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

Semantic-based Policy Composition for Privacy-demanding Data 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

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

***Abstract*— Record linkage can be used to support current and future health research across populations however such approaches give rise to many challenges related to patient privacy and confidentiality including inference attacks. To address this, we present a semantic-based policy framework where linkage privacy detects attribute associations that can lead to inference disclosure issues. To illustrate the effectiveness of the approach, we present a case study exploring health data combining spatial, ethnicity and language information from several major on-going projects occurring across Australia. Compared with classic access control models, the results show that our proposal outperforms other approaches with regards to effectiveness, reliability and subsequent data utility.**

Keywords-record linkage; association rules; policy composition; semantic web technology

I. INTRODUCTION

In the biomedical arena, the secondary use of electronic health records (EHRs) for research purposes can accelerate new discoveries including optimized medication and treatments, improved surgical procedures, through to population profiling and health benchmarking. Record linkage has been recognized as a key technique underpinning healthcare and public health research at the state and national levels, as it allows access to and use of cross-jurisdictional data such as hospital admissions data, treatment reports, prescriptions and death reports. Record linkage has been applied in numerous diverse projects, e.g. exploring the correlation between obesity and socio-economic status in Canada [1], understanding lung cancer treatment and the mortality of aboriginal people in the New South Wales (NSW) [2] amongst many other examples etc. Although anonymisation and confidentiality are essential considerations for biomedical data management, little technical work has been done to preserve privacy of linked records in an automated manner, which is cognizant of data leakage and potential inference risks. With the explosive growth of data, it is increasingly difficult to depend solely on stakeholders/ethics committees to identify all potential security issues and take measures to protect against them. Rather, linkage infrastructures ought to be designed to extend static data access requests with more dynamic query capabilities, ensuring the linkage risks are evaluated and minimized in an automatic manner. This is the motivation of this work.

II. RELATED WORK

Technically, access control represents the most commonly used technique to regulate security based on the paradigm of "*who can do what upon which resource*". Working in different contexts, access control policies have been defined reflecting a variety of stakeholders' needs and demands in protecting access to their resources and services. For instance, privileges are often associated with "roles" and assigned to individuals, e.g. through Role-Based Access Control (RBAC) systems [3] that are subsequently used to enforce access control decisions through local policies. To improve the applicability of RBAC policies [4], Attribute-based access control (ABAC) models were introduced in complex scenarios. Many of these are based on eXtensible Access Control Markup Language (XACML)-based policies. ABAC models can be used to provide context-aware access control, e.g. location/temporal-aware services in mobile *ad hoc* networks; purpose-based authorization decisions regarding access to medical information; relationship-based interactions in social networks, as well as organization-based exchange for e-Business purposes [5] [6] [7] [8] [9] [10]. However, this *black-and-white* authorization design is inadequate in supporting database system where data queries require "middle ground" security. Chaudhuri *et al.* (2011) claimed that "*authorizing users to access a subset of the data in request*" is becoming the mainstream model in most data management practices [11]. Niet *et al.* (2010) proposed advanced authorization solutions where privacy rules were extended through "Obligation" components that were enforced in policy decisions [12]. In this model, resource/user privacy preservation was seamlessly deployed in systems built upon RBAC/ABAC.

Policy composition is often necessary when dealing with distributed databases as conflicting behaviors can arise among policy domains (action/role/resource) of the collaborating parties. To tackle this issue, several frameworks have been used for policy conflict detection/resolution. For instance, Wang *et al.* (2014) devised a conflicting algorithm through building a "purpose tree model" based on the idea that "*privacy policies are concerned with which data object is used for what purpose*" [13]. Through matching the "*purpose*" and "*obligation*", they were able to identify conflicting policies that could be solved by use of obligations. A strategy-based approach was proposed including support for *Recency-Override*, *Specificity-Override* and *Deny-Override* [14]. Since the "precedence strategy" may not always be able to resolve conflicts issued by hierarchical authorities, conflict graphs can

2324-9013/18/31.00 ©2018 IEEE
DOI 10.1109/TrustCom/BigDataSE.2018.00060

be formed where resolutions are defined in a context-aware manner [15]. In addition to authorizing decisions, Lupu and Sloman (1999) identified modality conflicts through considering both authorized and obliged behaviors [16]. Specific to XACML-based applications, a standard resolution framework was designed including the conflict-resolution strategies such as *Permit-/Deny-Override*, *Only-One-Applicable* and *First-Applicable* [17]. Inspired by service discovery in cloud environments, Lin *et al.* (2013) proposed a policy similarity measurement in XACML. The principle here was based on the composition that occurs among similar policies and how this can minimize system resources while preserving the original purposes to the greatest extent. Specifically, policy candidates were decomposed into atomic elements e.g. rules, targets and target elements, and the similarity measured through a weighted distance aggregation [18]. For instance, the *Jaccard Similarity Coefficient* [19] was used for attribute closeness measurements [20]. However, privacy regulations represented as obligations should also be checked for hidden violations to any parties' requirements. This is frequently recognized in data linkage systems however support is limited in mainstream XACML-based applications.

It is possible to combine inference control within a formal policy framework to seamlessly deploy privacy-preserving functions in databases. Inference control techniques are used to tackle unintentional data disclosure inferred from access to seemingly non-sensitive items, e.g. postcodes or ethnicity. To prevent undesirable disclosure arising from such items, propagation through standard taxonomies can be checked, as "reachability" between concepts is considered as a contributing factor to privacy compromises [21]. In other words, sensitiveness can disperse along with hierarchical or other attribute inferences. To further refine inference risk management, Costante *et al.* (2013) showed how to evaluate the security cost from aggregated attributes by considering potential correlations between them [22]. Considering a users' personal information may be inferred from seemingly unrelated items, e.g. matching user age intervals or their gender with preference settings on the Google Ads Preference Manager [23], extended inference closures can be identified and subsequently used for evaluating potential threats, according to the sensitivity level of the inferred items [24].

Traditional policy-based access control applied to databases is often too static to satisfy the demands of many dynamic distributed applications, which rely on real-time integration of data sources or where an access decision depends on the results of queries. To support arbitrary linkage, a syntactic XACML policy is typically not able to answer requests specified using heterogeneous attributes. Furthermore, policy contents should be updated to prevent potential policy violations that might arise through any newly generated facts. To tackle these issues and challenges, existing solutions include semantics-based policy formulation and evaluation. Finin *et al.* (2008) explored the Web Ontology Language (OWL) to represent RBAC models through *role as classes* and *role as instances* [25]. In addition, Cirio *et al.* (2007) considered RBAC expressions through description logics (DL) to improve the semantic understanding needed for many access control scenarios [26]. Priebe *et al.* (2006) proposed an extended XACML architecture where an inference engine was built on a set of semantic rules and attribute ontologies [27]. Similarly, Kim (2013) applied the Resource Description Framework (RDF) to describe attributes that could be used to detect latent conflicts during policy aggregation leveraging semantic reasoning [28]. In terms of strategy utilization, Kolovski *et al.* considered reasoning aspects [29], while Liu *et al.* focused on extensibility and system scalability [30]. In these works, privacy issues were typically considered through obligations used for constraint checking and subsequent granting of access. Given that inference disclosure can often be detected by reasoning about association rules and extensible knowledge bases [31], we consider the enforcement of privacy-oriented security measures including generalization and suppression through associating them with policy obligations. The identification of such association rules across multiple data resources has not been explored and is the focus of this work.

III. RECORD LINKAGE ACCESS CONTROL FRAMEWORK

Record linkage refers to the activity of relating records that belong to the same entity across different data sets. Generally, record linkage refers to the activity of relating records in a data set that refer to the same entity across different data sources. In the biomedical field, linkage helps examine and understand public health issues typically outside of healthcare environments [32]. As a representative example, the Centre for Health Record Linkage (CHeReL) provides an infrastructure for EHR linkage and management [33]. As shown in Fig. 1, patients may be registered in multiple databases and thus have more than one Source Number (SN) (e.g. the A-01 and B-99). Through recognizing records belonging to same individuals, a central component can uniquely assign "*Master Linkage Keys* (MLK)" to data subjects [34]. Data sets may be used (linked) by meeting special needs – often related to anonymization concerns and ethically-driven research. According to Ritchie and Elliot (2015), existing linkage centers mainly rely on the Principle Based Model (PBM), i.e. all outputs should be evaluated by experienced staff since any pre-defined rules are thought to be insufficient to consider the full complexity of privacy [35]. For instance, CHeReL can only release datasets by obtaining approvals from custodians and ethics committees. Similar models have been deployed in Western Australia, Victoria, Southern Australia and Queensland [36] [37] [38] [39]. Undoubtedly, PBM offers maximum flexibility to researchers however it leads to a high cost in training professionals to assess the privacy risks of each linkage request. To avoid this, we propose a hybrid Rule Based Model (RBM) that can be used as a filter ruling out illegal requests and supporting decisions at the PBM level. Ultimately however data linkage should meet any/all overarching privacy needs associated with the individual policy rules from the original data providers.

Fig. 2 shows the linkage authorization within a proposed infrastructure, including the linkage center (service provider) and several EHRs repositories (data owners) during a given collaboration. Upon receiving requests such as *ReqA(Clinician, Read,[SourceID],[Attribute Bag])* the linkage center can locate the targeted datasets and subsequently evaluate the access request based on composite policies. Since policies are defined with heterogeneous information (often with their own namespaces, roles and data attributes), policy composition demands the semantic disambiguation of distributed knowledge. To achieve this, we build ontologies based on the metadata submitted by all of collaborating sites (step 0). The *[SourceID]* points to a fixed tabular row and the *[Attribute Bag]* further refines the columns (attributes names) requested (step 1). In this case, certain patients in registry D are matched (step 2). Through obtaining local policies, the composition occurs using the policy ontology (step 3 & 4). Afterwards, the center is expected to eliminate policy violations to ensure private constraints. This requires that individual policies are tested to ensure no explicit/implicit conflicts arise in granting permission to the data (step 5). At this stage, relevant metadata is used to support semantic reasoning for potential privacy risk disclosure of the combined data sets and the individual policies that are involved (step 6). Finally, anonymizing measures are enforced on data elements to reduce the leakage risks that may have been identified (step 7). At the stage of PBM checking, individual checkers need to check/evaluate the latent leakage specific to the topic and ultimately produce privacy-preserving datasets that can be returned to the user (step 8 & 9).

Cancer Registry			
SN	First name	Last name	Gender
A-01	Christine	Smith	Female

Hospitalisation Record			
SN	Full name	Gender	Postcode
B-99	Christine Smith	Female	3000

CHeReL linkage key		
Linkage ID	$SN_{cancer\ registry}$	$SN_{Hospitalisation}$
L-40	A-01	B-99

Figure 1. Record linkage in CHeReL.

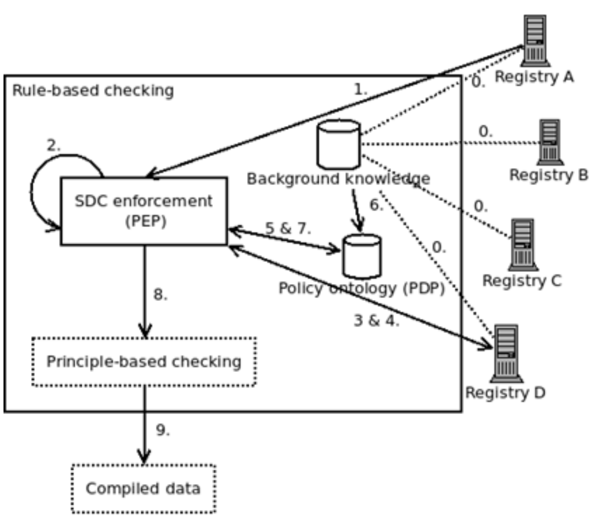


Figure 2. Example interaction for record linkage access.

A. Basic notions in formulating XACML policy

The scenario presented above requires policy expression and composition capabilities. XACML is the natural choice in this work. In [31], we demonstrated how XACML policies could be formulated and evaluated through reasoning over semantic contents. For instance, a pseudo policy (*Policy_1*) in Fig. 3 defines that *"EHRs of type-1 diabetes mellitus (T1DM) patients can be accessed by people who are authenticated as Clinicians who can read the 'de-identified version' for the specified research purpose". Policy_1* becomes applicable only if the requirements in *Tar_a* are satisfied. To generalize this work, we define the abstract XACML semantics as follows:

```
<Policy Id=Policy_1 Algorithm=deny-unless-permit>
 <Target Id = Tar_a>
  <Subject Id=Sub_a AttributeValue =Clinician/>
  <Resource Id=Res_a AttributeValue =T1DMPatients/>
 </Target>
 <Rule Id=Rule_1 Effect=Permit>
  <Target Id = Tar_b>
     <Action Id=Act_a AttributeValue =Read/>
     <Environment Id=Env_a AttributeValue
=ForResearch/>
  </Target>
  </Condition>
 </Rule>
 <Obligation Id = De-identification FulfillOnEffect =
Permit/>
</Policy>
```

Figure 3. Example XACML policy profile.

Definition-1. In a policy domain P_i, *Tar* is a set of targets, *Sub* is a set of subjects, *Act* is a set of actions, *Res* is a set of data elements, *Con* is a set of conditions and A^+/A^- are two types of authorization: *Permission* and *Prohibition*. The rules for constructing P_i are expressed as A_i^+ (P_i, tar_i, $<sub_i, act_i, res_i>$, +, $[con_i]$) and A_i^- (P_i, tar_i, $<sub_i, act_i, res_i>$, -, $[con_i]$) where $tar_i \in Tar$, $res_i \in Res$ and optionally, $con_i \in$ Con. As noted, in addition to authorization, it is often necessary to define obligation rules to refine post-authorization on results. For data-centric systems, obligations can act as a final 'privacy filter' to minimize inference attacks. For instance, informed consent is a typical obligation that has to be satisfied prior to linkage of EHRs for research purposes. In this case, *de-identification* is regarded as a way of supporting anonymizing measures over health information. Those behaviors are legislated by the *Privacy Act 1988* (Cth) (Privacy Principle *9*, *11* and *12*). Therefore, this mandatory operation is defined on target resources with an associated effect *permit* as the trigger event, i.e. they are only checked if the authorization check is "in principle" allow.

Definition-2. An obligation within a policy domain P_i can be expressed as O_i^+ (P_i, tar_i, $<sub_i, act_i, res_i>$, tri_i, fun_i) and O_i^- (P_i, tar_i, $<sub_i, act_i, res_i>$, tri_i, fun_i) where the combination effects within the policy domain act as a "trigger" of functions used for the operations *obliged to do* or *obliged not to do* based on the attributes (subject/resources).

Both the Definition-1 and Definition-2 provide the foundation of the XACML framework. Given demands for scalability in distributed systems, hierarchical models such as RBAC are often used for policy definition and management.

As such, policy composition demands policy extensions that can leverage hierarchy-based propagation when reasoning. In this paper we are primarily interested in the confidentiality of record linkage, hence we assume that access control actions refer to "read" operations.

B. *Propagation on Hierarchical Attributes*

Hierarchical Role Structure. A role hierarchy (RH) can be structured by referring to a standard taxonomy or organization structure [40]. Based on the hierarchical model policy, propagation can be defined as:

Definition-3. In the policy domain P_i the role hierarchy is depicted as RH ::= $\{role_i, \leq \mid i=1\ldots n\}$ where $role_i$ represents each role name in the hierarchy structure and $\leqslant$ stands for relations among these user groups in the role hierarchy. Therefore, the RH-based propagation of authorization and obligation can be expressed as:

$$Propagation_{A^+}^{RH} ::= \{A_i^+(role_i) \to A_i^+(role_j) | RH(role_i \leq role_j)\}$$
$$Propagation_{A^-}^{RH} ::= \{A_i^-(role_i) \to A_i^-(role_j) | RH(role_j \leq role_i)\}$$
$$Propagation_{O^{+(-)}}^{RH} ::= \{O_i^{+(-)}(role_i) \to O_i^{+(-)}(role_j) | A_i^{+(-)}(role_i) \to A_i^{+(-)}(role_j), Trigger(O_i) \equiv Effect(P_i)\}$$

where the condition $Trigger(O_i) \equiv Effect(P_i)$ restricts obligation rules to only trigger when matching the associated condition. Based on the positive rule shown in Fig. 3, Fig. 4 a) describes the propagation where global strategy "Deny-default" is applied. Given the policy *A1(Policy-1, Tar_b <Clinician, Read, T1DMPatients>, +), O1(Policy-1, Tar_a <Clinician, null, T1DMPatients>, +, De-identification)*, the permission can be propagated to any superior roles "Specialized Physician" however the subordinate roles like. "Hospital based dietician" and "Researcher" will be denied by default. In addition, when positively evaluating *Policy-1*, attached *O1* will be executed with *De-identification* to the targeted subject and resource - the Clinician accessing T1MD patient records in this case. Such propagation can also occur with *Permit* as the default result. In this case, role hierarchies are supposed to reflect the organizational authorities. As a result, *A2 (Policy-1, Tar_c <Clinical nurse specialist, Read, T1DMPatients>, -)* in Fig.4 b) should prevent *Clinical nurse specialist* and its subordinate roles reading the diabetes database however the *Specialized physician* will not be affected, i.e. they maintain the initial permission [3]. It is worth noting that adopting a default strategy in an access control system can help overcome possible conflicts in distributed systems however *Permit-default* is not a safe choice since it tends to make data access more readily available (and this is rarely needed).

a) Role hierarchy-based propagation (Deny-default)

b) Role hierarchy-based propagation (Permit-default)

Figure 4. Propagation with *Deny*- and *Permit-default* principles.

Semantically, hierarchical role structure can be formally represented through defining transitive predicates *seniorTo* and *juniorTo*. For example, the role hierarchy in Fig. 4 can be expressed as *seniorTo(Specialized physician, Clinician)*, *seniorTo(Clinician, Hospital based dietician) and juniorTo(Hospital based dietician, Clinician)* etc. With the designated effect, *Permit*, *Policy-1* can be stated with assertions including *hasPermission(Policy-1, A1)* and *hasObligation_P*(*Policy-1, O1*), which are then attached with specific attributes via *hasResource*(*Tar_b, T1DMPatients*) and *hasResource*(*Tar_a, T1DMPatients*).

Propagation based on role hierarchies can be finally expressed via using *enforceOn*, which is dynamically reasoned from semantic rules. For instance, both semantic rule *1-2* define the inner propagation among policy elements while the propagation of permission and positive obligation can be achieved by rules *3-4*. It is noted that *Permission* and *Obligation_P* are the subclass of *Authorisation* and *Obligation,* inheriting the basic propagation implied in rules *1-2*. Likewise, reasoning over negative results such as *Prohibition* and *Obligation_N* relies on the rules *5-6*.

1. *Authorisation(?a), hasTarget(?a, ?t), hasSubject(?t, ?s) → enforceOn (?a, ?s)*
2. *Obligation(?o), hasTarget (?o, ?t), hasSubject (?t, ?s), → enforceOn (?o, ?s)*
3. *Permission(?p), hasSubject(?t, ?s), hasTarget(?a, ?t), enforceOn(?p, ?s), seniorTo(?s', ?s) → enforceOn(?p, ?s')*
4. *Obligation_P (?o), hasSubject(?t, ?s), hasTarget(?o, ?t), enforceOn(?o, ?s), seniorTo(?s', ?s) → enforceOn(?o, ?s')*
5. *Prohibition(?p), hasTarget(?a, ?t), hasSubject(?t, ?s), enforceOn(?p, ?s), seniorTo(?s, ?s') → enforceOn(?a, ?s')*
6. *Obligation_N (?o), hasSubject(?t, ?s), hasTarget(?o, ?t), seniorTo(?r, ?r') → enforceOn(?o, ?r')*

Previous work introduced a semantic approach to reasoning about XACML policies based on heterogeneous attributes from different authorities [41]. Dealing with different policy domains, semantic-based formalization was used to support enhanced reasoning capabilities required when making access control decisions.

Definition-4. Suppose role hierarchies RH_x and RH_y are defined in policy domains P_i and P_j respectively. Using the relationship $RH_x(role_i) \equiv RH_y(role_j)$, propagation can be formed across hierarchies as:

$$Propagation_{A^+}^{RH_{x,y}} ::= \{A_i^+(role_k) \rightarrow A_i^+(role_j) | RH_x(role_k \leq role_i), RH_x(role_i) \equiv RH_y(role_j)\}$$
$$Propagation_{A^-}^{RH_{x,y}} ::= \{A_i^-(role_j) \rightarrow A_i^-(role_i) | RH_x(role_i \leq role_k), RH_x(role_k) \equiv RH_y(role_j)\}$$
$$Propagation_{O^{+(-)}}^{RH_{x,y}} ::= \{O_i^{+(-)}(role_i) \rightarrow O_i^{+(-)}(role_j) | A_i^{+(-)}(role_i) \rightarrow A_i^{+(-)}(role_j), Trigger(O_i) \equiv Effect(P_i)\}$$

For instance, Fig. 5 shows how role hierarchies RH1 and RH2 can be linked with equivalence roles *Diabetic nurse* and *Clinician*. By defining rules *7-8*, cross-domain authorization (obligations) can be realized by introducing equivalent concepts such as *equivalentWith(Diabetic nurse, Clinician)*. According to Definition 3 and Definition 4, more authorization rules can be identified through semantic reasoning such as *A1-extend (Policy-1, Tar_b <Diabetic nurse, Read, T1DMPatients>, +, null)* and *A1-extend (Policy-1, Tar_b <Diabetologist, Read, T1DMPatients>, +, null)*. As a result, the authorization coverage is extended through cross-RH propagation rules, which is especially useful in distributed environments since it offers far more flexibility and can exploit multiple ontologies.

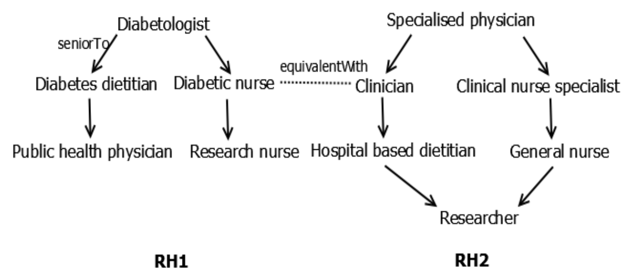


Figure 5. Example propagation cross role hierarchies.

7. *Authorisation(?a), enforceOn(?a, ?r), equivalentWith (?r', ?r) → enforceOn(?a, ?r')*
8. *Obligation(?o), enforceOn(?o, ?r), equivalentWith (?r', ?r) → enforceOn(?o, ?r')*

Hierarchical Data Model. As with role relationships, hierarchical resource profiles are part of the standard grammar of XACML [42]. They were originally defined to support the exponential growth of resource classes. EHRs are often structured as a set of attribute name–value pairs, and thus categorical attributes can be modelled according to their specificity. Hence policies can be refined by specifying the appropriate subset of resources.

Definition-5. In the policy domain P_i, attribute variables of resources can be formulated in *Data Hierarchies* (DH) where DH ::= {$value_i$, ≤ | i= 1…n}. Therefore, the propagation on hierarchical resources can be formed as:

$$Propagation_{A^+}^{DH} ::= \{A_i^+(value_i) \rightarrow A_i^+(value_j) | DH(value_j \leq value_i)\}$$
$$Propagation_{A^-}^{DH} ::= \{A_i^-(value_i) \rightarrow A_i^-(value_j) | DH(value_i \leq value_j)\}$$
$$Propagation_{O^{+(-)}}^{DH} ::= \{O_i^{+(-)}(value_i) \rightarrow O_i^{+(-)}(value_j) | A_i^{+(-)}(value_i) \rightarrow A_i^{+(-)}(value_j), Trigger(O_i) \equiv Effect(P_i)\}$$

It is worth noting that the authorization on *DH* will act on the "resource" while the obligation is specific to the function arguments. Consider an example with positive authorization/obligation applied with overall strategy given as *deny-default*. In this case, access restrictions on ethnicity information can be achieved through obligation *O1(Policy-1, Tar_a<null, null, T1DMPatients>, +,generalization(Ethnicity-1))*[1]. According to the specialty levels, these value hierarchies can be specified through *isA* assertions like *isA*(*3202-Bosnian, 32-South Eastern European*) and *isA*(*32-South Eastern European, 3-Southern and Eastern European*) etc. With this obligation, any data view including unit values *3202-Bosnian* can be replaced by the more general forms (e.g. *32-Southern and Eastern European*). In addition, such a tabular structure can be described through using *hasPatient*, *hasAttribute* assertions associated with row/column names, such as *hasPatient*(*T1DMPatient*, *Patient-1*) and *hasAttribute*(*Patient-1*, *3202-Bosnian*). By reasoning over rules *9 -10* it is possible to associate access control rules to database contents. Through propagation of hierarchical values, such rules can be dynamically executed through reasoning. In this work we focus on access to data resources through authorization and data privacy preservation through obligations. Therefore, Rule *11* focuses on releasing objects (containing data items) which implies a general structure of the contents in the database. For special privacy requirements, Rule *12* is used to target elements, which can then propagate to more specific entities. For instance, *O1* should be used for ethnical contents such as *32-South Eastern European* and according to the reasoning, contents like *3202-Bosnian* should be treated in the same way.

9. *Authorisation(?a), hasTarget(?a, ?t), hasResource (?t, ?r), hasPatient(?r, ?p), hasAttribute(?p, ?e) → enforceOn(?a, ?e)*
10. *Obligation(?o), hasTarget(?o, ?t), hasResource(?t, ?r) hasPatient(?r, ?p), hasAttribute(?p, ?e) → enforceOn(?a, ?e)*
11. *Permission(?a), Ethnicity(?e), Ethnicity(?e'), isA(?e', ?e), enforceOn(?a, ?e') → enforceOn(?a, ?e)*
12. *Obligation_P(?o), Ethnicity(?e), Ethnicity(?e'), isA(?e', ?e), enforceOn(?o, ?e) → enforceOn(?o, ?e')*

C. Inference Disclosure Prevention

Policy composition based on propagation assumes that the information is stable. However, the ever-increasing amount of digital information now available poses threats to privacy protection. Linking records from different custodians can cause privacy issues since heterogeneous policies for datasets can be composed where violations cannot be detected in a timely manner. For instance, obligation *O1(Policy-1, Tar_a <null, null, T1DMPatients>, +, generalization(Postcodes-1))* may not be completely enforced by disclosing spatial information, e.g. the postcode-Statistical Area (SA1) mapping [53] is available to the public and thus may cause inference leakage problems.

Definition-6. Suppose a set of values associated with explicit mappings across DH is expressed as:

$$RV_{er}(value_i, value_j) ::= \{DH_x(value_i) \circ_{er} DH_y(value_j) | \circ_{er} \in ER\}$$

where *Explicit Relation* (ER) refers to the set of explicit mappings in the policy domain. Based on such auxiliary

1 Function *generalization(Ethnicity-1)* is to prevent the access to the unit values in the *Ethnicity* column.

knowledge, it is possible to realize authorization propagation cross the hierarchical resources with obligations formed as:

$$Propagation_{A^{+(-)}}^{DH_{x,y}} ::= \{A_i^{+(-)}(value_i) \rightarrow A_i^{+(-)}(value_j) | RV_{er}(value_i, value_j), \circ_{er} \in \{A_i^{+(-)}, \rightarrow_{authorisation}\}\}$$

$$Propagation_{O^{+(-)}}^{DH_{x,y}} ::= \{O_i^{+(-)}(value_i) \rightarrow O_i^{+(-)}(value_j) | RV_{er}(value_i, value_j), \circ_{er} \in \{O_i^{+(-)}, \rightarrow_{obligation}\}\}$$

Conditions such as $\circ_{er} \in \{A_i^{+(-)}, \rightarrow_{authorisation}\}$ and $\circ_{er} \in \{O_i^{+(-)}, \rightarrow_{obligation}\}$ refer to the predicate $\circ_{er}$ specified in the semantic rules indicates the authorization/obligation propagation. Fig. 6 shows an example of cross-DH propagation. A good practice in formulating pragmatic domain knowledge is to reuse well-known RDF vocabularies such as FOAF [43], SKOS [44], GeoName [45], vCard [46] or Dublin Core [47]. Domain experts should only devise new terms only if existing vocabularies are not sufficient to express the required concepts. For instance, geographical concepts *Postcode-4117* and *SA1-31103131212* are associated by the inclusive relations *dc:isPartOf* and *dc:hasPart*. Through formulating semantic rules for obligation enforcement, it is possible to propagate operations to related contents. Suppose the obligation is defined to enforce one-level generalization (e.g. *Postcode-4117* → *Postcode-411**). Through reasoning Rule 13, security measures can be enforced by replacing the unit content with a more aggregated level (e.g. *SA2-311031312*). Since record linkage should allow data sets to be combined arbitrarily, implied relations can be identified across data models based on value distributions in linkage sets.

13. *Obligation(?o)_P, Postcode(?p), SA(?s), hasPart(?p, ?s), enforceOn(?o, ?p) → enforceOn(?o, ?s)*

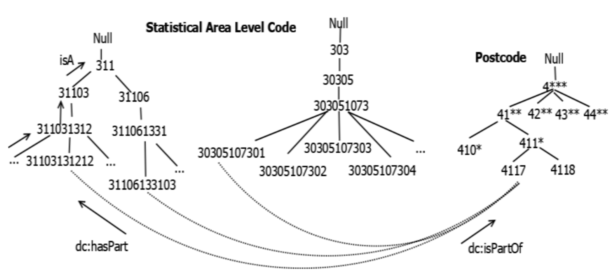


Figure 6. Associated vocabularies with semantic predicates.

Privacy may be threatened by arbitrary linkage where inferences can unintentionally arise. Instead of directly defining policies, the priority is dealing with *implicit associations* that give rise to undesirable inference channels that contribute to latent disclosure leaks. For instance, Chinese children (0-14) are rarely diagnosed with T1DM and thus the appearance of *6101-Chinese* is much lower than the average [48]. Considering arbitrary combinations of linkage requests, special attention should be given to data value distributions. Utilizing mining of association rules, potential associations among heterogeneous variables can be found from evolving data corpora. In this case, personal attributes in the linkage set need to be evaluated with specific attribute combinations and overlapping sizes. Such relations are not limited to single domains and in most cases, they can be used to bridge concepts across domains. Data-centric propagation can subsequently be specified as follows.

Definition-7. Suppose a set of unit-level variables associated with implicit mappings across DHs is expressed as:

$$RV_{ir}(value_i, value_j) ::= \{DH_x(value_i) \circ_{ir} DH_y(value_j) | \circ_{ir} \in IR\}$$

where the *Implicit Relation (IR)* refers to the inference channel impacts on policy decisions. Using auxiliary knowledge, it is possible to realize authorization propagation across sources with obligations formed as:

$$Propagation_{A^{+(-)}}^{DH_{x,y}} ::= \{A_i^{+(-)}(value_i) \rightarrow A_i^{+(-)}(value_j) | RV_{ir}(value_i, value_j), \circ_{ir} \in \{A_i^{+(-)}, \rightarrow_{authorisation}\}\}$$

$$Propagation_{O^{+(-)}}^{DH_{x,y}} ::= \{O_i^{+(-)}(value_i) \rightarrow O_i^{+(-)}(value_j) | RV_{ir}(value_i, value_j), \circ_{ir} \in \{O_i^{+(-)}, \rightarrow_{obligation}\}\}$$

Considering privacy, such associations are produced at the trusted party where the linkage is conducted. Once pairwise values satisfy the propagation formula, they should be assigned bi-directional associations formed as $DH_x(value_i) \circ_{ir} DH_y(value_j)$. Different from explicit mappings from domain knowledge, such implicit relations are effective for ad hoc and evolving data linkage scenarios.

Definition-8. For linkage set D constructed by linking $dataset_A$ and $dataset_B$, the association rules like $Itemset_A \rightarrow Itemset_B$ holds if the following conditions are established:

$$\frac{|R(Itemset_A)|}{|R_{linkage}|} \geq ms$$

$$\frac{|R(Itemset_B)|}{|R_{linkage}|} \geq ms$$

$$\frac{|R(Itemset_A \cup Itemset_B)|}{|R_{linkage}|} \geq ms$$

and the association rule formed as $Itemset_A \rightarrow Itemset_B$ having the confidence value satisfying:

$$\frac{|R(Itemset_A \cup Itemset_B)|}{|R(Itemset_A)|} \geq mc_B$$

Here |R(x)| returns the number of records where the variable x appears. Minimum support (*ms*) is defined by the linkage domain to filter out item sets that are not necessary to explore associations. In addition to statistical significance, the strength of associations can be evaluated using local confidence levels, such as minimum confidence required by dataset B (mc_B). For instance, a subset of attributes from two different registries is shown in the Fig. 7 where the "language spoken at home" is *2201-Greek* and "ethnicity" is *3205-Greek*. The numbers in the parenthesis refers to the co-occurrences and the respective appearances in the datasets. As defined, the dependence of *2201-Greek* to *3205-Greek* is 100% (42/42) while only 8.4% (42/500), i.e. only 8% of Greek people speak Greek at home, but of all those that do, they have an ethnicity of Greek. Given a minimum confidence of *0.8*, the inference from *2201-Greek* to *3205-Greek* is accepted for further evaluation of policies specified with the language and ethnical variables. Specially, the Rule *14* is defined to reason about obligation enforcement along with such associated variables.

14. *Obligation_P(?o), Language(?l), Ethnicity(?e), ir(?l, ?e), enforceOn(?o, ?e)* → *enforceOn(?o, ?l)*

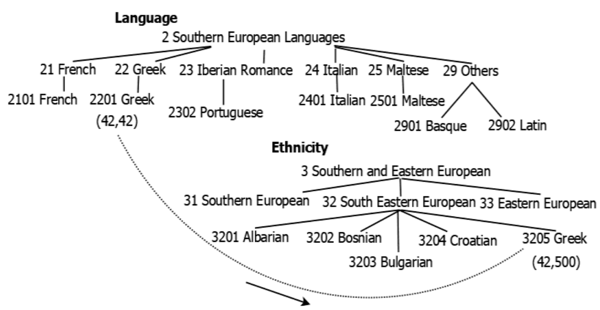


Figure 7. Mining associations cross vocabularies (*Language* → *Ethnicity*).

Any disclosure incurs a privacy cost. As such a key goal for privacy preservation is to minimize privacy loss while maintaining a given level of utility. To evaluate a privacy-awareness policy framework, certain metrics need to be defined to quantify such indicators. Instead of using 0/1 to signify whether data should be disclosed or not, we propose a refined method by taking distinctive "specialness" into consideration. For data linkage, it is necessary to measure the significance of results to local datasets, e.g. the percentage of patients falling in the linkage set is a key measure.

Definition-9. For records shared by party A and B, the Overlapping Rates *(OR)* can be computed by:

$$OR_A = \frac{|R_A \cap R_B|}{|R_A|} \left(OR_B = \frac{|R_A \cap R_B|}{|R_B|} \right)$$

where $|R_n|$ refers to the number of source records while $|R_n \cap R_m|$ counts the size of the resultant data set of $linkage_{A\text{-}B}$. On this basis, the privacy cost can be computed by:

$$PC_A = \frac{OR_A \cdot \sum_{i=1}^{n} \triangle L_i \cdot [percentage_i]}{N_A} \left(PC_B = \frac{OR_B \cdot \sum_{j=1}^{m} \triangle L_j \cdot [percentage_j]}{N_B} \right)$$

Here N_i refers to the number of local attributes; ΔL_x represents the differences between "expected specialness" and "resultant specialness" and $percentage_x$ refers to the proportion of related records in the overlapping set (linkage). It is noted that *PC* can be computed once all associations are identified and applied.

IV. Case Study-Type-1 Diabetes Analytics

A. Background

To demonstrate the benefits of this approach, we consider a linkage scenario involving two major projects currently ongoing involving the University of Melbourne. To promote and understand public health in Victoria, the Department of Health and Human Services (VicHealth – https://www.vichealth.vic.gov.au/) undertakes a survey involving 25,000+ Victorians with regards to their overall health, work-life balance, drinking and smoking habits and basic demographics. VicHealth aggregates results using standard geospatial regions, typically local government areas (LGAs) or statistical local areas (SLAs). It is noted that arbitrary aggregation using unit level data is also possible using geo-spatial privacy technologies as described in [49]. Thus, the unit-level point-based data (respondents' addresses) can be aggregated to Statistical Area levels (SA1-SA4), e.g. the people in an SA1 that live within a given distance of a park or a bottle shop [53]. Fig. 8 shows the VicHealth data aggregated at the SLA level for Greater Melbourne showing the amount of monies spent per week (in dollars) on alcohol for given SLAs. The darker colors on the choropleth map reflect an increase in alcohol spends. The actual data is shown in tabular format also (aggregated at the SLA level).

The Australian Diabetes Data Network (ADDN – www.addn.org.au) has established a national type-1 diabetes platform for Australia. This facility comprises (at present) over 13,000 patients from major diabetes centers across Australia as shown in the Fig. 9. A rich range of information on these patients is available including their demographic details, their treatments and visit information. This system includes both pediatric and adult patient data and supports the Australian Diabetes Society (ADS) and Australian Pediatric Endocrine Group (APEG).

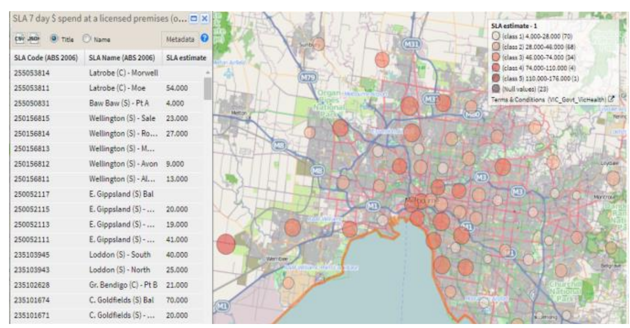


Figure 8. SLA-based alcohol spending patterns across Greater Melbourne.

ADDN Participants – Center Summary							
Center	Total	Active				Inactive	Data Last Updated
		All Active	Type1	Type2	Others		
NSW-NTL-JHC	350	282	282	0	0	68	25/07/2016
NSW-SYD-CHW	1133	1133	1038	28	67	0	05/04/2016
QLD-BNE-LCH	654	479	465	2	12	175	28/04/2016
SA-ADL-WCH	758	629	612	10	7	129	30/06/2016
VIC-MEL-MCH	1138	703	664	15	24	435	12/07/2016
VIC-MEL-RCH	1418	1122	1091	12	19	296	29/04/2016
VIC-MEL-RMH	1316	582	582	0	0	734	15/06/2016
WA-PER-PMH	1252	966	900	40	26	286	25/07/2016
All Centers	8019	5896	5634	107	155	2123	

Figure 9. Diabetes patient recruitment in ADDN.

Fig. 10 shows a fragment of the ADDN data dictionary. Prior to data exchange/linkage, repositories submit their data schema, which is abstracted from (wherever possible) standardized sources. Both the language [50] and ethnicity [51] hierarchies are standard taxonomies defined by the Australian Bureau Statistics (ABS – www.abs.gov.au). Geographic classifications such as Postcodes [52] and Statistical Area level [53] (SA1-SA4) codes have been defined by the Australian Statistics Geography Standard (ASGS). Different from categorical attributes, numeric and other variables can require *ad-hoc* transformations. For instance, patient ages can be constructed based on exact values (age = 6) or based on intervals (0 < age < 5 years). For quantification, values such as $0, \frac{1}{3}, \frac{2}{3}$, or 1 can be attached to capture the local specificity of different concept clusters where the unit level is recognized as "*1*" and "*0*" refers to empty [54]. Through this, the granularities of different clusters can be used to understand the effect when measuring "inference channels".

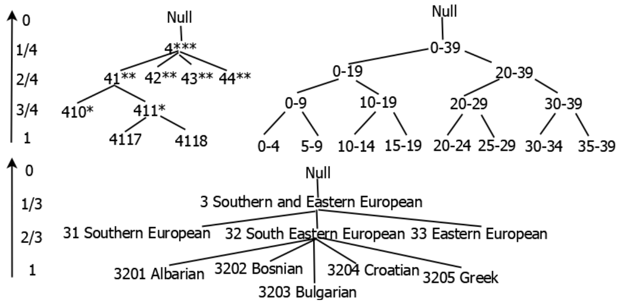


Figure 10. Quantified hierarchical variables.

Both ADDN and VicHealth can deal with health-related data with standardized information wherever possible, e.g. geospatial data. In this case study we assume that there exists a set of patients that have type-1 diabetes in ADDN that were also involved in the VicHealth survey. The individual identity of the patients should obviously not be disclosed, but importantly the danger of potentially identifying an individual should also be protected against. To demonstrate how a given policy violation can be detected from these two data rich resources, we select EHRs from ADDN (*2000*) and VicHealth (*2500*) with *1000* shared patients (respondents) existing in both registries with completely different attributes. After cleaning the incomplete records, the remaining *996* records were used as inputs to the analysis. Table I shows the attributes and sources of knowledge used. These data models and sources build upon standards and ontologies.

TABLE I. SAMPLE REGISTRIES AND THEIR ATTRIBUTE INFORMATION

Registry	Role hierarchy	Variable	Source	Instance
VicHealth	*Admin* ***Researcher***	*Age* *Language* *Statistical area*	*5-year* *ASCL* *ASGS*	*15* *36* *95*
ADDN	*Diabetologist* ***Clinician*** *Nurse*	*Postcode* *Ethnicity* *Gender*	*AU Post* *ASCCEG*	*83* *38* *4*

As shown in Fig. 11, both VicHealth and ADDN have defined policies based on geospatial distributions that require special protection when releasing non-geospatial attributes (e.g. *Age* or *Ethnicity*). For instance, both VicHealth and ADDN prevent geo-spatial leakage based on the number of patients within a given postcode, e.g. *at least 3 or 5 individuals need to be located in the same spatial level for the aggregated data to be released*. To achieve this, the SA1 codes are transformed to more aggregated SA2 level areas when there are insufficient numbers of respondents (<*3*). Similarly, postcodes in ADDN can be aggregated before allowing disclosure/linkage. In this scenario we consider a clinician (***ADDN Clinician***) requesting to access information related to patients existing in both data resources. Based upon the cross-hierarchy associations between Clinician and Researcher (***VH Researcher***), data elements in the linkage can be released once the privacy obligations are successfully completed. With the relation ***VH Researcher*** $\equiv$ ***ADDN Clinician***, data elements in the linkage can be released once the privacy obligations are successfully met. Specifically, Fig. 12 shows the composition rules for how an ADDN clinician can access the associated VicHealth data elements through linkage.

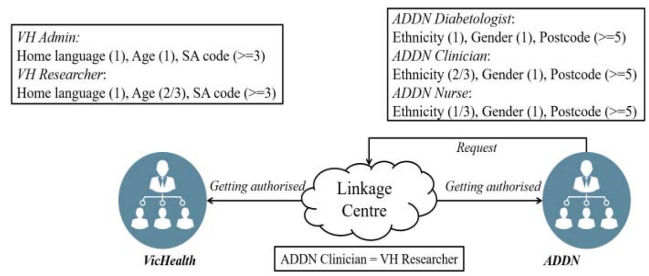


Figure 11. Access patterns in VicHealth and ADDN.

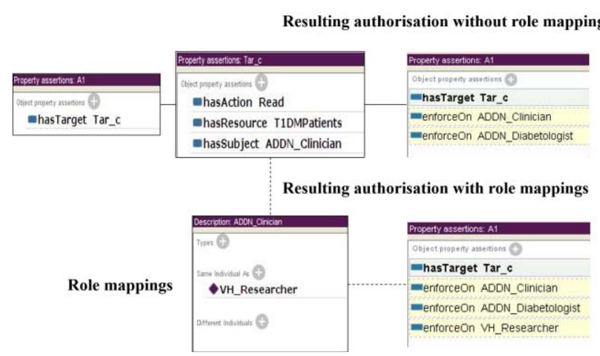


Figure 12. Composing policies by reasoning via role hierarchies

As discussed, just using authorization decisions only raises potential inference risks. For instance, it cannot guarantee the release contains at least 5 records in each postal area as defined at ADDN side. In other words, less than *5* individuals in the postal region should result in no data being released. However, disclosing the VicHealth-ADDN linkage data set shown in Fig. 13 can violate the protection intent since a smaller population can be identified from the group, e.g. 6 patients are distributed in two SA1 regions, *31103131619* and *31103131212* which belong to two postal areas, *4118* and *4117*, respectively. Based on the geo-spatial concept mappings, the protected zip code *411** will be refined, which can breach the ADDN policy. Based on the definition of *Rule 12*, the obligation enforced to generalize *4117* and *4118* as *411** should be propagated to the SA1 codes *31103131619* and *31103131212*. Consequently, SA codes should be generalized until at least 5 patients are located in one postal region. In this case, the SA3 code "*31103*" will be released in *Linkage_1* to *Linkage_6*.

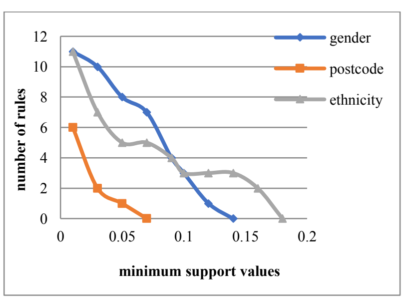
ADDN-VicHealth Linkage			
Linkage-ID	**Postcodes**	**SA codes**	...
Linkage _1	411*	31103131619	...
Linkage _2	411*	31103131619	...
Linkage _3	411*	31103131619	...
Linkage _4	411*	31103131212	...
Linkage _5	411*	31103131212	...
Linkage _6	411*	31103131212	...
...	...	...	...

Figure 13. Geospatial privacy of VicHealth-ADDN linkage.

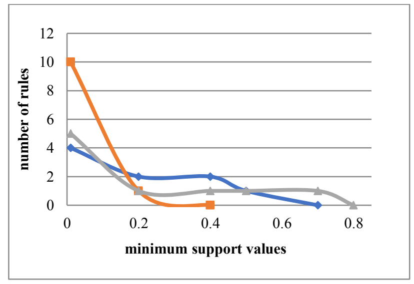
B. Result Analysis

1) Association Rule Distribution

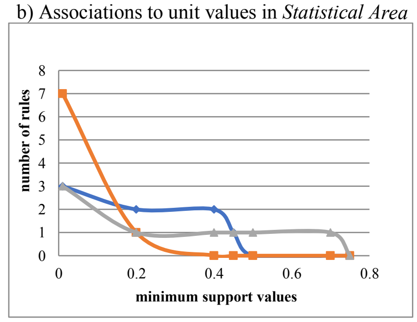
As discussed, based on associations among attributes identified via linked datasets, semantic reasoning can be implemented to support policy composition in distributed environments. As shown in Table I, each site collects patient details from three different perspectives. Given the principle requiring that only one item can be contained in the rule head/body, association rules can be evaluated in different dimensions. In this case for each attribute, three templates can be defined to construct rules. For example, taking ADDN variables as the "consequences" gives nine double-attribute templates ($C_3^1 \cdot C_3^1$). To support association rule mining, we implement a process based on *Apriori* [55] – a mining algorithm used to find frequent items from transaction datasets and association rules for business purposes. The idea involves computing the frequency of item sets and identifying those above a "minimal threshold of occurrence" as "large item sets". Instead of Boolean values, categorical attributes with related semantic meanings can also be considered. On this basis, elements of "large item sets" can be formed as association rules if they co-occur in the same records. Optionally, such rules can be filtered based on "minimal confidence" levels. In this case for linkage scenarios we implement processes distinguishing 'requestor' and 'responder'.



a) Associations to unit values in *Age*



b) Associations to unit values in *Statistical Area*



c) Associations to unit values in *Language*

Figure 14. Number of association rules on ADDN items.

As shown in Fig. 14, the rule numbers are plotted with increasing support values. As seen, all combinations exhibit a downward trend as the minimum support grows, however particularities can be found with different combinations. For instance, Fig. 14 (a) shows Postcode variables are least associated with age variables (number = *6*, minimum support = *0.01*) whereas they become the most associated variable when it comes to SA codes (number = *10*, minimum support=*0.01*) whilst Home Language Spoken (number = *7*, minimum support = *0.01*) are shown in Fig. 14 (b) and Fig. 14 (c). It is reasonable to expect associations between statistical areas and postcodes since explicit mappings exist between spatial extents. The results also highlight patients in different age intervals evenly distributed however language impacts on where to live in Victoria. Such linkage patterns indicate the association from ethnicity to statistical areas and languages (minimum support = [*0.2*, *0.7*]) in Fig. 14 (b) and Fig. 14 (c). These indicate that more than one half of the cohort are featured in such co-occurrences and thus there is an increased chance for new fact identification. Such results are returned to custodians who may update the minimal confidence requirement to balance the associated external risk and subsequent utility. In addition, data providers can define minimum association strengths. Fig. 15 shows how rules mined in different templates can be filtered by increasing minimum confidence levels. When the value is *0.4*, zero associations can be found to any age group. This is due to the even distribution of gender variables within other auxiliary knowledge, i.e. the "Gender" variables are relatively safe to disclose without privacy disclosure risk issues arising.

Based on implicit associations mined from arbitrary linkages, further access control rules can be generated through data scaling. As shown in Table II, parameters (minimum support = *0.1%*; minimum confidence = *1.0*) are set at both extremes to allow minor variations of Ethnicity and Home spoken language to be identified. Through computing the frequency and co-existence, identified associations from the linkage set may be applied to affect policy decisions in policy composition.

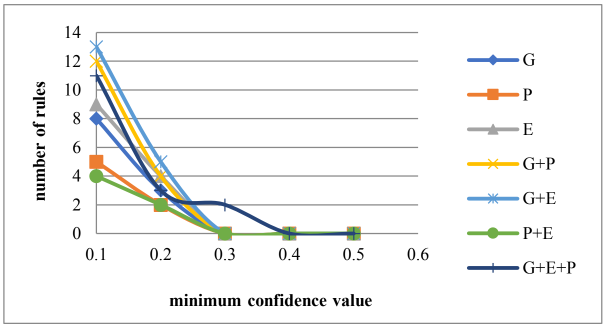


Figure 15. Numbers of association rules to *Age*.

TABLE II. ASSOCIATION ANALYSIS BETWEEN "ETHNICITY" AND "LANGUAGE"

Associations between *Ethnicity* and *Language* (conf = 1.0)		
Frequent items in "Ethnicity"		
1103-Australian South Sea Islander; 1202-Kiwi; 2301-Austria; 2306-West German; 2307-Swiss; 2405-Swedish; 2311-Belgian; 3016-Spanish; 3203-Bulgarian; 3205-Greek; 3215-Cyprian; 3307-Polish; 3308-Russian; 4106-Lebanese; 4907-Turkish; 5201-Filipino; 5214-Singaporean; 6901-Japanese; 7106-South African Indian; 7112-Pakistani; 7126-Sri Lankan; 8102-American; 9200-East African; 8204-Chilean;		
Frequent items in "Home Language"		
1201-English; 1301-German; 1401-Dutch; 1403-Afrikaans; 2101-French; 2201-Greek; 2302-Portuguese; 2303-Spanish; 3602-Polish; 4202-Arabic; 4301-Turkish; 4206-Assyrian Neo-Aramaic; 3402-Russian; 5103-Tamil; 5104-Telugu; 5202-Gujarati; 5207-Punjabi; 5211-Sinhalese; 5212-Urdu; 6511-Tagalo; 7201-Japanese; 9101-American; 9304-Maori (New Zealand);		
Number of Implicit Associations		
Language → Ethnicity	Ethnicity → Language	Bi-direction
14	3	2

2) Policy Performance Evaluation

To achieve policy compliance in linkages, we consider XACML as the fundamental framework in which policies can be defined and evaluated through a range of different models:

- *Model 1*. Policies are evaluated without structured data;
- *Model 2*. Policies are evaluated against hierarchical data structures;
- *Model 3*. Policies are evaluated against hierarchical data structures with explicit inferences;
- *Model 4*. Policies are evaluated against hierarchical data structures with both explicit and implicit inferences.

To evaluate access control policies, Paci and Zannone (2015) introduced a set of metrics regarding *effectiveness* and *efficiency* evaluation [21]. Specifically, these metrics are defined by comparing the gap between "data with expected protection" and "data with resultant protection". On this basis, we introduce an evaluation framework with adjustments applicable to dynamic data linkage applications.

Metric-1. Here the policy effectiveness refers to the completeness with which users achieve specified (protection) goals [56]. To tackle comprehensive concerns related to risk detection, protected data in access patterns need to have a "ground truth". In this case, the effectiveness can be evaluated by comparing privacy costs caused by policy models (Definition 9).

Metric-2. In relation to the effectiveness, efficiency refers to the resource utilization in relation to achieving system goals [56]. Through transforming data according to the ground truth of protection, it is possible to utilize different numbers of rules enforced based on *Models 1-4*. During this process, the more statements is required, the less efficient the model is.

Metric-3. Utility is another factor essential to consider. With regard to the databases, data utility gains are inversely related to "information loss", which can be measured through computing Sum of Square Error (SSE)/Total Sum of Square Error (SST) [57]. In this case, *n* records composed by *m* attributes were masked by replacing the original variable x_{ij} by its replacement $x_{ij}^{'}$. The SSE can be calculated by aggregating variable distances $\mathrm{d}(x_{ij}, x_{ij}^{'})$ and level (x_{ij}) where SST reflects the maximal details contained in each attribute (e.g. *0* →*1*).

$$\frac{\text{Sum of Square Error (SSE)}}{\text{Total Sum of Squares (SST)}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} d(x_{ij}, x_{ij}')^2}{\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{level}(x_{ij})^2} \quad (1)$$

TABLE III. COMPARISON OF POLICY MODELS (MODEL 1-4)

	Effectiveness (Security cost %)	**Efficiency (Rules specified for data privacy)**	**Utility Loss**
Model 1	*4.32%*	*1024*	*13.9%*
Model 2	*4.32%*	*29*	*11.4%*
Model 3	0.156%	28	
Model 4	-	-	

Table III shows the performance of policies defined using *Data Models 1-4* for linkage. All these indicators are adversely affected by the numeric results, i.e. security costs, number of statements and information loss. After calculating associations between value pairs, the ground truth can be extended by adding further knowledge. On this basis, we are able to see that the highest security cost through *Model 1* and *Model 2*, is due to a lack of semantic mapping between data hierarchies such as "*SA codes* to *Postcodes*" and "*Language* to *Ethnicity*". Therefore, the security cost is measured by calculating the PC_{SA1} and $PC_{Language}$. Through extending the correspondence information in *Model 3*, the disclosure risk of SA1 codes is addressed by semantic reasoning while the issues caused by the language use remains. In addition, through extending temporal associations and enabling semantic reasoning, the expected data profile can be realized using *Model 4*.

Based on existing knowledge and implicit associations in Table II, we compare the *resources* (privacy statements) used and their different efficiencies. Due to a lack of propagation in *Model 1*, at least *1024* statements are required to process the

SA and Language variables in records (SA1 codes in all *996* records plus language values in *28* of them need processing due to the explicit/implicit associations). Through using hierarchical structures in *Model 2*, the SA code generalization can be realized by adding one more statement such as *generalization(SA-1)* from the ADDN side (*29* statements in total). When it comes to *Model 3*, the manual operation on SA1 codes is not necessary for the knowledge extension and obligation propagation leveraging explicit mappings. Since *Model 4* includes both types of relations, no additional rule for enforcement is demanded.

Through calculating the SSE/SST and noting that a higher result implies less useful data results, we compare the utility of protected data through different policy models. In this stage, data samples are divided into two groups: data processed with/without data hierarchies based on the aggregation and suppression techniques. In the suppression case, the distance between masking and original attributes can only take binary values, i.e. *0* if they are equal and *1* otherwise. The result shows that knowledge-based aggregation maintains a higher overall utility level.

V. CONCLUSION

In this paper, we present a semantic approach to compose security policies for privacy-demanding record linkage. Through analyzing privacy issues in typical scenarios, we present a framework where inference control can be delivered through reasoning about knowledge models and associated semantic rules. We show how dynamic correlations among various attributes generated through arbitrary linkages can increase the possibility of policy violations. To tackle this, we propose it is necessary to calculate the associations between pairwise attributes by counting the occurrence and co-occurrence of data items in overlapping data sets. We show how improved performances in terms of *effectiveness*, *efficiency* and *utility* can be achieved by enriching auxiliary data models. Based on the results, we conclude that specifying policies based on structured data can minimize the loss of information while reducing the risk of privacy disclosure when semantically composing policies. When more types of associations are considered, improved security performance and minimizing risk disclosure can also be achieved. For future work, we intend to explore the reconciliation of conflicting disclosures including policy negotiation based on attribute types/values that can constitute a key step towards resolving privacy leakage in large-scale distributed systems.

ACKNOWLEDGMENT

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.

[2] Gibberd, A., Supramaniam, R., Dillon, A., Armstrong, B. K., & O'Connell, D. L. (2016). Lung cancer treatment and mortality for Aboriginal people in New South Wales, Australia: results from a population-based record linkage study and medical record audit. BMC cancer, 16(1), 1.

[3] Sandhu, R. S., Coyne, E. J., Feinstein, H. L., & Youman, C. E. (1996). Role-based access control models. Computer, (2), 38-47.

[4] Yuan, E., & Tong, J. (2005, July). Attributed based access control (ABAC) for web services. In Web Services, 2005. ICWS 2005. Proceedings. 2005 IEEE International Conference on. IEEE

[5] Ray, I., Kumar, M., & Yu, L. (2006). LRBAC: a location-aware role-based access control model. In Information Systems Security (pp. 147-161). Springer Berlin Heidelberg.

[6] Joshi, J. B., Bertino, E., Latif, U., & Ghafoor, A. (2005). A generalized temporal role-based access control model. Knowledge and Data Engineering, IEEE Transactions on, 17(1), 4-23.

[7] Luo, H., Kong, J., Zerfos, P., Lu, S., & Zhang, L. (2004). URSA: ubiquitous and robust access control for mobile ad hoc networks. IEEE/ACM Transactions on Networking (ToN), 12(6), 1049-1063.

[8] Sun, L., Wang, H., Soar, J., & Rong, C. (2012). Purpose based access control for privacy protection in e-healthcare services. Journal of Software, 7(11), 2443-2449.

[9] Rizvi, S. Z. R., & Fong, P. W. (2016, March). Interoperability of Relationship-and Role-Based Access Control. In Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy (pp. 231-242). ACM.

[10] El Kalam, A. A., & Deswarte, Y. (2006, November). Multi-OrBAC: A new access control model for distributed, heterogeneous and collaborative systems. In 8th IEEE International Symposium on Systems and Information Security.

[11] Chaudhuri, S., Kaushik, R., & Ramamurthy, R. (2011, January). Database access control and privacy: Is there a common ground?. In CIDR (pp. 96-103).

[12] Ni, Q., Bertino, E., Lobo, J., Brodie, C., Karat, C. M., Karat, J., & Trombeta, A. (2010). Privacy-aware role-based access control. ACM Transactions on Information and System Security (TISSEC), 13(3), 24.

[13] Wang, H., Sun, L., & Bertino, E. (2014). Building access control policy model for privacy preserving and testing policy conflicting problems. Journal of Computer and System Sciences, 80(8), 1493-1503.

[14] Jin, J., Ahn, G. J., Hu, H., Covington, M. J., & Zhang, X. (2011). Patient-centric authorization framework for electronic healthcare services. computers & security, 30(2), 116-127.

[15] Masoumzadeh, A., Amini, M., & Jalili, R. (2007, May). Conflict detection and resolution in context-aware authorization. In Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on (Vol. 1, pp. 505-511). IEEE.

[16] Lupu, E. C., & Sloman, M. (1999). Conflicts in policy-based distributed systems management. Software Engineering, IEEE Transactions on, 25(6), 852-869.

[17] Hu, H., Ahn, G. J., & Kulkarni, K. (2013). Discovery and resolution of anomalies in web access control policies. Dependable and Secure Computing, IEEE Transactions on, 10(6), 341-354.

[18] Lin, D., Rao, P., Ferrini, R., Bertino, E., & Lobo, J. (2013). A similarity measure for comparing XACML policies. Knowledge and Data Engineering, IEEE Transactions On, 25(9), 1946-1959.

[19] Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists (Vol. 1, No. 6).

[20] Li, Y., Cuppens-Boulahia, N., Crom, J. M., Cuppens, F., Frey, V., & Ji, X. (2015). Similarity Measure for Security Policies in Service Provider Selection. In Information Systems Security (pp. 227-242). Springer International Publishing.

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

[22] Costante, E., Vavilis, S., Etalle, S., den Hartog, J., Petkovic, M., & Zannone, N. (2013, July). Database anomalous activities detection and quantification. In Security and Cryptography (SECRYPT), 2013 International Conference on (pp. 1-6). IEEE.

[23] Castelluccia, C., Kaafar, M. A., & Tran, M. D. (2012, July). Betrayed by your ads!. In International Symposium on Privacy Enhancing Technologies Symposium (pp. 1-17). Springer Berlin Heidelberg.

[24] Accorsi, R., & Müller, G. (2013, May). Preventive inference control in data-centric business models. In Security and Privacy Workshops (SPW), 2013 IEEE (pp. 28-33). IEEE.

[25] Finin, T., Joshi, A., Kagal, L., Niu, J., Sandhu, R., Winsborough, W., & Thuraisingham, B. (2008, June). R OWL BAC: representing role based access control in OWL. In Proceedings of the 13th ACM symposium on Access control models and technologies (pp. 73-82). ACM.

[26] Cirio, L., Cruz, I. F., & Tamassia, R. (2007, November). A role and attribute based access control system using semantic web technologies. In On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops (pp. 1256-1266). Springer Berlin Heidelberg.

[27] Priebe, T., Dobmeier, W., & Kamprath, N. (2006, April). Supporting attribute-based access control with ontologies. In Availability, Reliability and Security, 2006. ARES 2006. The First International Conference on (pp. 8-pp). IEEE.

[28] Kim, J. (2013, July). Hybrid Authorization Conflict Detection by Inferring Partial Data in RDF Access Control. In Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual (pp. 89-94). IEEE.

[29] Kolovski, V., Hendler, J., & Parsia, B. (2007, May). Analyzing web access control policies. In Proceedings of the 16th international conference on World Wide Web (pp. 677-686). ACM.

[30] Liu, A. X., Chen, F., Hwang, J., & Xie, T. (2008, June). Xengine: a fast and scalable XACML policy evaluation engine. In ACM SIGMETRICS Performance Evaluation Review (Vol. 36, No. 1, pp. 265-276). ACM.

[31] 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.

[32] Evans, J. M. M., McMahon, A. D., McGilchrist, M. M., White, G., Murray, F. E., McDevitt, D. G., & MacDonald, T. M. (1995). Topical non-steroidal anti-inflammatory drugs and admission to hospital for upper gastrointestinal bleeding and perforation: a record linkage case-control study. Bmj, 311(6996), 22-26.

[33] Centre for Health Record Linkage (CHeReL). Available on http://www.cherel.org.au/.

[34] 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.

[35] Ritchie, F. and Elliot, M. (2015) Principles- versus rules-based output statistical disclosure control in remote access environments. IASSIST Quarterly, 2015 (Summer). pp. 5-13. ISSN 0739-1137 Available on http://eprints.uwe.ac.uk/28489

[36] Data Linkage. Western Australia. Available on http://www.datalinkage-wa.org.au/.

[37] SA-NT DataLink. Available on https://www.santdatalink.org.au/

[38] Victorian Data Linkages (VDL). Available on https://www2.health.vic.gov.au/about/reporting-planning-data/victorian-data-linkages

[39] Queensland Data Linkage Framework. Available on http://www.health.qld.gov.au/hsu/

[40] Van, H. D. S., Dang, T. A., & Dang, T. K. (2014). Supporting Authorization Reasoning Based on Role and Resource Hierarchies in an Ontology-Enriched XACML Model. International Journal of Computer and Communication Engineering, 3(3), 155.

[41] Lu, Y. and Sinnott, R. O. "Semantic Security for E-Health: A Case Study in Enhanced Access Control," IEEE 12th Intl Conf on Autonomic and Trusted Computing, Beijing, China, 2015, pp. 407-414.

[42] Anderson, A. Hierarchical resource profile of XACML v2. 0. OASIS Standard, February 2005. Available on https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

[43] Brickley, D., & Miller, L. (2000). The Friend of a Friend (FOAF) project.

[44] Isaac, A., & Summers, E. (2009). SKOS Simple Knowledge Organization System. Primer, World Wide Web Consortium (W3C).

[45] Halpin, Harry, et al. "Representing vCard objects in RDF." W3C Member Submission 20 (2010): 3499-3508.

[46] Krummenacher, R., Norton, B., & Marte, A. (2010, September). Towards linked open services and processes. In Future internet symposium (pp. 68-77). Springer, Berlin, Heidelberg.

[47] Dublin Core Metadata Initiative. (2012). Dublin core metadata element set, version 1.1.

[48] Diabete UK. List of countries by incidence of Type 1 diabetes ages 0 to 14. Diabetes UK. Last updated on January 2nd, 2013. Available on https://www.diabetes.org.uk/About_us/News_Landing_Page/UK-has-worlds-5th-highest-rate-of-Type-1-diabetes-in-children/List-of-countries-by-incidence-of-Type-1-diabetes-ages-0-to-14

[49] Sinnott, R. O., Bayliss, C., Bromage, A., Galang, G., Gong, Y., Greenwood, P., ... & Pursultani, H. (2016). Privacy Preserving Geo-Linkage in the Big Urban Data Era. Journal of Grid Computing, 14(4), 603-618.

[50] Australian Bureau Of Statistics. Australian Standard Classification of Languages (ASCL). 2016 Version.

[51] Australian Bureau Of Statistics. Australian Standard Classification of Cultural and Ethnic Groups (ASCCEG). 2016 Version.

[52] Find a postcode. Australia Post. Available on http://auspost.com.au/postcode

[53] ABS. Australian Statistical Geography Standard (ASGS). 2016 Version.

[54] Al-Mubaid, H., & Nguyen, H. A. (2009). Measuring semantic similarity between biomedical concepts within multiple ontologies. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 39(4), 389-398.

[55] Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB conference, pp 487–499.

[56] ISO/IEC 25010:2011: System and Software Engineering – Systems and Software Quality. Requirements and Evaluation (SQuaRE) – System and Software Quality Models (2011)

[57] Lixia, W., & Jianmin, H. (2009, August). Utility evaluation of k-anonymous data by microaggregation. In Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on (Vol. 4, pp. 381-384). IEEE.