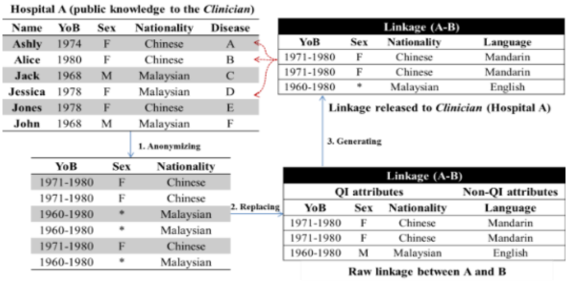
[1] Corresponding Author: PhD candidate Yang Lu, Department of Computing and Information System, The University of Melbourne, Parkville VIC 3010; Email: luy4@student.unimelb.edu.au.

1. Privacy Preservation for Record Linkage

Security models designed for the health records are typically based on the Health Insurance Portability and Accountability Act of 1996 (HIPAA) involve removing or obfuscating identifying information, limiting unnecessary access and separating attributes that can be used for potential individual disclosure [2]. However by using background knowledge from disclosure files (DFs) it is the case that individuals in such data can be inferred (re-identified) by internal users[2]. As one example, Mr. Smith is the only patient over 80 years old in a given cancer registry. If his clinicians know this by accessing his raw records, then such minor facts about non-identifiable attributes (e.g. Age>80) may lead to re-identification. To tackle this background leakage issue, Sweeney (2002) proposed the *classic k-anonymity* processing quasi-identifiers (*QIs*) to satisfy privacy requirements, i.e. any individuals represented in a released data set must be indistinguishable from at least *k-1* other individuals [3]. To achieve this, attributes need to be generalised (suppressed) until there exist at least k identical records before the dataset can be released. To reduce the impact on the quality of information [4], we propose *linkage k-anonymity* (LA) by which (obfuscated) individuals in a released linkage set are required to be indistinguishable from at least *k-1* other individuals in the local dataset. The idea behind this is that most linkage cases do not include all local patients and thus not all modifying data for privacy-preserving purposes is used. To explain this, Figure 1 shows a scenario where record linkage is used through the *LA* method. Suppose clinicians working at Hospital A apply to have the linkage between their dataset 'Hospital A' and the external data set 'Pharmacy B' supported. Instead of processing the linkage on the *QI* union {*Year of Birth (YoB)*, *Sex*, *Nationality*, *Language*} to meet the requirement $k_{linkage}$ composed of local *k* values[3], *LA* will only transform the local dataset that may be possibly known by the requestors, e.g. executing *3-anonymity* on the local *QI* attributes {*YoB*, *Sex*, *Nationality*} in Hospital A and replacing the raw tuples in the linkage set with generalised records so that users have 1/3 chance (at most) to re-identify patients by matching with local records. For the tuple <*1971-1980, F, Chinese, Mandarin*> in the linkage set, three individuals (Ashly, Alice and Jessica) are matched at Hospital A and thus meet the requirement $k_{linkage}$=3. Therefore, *LA* provides the same privacy-preserving effect as the classic anonymity method by distinguishing *QI* and *Non-QI* attributes (i.e. *QI* attributes only in Pharmacy B) on a case-by-case basis, whilst using classic *k-anonymity* on the linkage set results in more-transformed tuples, e.g. <*1960-1980, *, Asian, **> and causes more data loss.

[2] *Internal user* with regards to a linkage project refers to requestors who are authenticated by related databases and thus have access to certain information of data owners (patients).

[3] $k_{linkage}$ refers to the maximum *k* among the member datasets, i.e. $max\{k_1, ..., k_n\}$.



Hospital A (public knowledge to the *Clinician*)

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

1. Anonymizing

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

2. Replacing

Linkage (A-B)

QI attributes			Non-QI attributes
YoB	Sex	Nationality	Language
1971-1980	F	Chinese	Mandarin
1971-1980	F	Chinese	Mandarin
1960-1980	M	Malaysian	English

Raw linkage between A and B

3. Generating

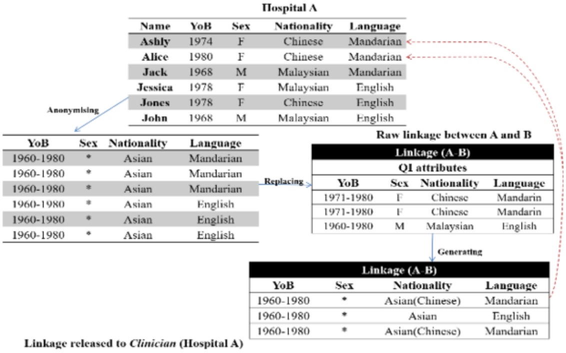
Linkage (A-B)

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

Linkage released to *Clinician* (Hospital A)

Figure 1. Linkage processed with *linkage 3-anonymity.*

Applying syntax-based transformation alone may not be sufficient to prevent privacy leakage occurring since any changes in privacy policies at local sites may impact the linkage anonymity in terms of *k* values and *QIs*. For instance, from the linkage released in Figure 1, it is not difficult for users to identify the association *Mandarin (**Language**) → Chinese (**Ethnicity**)*. As a result, Hospital A could request the same linkage while additionally using *Language* as the fourth *QI* locally. As shown in Figure 2, by executing the *LA* on the full scheme, linkage tuple <*1960-1980, *, Asian, Mandarin*> can be generated to match three individuals (Alice, Ashly and Jack). However, based on the association, the tuple can be refined as <*1960-1980, *, Chinese, Mandarin*>. As a result, the previous linkage release can cause privacy violations by increasing the chance of re-identification from 1/3 to 1/2. Although the *Language* itself does not help re-identify patients, N-gram associations can be utilised to refine values and subsequently increase the risk of potential re-identification of individuals.



Hospital A

Name	YoB	Sex	Nationality	Language
Ashly	1974	F	Chinese	Mandarian
Alice	1980	F	Chinese	Mandarian
Jack	1968	M	Malaysian	Mandarian
Jessica	1978	F	Malaysian	English
Jones	1978	F	Chinese	English
John	1968	M	Malaysian	English

Anonymising

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

Replacing

Raw linkage between A and B

Linkage (A-B)

QI attributes			
YoB	Sex	Nationality	Language
1971-1980	F	Chinese	Mandarin
1971-1980	F	Chinese	Mandarin
1960-1980	M	Malaysian	English

Generating

Linkage (A-B)

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

Linkage released to *Clinician* (Hospital A)

Figure 2. Linkage processed with *linkage 3-anonymity* (scheme updated).

2. Method - Semantic-based Linkage Anonymity

General solutions for inference disclosure involve ruling out risky associations from previous linked data releases. Current research in this direction focuses on association rule mining which deals with transaction records with “0/1” values marking the appearance of items and numerically calculating the confidence of the association evaluation [5]. The Eq. (1) is defined to identify the association rule such as X→Y by satisfying certain conditions.

$$\text{Support}_x = \frac{|R_x|}{|R_{\text{linkage}}|} \geq S\%;\ \text{Confidence}_{x \to y} \geq C\% \qquad (1)$$

where $|R_{\text{linkage}}|$ = the number of records in the linkage set
$|R_{x\ (x,\ y)}|$ = the number of records contain the tuple x (or both tuple x and y)
S% = the minimum numeric support level
C% = the minimum numeric confidence level

Local schemes can be freely and frequently updated. Therefore, it is necessary to enable semantic-based verification to anonymous tuples. Figure 3 shows a class diagram modelling components of *linkage k-anonymity (LA)* framework. Based on the rules defined in Table 1, two procedures - LA scheme formation and semantic-based risk analysis are implemented to support linkage de-identification. Upon receiving a request to link databases *DB1* and *DB2* as linkage *Ln1*, semantic rules will be used for reasoning to track and compose the *classic k-anonymity* requirements in the related databases, e.g. *hasAnoReq(DB1, 2) hasAnoReq(DB2, 3)* where 2 and 3 are their respective k values. With regards to completeness, Rule 2 speculates that linkage will be processed with the highest requirement of all datasets involved. As a result, the $k_{linkage}$ will be calculated and then enforced to the linkage case *Ln1* such as *hasLnAnoReq(Ln1, 3)*. Instead of taking all *QIs* of databases, the linkage *QIs* should be determined so as to reduce the amount of data generalisation. Therefore, Rules 3-4 are reasoned about to identify linkage *QIs* based on the relationship between requestor roles and databases. For example, dealing the linkage request *req1* with the facts like *hasRole(req1, Clinician)*, *hasResource(req1, Ln1)*, *authenticate*(*DB1, Clinician*) and *linkFrom(Ln1, DB1)*, only the *QIs* in *DB1* will be utilised as linkage *QIs*, such as *hasQI(DB1, Gender)* and *hasLnQI(Ln1, Gender)*. As a consequence of semantic reasoning, the linkage set is anonymised with ever-changing *QIs*, which are aligned with *classic k-anonymity* approaches (see details in Section 1).

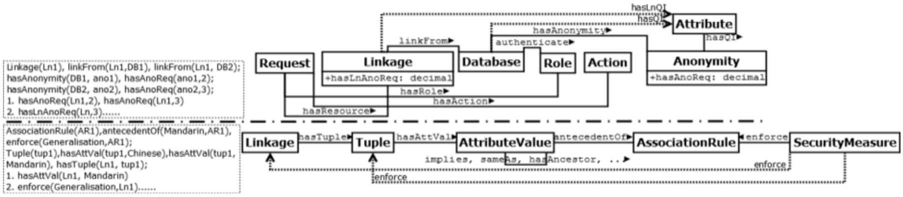

hasLnQI
hasQI
Attribute
hasAnonymity
hasQI
linkFrom
authenticate
Linkage(Ln1), linkFrom(Ln1,DB1), linkFrom(Ln1, DB2);
hasAnonymity(DB1, ano1), hasAnoReq(ano1,2);
hasAnonymity(DB2, ano2), hasAnoReq(ano2,3);
1. hasAnoReq(Ln1,2), hasAnoReq(Ln1,3)
2. hasLnAnoReq(Ln,3)......
Request
Linkage
+hasLnAnoReq: decimal
Database
Role
Action
Anonymity
+hasAnoReq: decimal
hasRole
hasAction
hasResource
AssociationRule(AR1),antecedentOf(Mandarin,AR1),
enforce(Generalisation,AR1);
Tuple(tup1),hasAttVal(tup1,Chinese),hasAttVal(tup1,
Mandarin), hasTuple(Ln1, tup1);
1. hasAttVal(Ln1, Mandarin)
2. enforce(Generalisation,Ln1)......
Linkage
hasTuple
Tuple
hasAttVal
AttributeValue
antecedentOf
AssociationRule
enforce
SecurityMeasure
implies, sameAs, hasAncestor, ..
enforce
enforce

Figure 3. Semantic notations for *LA* execution.

Privacy violations due to inference can be avoided through enforcing semantic-based risk analysis on anonymous tuples. As introduced, associated attributes identified from 'previous release' such as *Mandarin → Chinese* are used to verify the privacy of attribute values of 'current release'. For instance, hierarchical attributes about Ethnicity[4] and Language[5] provides semantic notations expressed as ***Language****(7104-Mandarin)*, ***Ethnicity****(6101-Chinese),* *hasAncestor*(*7104-Mandarin,71-Chinese*),

[4] Australian Statistic Bureau. Australian Standard Classification of Cultural and Ethnic Groups, 2016.

[5] Australian Statistic Bureau. Australian Standard Classification of Languages, 2016.

hasAncestor(*6101-Chinese,61-Chinese Asia*) as well as potential associations *implies(7104-Mandarin, 6101-Chinese)*. Suppose some rule antecedents (e.g. *7104-Mandarin*) are required to be generalised once they appear in linkage tuples. Therefore, we can attach the security measure to association rules such as ***AssociationRule****(AR1)*, *antecedentOf(7104-Mandarin, AR1)* and *enforce(Generalisation, AR1)*. Through locating attribute values in the tuple *tup1* (Rule 5), both Rule 6 and Rule 7 are reasoned about to eliminate potential leakage from anonymised linkage. With the results *enforce(Generalisation, tup1)* and *regarding(tup1, 7104-Mandarin)*, it is suggested replacing the value *7104-Mandarin* with *71-Chinese* to protect ethnicity details from being refined in the relevant records, such as *61-Chinese Asia (includes Mongolia)*.

Table 1. Semantic rules for linkage anonymity.

Purpose	Semantic Rules
Scheme Formation	1. Linkage(?ln), Database(?db), linkFrom(?ln,?db), hasAnonymity(?db,? ano), hasAnoReq(?ano, ?n#decimal) →hasAnoReq(?ln, ?n#decimal)
	2. Linkage(?ln), hasAnoReq(?ln,?n1#decimal), hasAnoReq(?ln,?n2#decimal), greaterThan(?n1#decimal, ?n2#decimal) → hasLnAnoReq(?ln,?n1#decimal)
	3. Database(?db), hasAnonymity(?db,?ano), hasQI(?ano,?qi) →hasQI(?db,?qi)
	4. Request (?req), hasResource (?req, ?res), hasRole(?req,?role), linkage(?res), linkFrom(?res,?db), authenticate(?db,?role), hasQI(?db,?qi) → hasLnQI(?res,?qi);
Risk Analysis	5. Linkage(?ln), hasTuple(?ln,?tup), hasAttrVal(?ln,?val) → hasAttVal(?ln,?val)
	6. Linkage(?ln), hasAttVal(?ln,?att), AssociationRule(?ar), antecedentOf(?att,?ar), enforce(?me,?ar) → enforce(?me,?ln)
	7. Linkage(?ln), enforce(?me,?ln), hasAncestor(?val1,?val_1), hasTuple(?ln,?tup), hasAttVal(?tup,?val2), hasAttVal(?tup,?val_1), implies(?val2,?val1)→enforce(?me,?tup), regarding(?tup,?val2)

3. Case Study – Simulated linkage between ADDN and VicHealth Clients

The experimental data used in this case study was based on 1000 patients collected from the Australasian Diabetes Data Network (ADDN, http://www.addn.org.au/) and 1850 respondents from the Victorian Health Promotion Foundation (VicHealth, https://www.vichealth.vic.gov.au/). Through allocating 500 individuals in both systems with different attributes, experiments were performed on a laptop with Windows 7 operation system (3.20 GHz Intel Core processor and 8GB Memory) to compare the performance of *linkage k-anonymity (LA)* and *semantic-based linkage k-anonymity (SLA)* through using *Scheme1* and *Scheme2* at *time1* and *time2* (*time1<time2*). Due to the attribute accumulation, association rules mined from *time1* release can be included into the knowledge base to deal with the repeated request at *time2*.

3.1. Metrics of Privacy Cost and Utility

Privacy cost refers to the chance of re-identification by adversaries. Given the 'unknown presence' in the linkage set, Eq. (2) can be defined to measure the disclosure risk of each individual shared by all databases. In particular, $Pr(CI_i)$ stands for the possibility of uniquely identifying CI_i from the equivalence class, $Class_i$ to which the linkage record matches.

$$\mathrm{Pr}(\mathrm{CI_i}) = \frac{1}{|\mathrm{Class_i}|} \tag{2}$$

Based on the individual possibility, the average possibility can be calculated as the measure of privacy cost incurred by the linkage set, as defined in Eq. (3). Specially, *n* refers to the number of common individuals in the linkage set.

$$\mathrm{Risk}= \frac{1}{\mathrm{n}}\sum\nolimits_{\mathrm{n}}\mathrm{Pr}(\mathrm{CI_i}) \tag{3}$$

The function 'Sum of Squared Error (SSE)/Sum of Squared Table (SST)' was applied to measure the information loss of micro-aggregating values in equivalence groups [8]. Since the within-groups SSE is never greater than SST, reported values (%) are in the range of [0,100]. Dealing with categorical attributes, the original and modified values can be quantified by the level of hierarchies they represent. Specially, the SSE (SST) is calculated by the Eq. (4).

$$\mathrm{SSE}=\sum\nolimits_{\mathrm{n}}\sum\nolimits_{\mathrm{m}} \mathrm{dis}(\mathrm{x_{ij}}, \mathrm{x'_{ij}})^2 \tag{4}$$

where $\mathrm{x_{ij}}$ = original attribute value
$\mathrm{x'_{ij}}$ = anonymised attribute value ($\mathrm{x'_{ij}} = 0$ once calculating SST)
dis() = distance between attributes within the hierarchical structure
m = the number of quasi-identifiers in the linkage set

3.2. Evaluation Result and Discussion

Table 2 compares the performance of de-identifying linkage with *LA* and *SLA*. With different sets of *linkage QIs*, the impact caused by local changes is apparent. At *time1*, both approaches perform identically in terms of privacy preservation since there is no available knowledge at that given time. The anonymity requirement can never decrease, i.e. once data is linked and released the risk can never diminish, hence the disclosure risk can be calculated based on the associations mined from the previous *2-anonymised* linkage at *time1*. With *Language* information representing the fourth *QI* in ADDN, ADDN-VicHealth linkage will be anonymised based on *Scheme2*. As shown, the actual disclosure risk by using *LA* is equal to or higher than the *SLA*, with the result comparison 14.7% vs 14.7% (k=2), 5.0% vs 4.3% (k=3) and 4.7% vs 3.4% (k=4). This shows that with privacy verification, certain risky information will be detected and processed by using *SLA* to anonymise linkage. Similar to the privacy analysis, the utility comparison between *LA* and *SLA* is conducted under the temporal consideration. There is an increased data loss incurred with higher requirements under both schemes. Based on mining previous releases, the semantic approach becomes effective for risky values whenever the same linkage request is applied again, i.e. *LA* may be able to preserve more information however it runs a higher risk of re-identification than *SLA*-based linkage. Although the verification results in data generalisation, the major improvement in data quality is due to the linkage *QI* attribute filter.

Table 2. Privacy Cost and Information Loss by using *LA* and *SLA*.

LA and SLA		**Time1: QI= {Gender, Ethnicity, Postcode}**			**Time2: QI={Gender, Ethnicity, Postcode, Language};**		
		k=2	k=3	k=4	k=2	k=3	k=4
LA	Disclosure risk (%)	14.7	5.2	3.5	14.7	5.0	4.7
	SSE/SST (%)	13.9	20.1	22.1	25.3	27.0	28.5
SLA	Disclosure risk (%)	14.7	5.2	3.5	14.7	4.3	3.4
	SSE/SST (%)	13.9	20.1	22.1	25.3	27.3	28.9

4. Conclusions and Future work

In this paper, we propose a *semantic-based linkage k-anonymity* (*SLA*) approach based on *k-anonymity* and linkage properties with the aim of eliminating privacy disclosure risks while preserving data utility. In the future, we will further explore semantic approaches in privacy preserving record linkage (PPRL), while protecting patient privacy, where the accuracy of record matching should be maintained.

References

[1] F. Young, A. J. Dobson & J. E. Byles (2001). Health services research using linked records: who consents and what is the gain?. *Australian and New Zealand journal of public health*, 25(5), 417-420.

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

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

[4] G. Loukides, A. Gkoulalas-Divanis & J. Shao (2011). On balancing disclosure risk and data utility in transaction data sharing using RU confidentiality map. *Proceedings of the Joint UNECE/Eurostat Working Session on Statistical Data Confidentiality*, 19.

[5] R. Agrawal, T. Imieliński & A. Swami (1993, June). Mining association rules between sets of items in large databases. *ACM Sigmod Record*, 22(2), 207-216. ACM.

[6] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz & F. Sebé (2006). Efficient multivariate data-oriented micro-aggregation. *The VLDB Journal—The International Journal on Very Large Data Bases*, 15(4), 355-369.