Propagation Control Services for WebID Analytics on the Decentralized Social Web

Yuh-Jong Hu

Abstract A WebID is a single sign-on token for a user's authentication at multiple servers. In this chapter, we allow boundless WebIDs to be collected, shared, and integrated for analytics on the decentralized Social Web. The primary stakeholders in WebID analytics are the data owner, data controller, and data user. All three types of stakeholders are sufficiently aware of propagation control services so that WebIDs have best protection and usage. Types of semantics-enabled policy are proposed and enforced by data controllers to enable access control, data handling, and data releasing actions on the WebID datasets. The policy enforcement should be accountable and transparent at the data controllers to provide WebID propagation control services. Each data controller enforces a data handling policy to anonymize massive WebIDs. Moreover, the super data controller enforces access control and data releasing policies to ensure that the data owners receive the privacy-preserving WebID analytics services. Finally, we point out how to resolve WebID protection and utility conflict through different types of semantics-enabled policy to call for WebID propagation control services at the data controllers of an information value chain.

1 Introduction

Personal data can be considered a new asset class that provides valuable insights when placed under effective analytics and interpretation [37]. Big data analytics has become one of the emerging research issues in the computer science field and other related fields, such as statistical analysis and data-driven decision making [29]. We

Yuh-Jong Hu
Emerging Network Technology (ENT) Lab
Department of Computer Science
National Chengchi University, Taipei, Taiwan
e-mail: hu@cs.nccu.edu.tw

face several research challenges when providing socially aware data analytics in online social networks (OSNs). First, the data volume on an OSN is so large and its velocity moves so fast that it exceeds the processing capacity of conventional data management systems. Second, the data come from heterogeneous sources in a variety of data formats and semantics, so it is extremely difficult to provide effective data integration. Third, current centralized OSNs are all *walled gardens*, which makes seamless data integration almost impossible.

Most of the current big data analytics studies are mainly dealing with the three v's challenges to resolve the data volume, velocity, and variety problems [40]. A special report, titled "Data, Data, Data Everywhere", explores the problem of vast information collection with the complex issues of data archiving, accessing, managing, and securing [16]. This report suggests that we should consider using metadata, or data about data, for effective machine processing to glean implicit values through data analytics.

We also need new rules to regulate the big data analytics processes and furthermore to ensure the compliance of privacy protection principles. We therefore considered the emerging research issue of big data protection [42] in order to reveal the complete landscape of privacy-preserving data analytics on social networks. Otherwise, we might face a new barrier when applying integral data analytics services across legal domains of data sources.

An inter-disciplinary study of big data privacy was recently presented at the workshop of CSAIL, MIT[1]. In this workshop, academia and industries pointed out their concerns about the lack of privacy services for big data. In fact, several well-known cryptography and statistical techniques, e.g., differential privacy [14] and fully homomorphic encryption [19], have been proposed to enable output perturbation and data encryption while providing private data management services for analytics in the open outsourcing cloud computing environment [18].

The risk of re-identifying personally identifiable information (PII) across multiple data sources was addressed in [41]. In addition, the original concepts on k-anonymity and its enforcement through generalization and suppression techniques were actually defined in [35]. We must ensure that the quasi-identifier has at least k-anonymous PII in a dataset to avoid re-identification risk. Therefore, we mask PII attributes in a quasi-identifier to de-identify each PII before disclosure to a data analyst. However, k-anonymity did suffer from a privacy protection insufficiency problem against a PII cross-linkage attack when we had an unknown number of available external data sources.

The studies on differential privacy aim to achieve the ambition of bringing theoretical soundness of cryptography for statistical disclosure control (SDC) with query outputs perturbation by noise [15]. In fact, the research in differential privacy seems to be more focused on controlling the re-identifiable risk of data than on providing analytics utility [12]. Conflict always exists between data protection and usage utility. How to balance these two objectives is an eminent challenge for the big data research community [22].

[1] Big Data Privacy Workshop: Advancing the State of Art in Technology and Practice.

In this study, we consider using socially aware anonymized WebID datasets for analytics. WebID-TLS, known as the Friend-of-a-Friend (FOAF) + Transport Layer Security (TLS) protocol, uses client-side certificates of WebIDs for a Web user's authentication. A Web server requests an X.509 certificate from a Web user over the TLS to enable secure data communication and service access authentication [24]. A WebID [36], including a Web user's Profile with its certificate, and the social relationship information, are described as the RDF(S)-based FOAF ontology. The WebID Profile attributes of PII and a quasi-identifier must be anonymized before disclosure to prevent a data owner's privacy from violation. Similarly, the data owner's social relationships are also anonymized to preserve the owner's privacy.

The SDC methods were classified as conceptual, query restriction, data perturbation, and output perturbation [1]. In this study, three types of semantics-enabled policy are proposed and enforced to enable access control, data handling, and data releasing actions for appropriate propagation control services. These actions correspond to the original SDC methods for query restriction, data manipulation and perturbation, and output perturbation for microdata protection [11].

The concepts of appropriate propagation control services are described as RDF(S)-based ontologies, and are enforced as SPARQL. In fact, we leverage the power of Semantic Web techniques, including RDF(S), FOAF, and SPARQL, and apply three types of semantics-enabled policy enforcement to call for appropriate WebID propagation control services at the data controllers. For more details, please see Section 4.2.

1.1 Research Issues and Contributions

Main research goals. In this study, we argue why we should consider applying propagation control services for WebID analytics on the decentralized Social Web. WebIDs will be collected and propagated at each data controller and will be available later at the super data controller for big data analytics. We must ensure that each data owner's privacy rights are well-respected and free from any usage violations. Moreover, we must ensure transparent and accountable propagation control services at the data controllers and super data controllers along with the entire WebID provenance propagation path.

The WebID secure management services for access control, dissemination, and disclosure are enacted as parts of the WebID propagation control services. For example, an access control policy calls for query restriction services, and a data handling policy calls for data manipulation and anonymizing services. Finally, a data releasing policy calls for output perturbation services. More specifically, this paper addresses the following major *research issues*:

1. How do we restructure the current centralized online social network architecture into the decentralized Social Web to provide wide-scale WebID capturing, recording, anonymizing, sharing, integration, modeling, and analytics services?

2. How do we provide *transparent* and *accountable* WebID propagation control services at the data controllers to assure WebID protection for the data owner and usage utility for the data user?
3. How do we provide WebID protection and usage utility through types of semantics-enabled policy enforcement to call for WebID propagation control services at the data controllers of an information value chain?

Our contributions. Our main contributions are (i) restructuring the centralized online social network architecture into the decentralized Social Web for wide-scale WebID collection and analytics, (ii) demonstrating how to provide transparent and accountable propagation control services at the data controllers to assure WebID protection for the data owner and usage utility for the data user, and (iii) modeling how to provide WebID protection and utility through types of semantics-enabled policy enforcement to call for WebID propagation control services at the data controllers of an information value chain.

Outline. This paper is organized as follows. In Section 1, we give an introduction. Then, we provide background information in Section 2. In Section 3, we explain why we restructured the centralized online social networks into the decentralized Social Web. In Section 4, we present the concepts of transparent and accountable propagation control services for WebID sharing, integration, and protection. In Section 4.2, we also point out the reasons for choosing RDF(S)-based ontologies and SPARQL queries to enable propagation control services. In Section 5, we present three types of semantics-enabled policies that call for WebID propagation control services on the privacy-aware Social Web. In addition, we explain how the big volume of WebID hybrid analytics services can be implemented in the RHadoop platform. In Section 6, we address related work. Finally, we conclude this paper with possible future work in Section 7.

2 Background

We first exploited the centralized social network's architecture and restructured it into the decentralized Social Web to provide wide-scale data sharing. The research issues of privacy in social networks are not the same as the research issues in data protection in the relational database management system [49]. Given a complete information value chain, we intend to apply types of semantics-enabled policy for information propagation and control to assure the information quality and privacy protection criteria.

We allow the big data analytics process to be operated in the entire information value chain, and the semantics-enabled policies are enacted transparently and accountably at the data controller, which ensures that each data owner's privacy concerns is respected and each data user's usage utility is preserved.

Any available data manipulation techniques, such as sanitation, obfuscation, and anonymity, are applied to the WebID datasets to de-identify the PII, quasi-identifier, and sensitive social relationships. Moreover, we also allow upstream data owners

and downstream data users using data provenance techniques [30] to trace and examine the data protection and usage criteria at each data controller checkpoint along with the WebID propagation path of the information value chain. The final goal of privacy-preserving WebID analytics is to empower the balancing between WebID protection and usage utility on the decentralized Social Web.

Based on [28], we propose a six-stage WebID analytics process: (1) acquisition and recording; (2) profile attributes and social relationships extraction with semantic annotation for anonymizing; (3) integration, aggregation, and representation; (4) modeling and analysis; (5) query processing and disclosure for analytics; and (6) interpretation (see Figure 1).

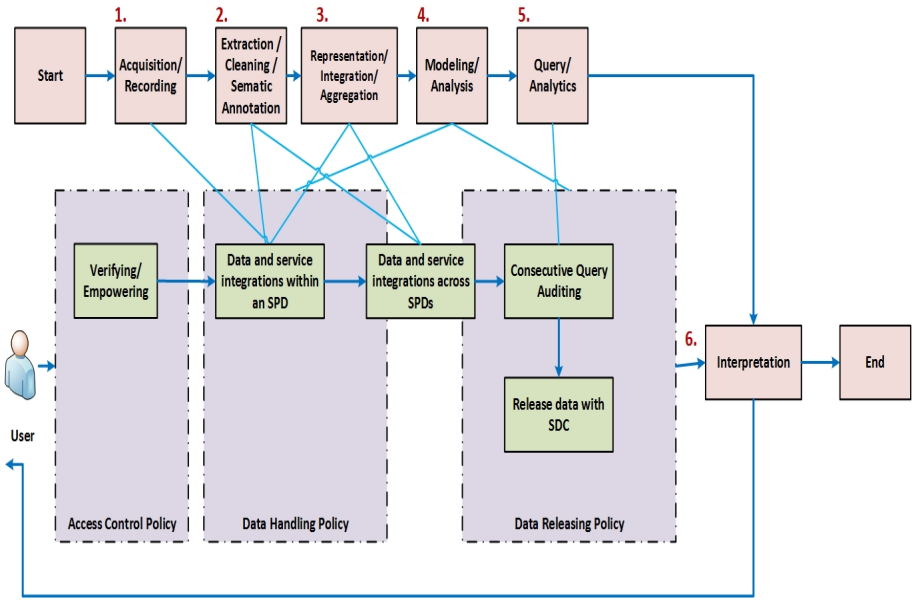

Fig. 1 The WebID analytics process is shown as a six-stage lifecycle, where three types of semantics-enabled policy are defined and enforced to call for WebID propagation control services and to achieve access control, data handling, and data releasing objectives.

Different stakeholders, including data owners, data controllers, and data analysts, are involved in the WebID analytics process. They are all aware of the status of WebID propagation control service execution. The WebID risk-utility problem for disclosure can be investigated through the *transparent* and *accountable* execution of propagation control services from a limited number of data controllers to numerous data owners and data users.

3 From a Centralized to Decentralized Social Web

A centralized social network is a data silo of walled gardens, where a data controller is responsible for his/her own data protection and analytics services. On the one hand, a data owner does not have full control over his/her own PII or social relationship information collection and disclosure. On the other hand, a data analyst can directly proceed PII analytics without the data owners' explicit awareness and consent. We envision a new decentralized Social Web architecture emerging, where a data owner, with a single sign-on access token, can flexibly select one of many well-known and trusted data controllers to manage his/her PII and sensitive relationship information for numerous Social Web sites [3].

In addition, each data owner's digital footprint contexts are recorded, interconnected, aggregated, and analyzed only for the good of the data owner. For example, ontology-based context fusion techniques can be applied to multiple social network platforms and the algorithms can find out the hidden relationships of contexts for mobile users' recommendation services [25]. However, the entire big data propagation and disclosure control process for analytics still must be fully compliant with the privacy protection principles. On the decentralized Social Web, we also have incentives to invite other major stakeholders, e.g., data controllers and data users, to participate, because they can obtain the value of wide-scale interconnected datasets for sharing, aggregation, and analysis.

3.1 The Decentralized Social Web

The World Wide Web Without Walls (W5) ecosystem concept was proposed to resolve the data protection and interoperable problems of current centralized OSNs[27]. The W5 breaks the data silo of a walled garden situation with *aggregates*. An aggregate is similar to a data controller concept, because it uses a single virtual logical machine to host a large collection of services from developers and commingled data from many Web users. In W5, numerous aggregates form the decentralized Social Web.

We first restructure a centralized online social network into a decentralized Social Web, where a super-peer domain (SPD) is circumvented with an independent logical boundary of the WebID repository framework to enforce the appropriate propagation control services at the (super) data controllers. In reality, the SPD can be declared as a policy-aware legal domain in the outsourcing of a socially aware data cloud storage environment [7].

In this decentralized Social Web architecture, each Web user can be authenticated at a data controller by using a single sign-on WebID token before a service is requested. Each data owner can flexibly select one of the trusted data controllers as his/her WebID guardian, and the WebIDs are easily interconnected and aggregated for analytics.

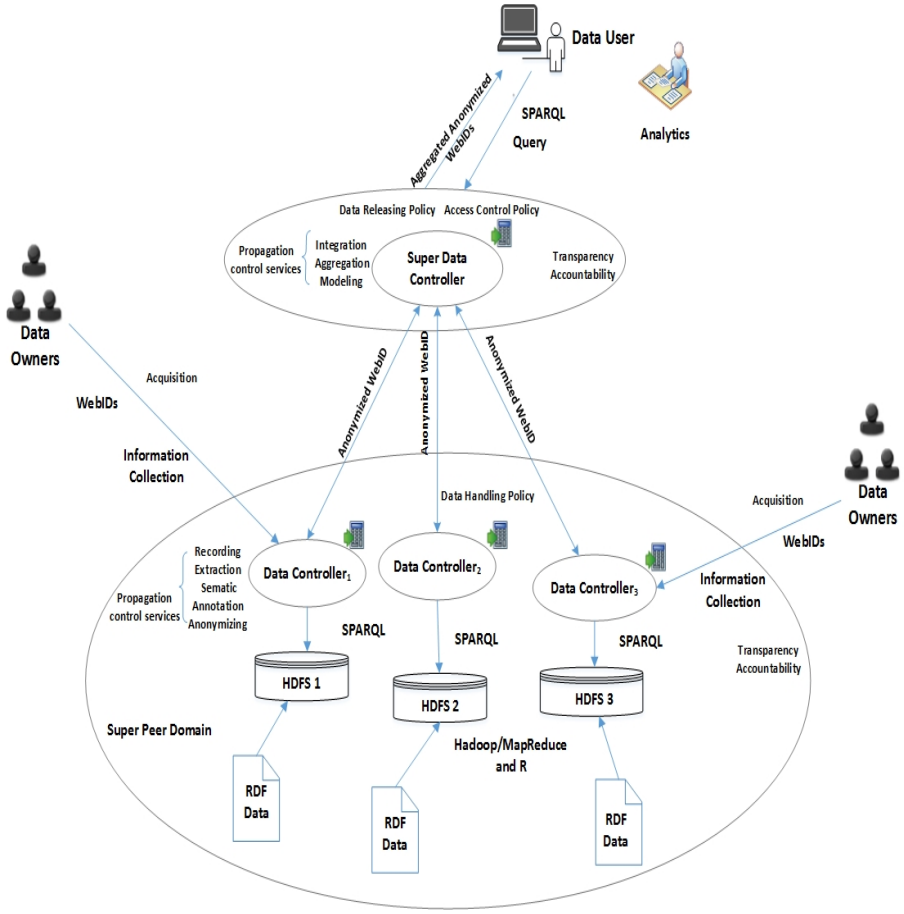

Fig. 2 The super-peer domain (SPD) data cloud for the decentralized Social Web, where a data owner hand-picks a trusted data controller to record and mask his/her WebID in an anonymized WebID dataset. Later, a data controller forwards this anonymized dataset to the super data controller to enforce the WebID access control and disclosure operations.

A decentralized Social Web user is fully in charge of his/her own WebID sharing and dissemination, because a data owner is endowed with self-control over the privacy protection policy configuration. The data owner entrusts a data controller to manipulate his/her WebID, which implies that the semantics-enabled policy that was established and enforced at the data controllers should be transparent and accountable for any stakeholder to examine and verify its trustworthiness. In contrast, a Web user lacks these features in the centralized OSNs.

3.2 WebID Linked Data

A WebID is shown as a FOAF ontology, and it uniquely describes a personal profile with the user's digital certificates and social relationship information [24]. A WebID is also portable and linkable, because it applies the URI *Webizing* technique to refer to a user's identity on the decentralized Social Web.

The incentives to use WebID linked data are as follows. First, RDF(S) linked data can provide the seamlessly wide-scale WebID integration from heterogeneous data sources without being hindered by the inconsistency of their schema [20]. RDF(S) is a graph-based ontology language for representing the FOAF ontology of WebID linked data [47]. SPARQL is a query language, and it can be used for WebID linked data access control, dissemination, aggregation, and disclosure.

Second, WebID linked data have various interchange languages for their interoperability. Turtle is a primary RDF(S) concrete syntax for WebID representation [6]. JSON is an interchange language for data serialization and de-serialization of Graph API outputs captured from the NoSQL data-stores [44]. Another emerging JSON-LD is an interchange language to become an intermediate format between JSON and RDF. JSON-LD uses `@context` to describe JSON data schema and vocabulary sources; `@type` to describe the data type of a vocabulary; and `@id` to represent a vocabulary as an identifier [39]. Therefore, once JSON objects become JSON-LD objects, they are interoperable and reusable through these additional vocabularies.

Third, each Web user's WebID is collected, linked, and anonymized at a data controller. Later on, these anonymized WebID datasets are disseminated and aggregated at the super data controller for integral analytics. This process simplifies the wide-scale WebID query processing and protection on the decentralized Social Web.

Finally, each stage of the WebID analytics process is enacted by one of the responsible actors, a data controller or the super data controller, to empower the appropriate propagation control services within an SPD. The entire WebID analytics lifecycle is activated through various types of semantics-enabled policy enforcement to call for WebID propagation control services and to achieve restricted query, manipulation, anonymizing, dissemination, and disclosure of WebID datasets. Moreover, WebID dataset sharing and propagation control across SPDs is still possible. In this case, the semantics-enabled policies established at the different super data controllers are unified to enable integral WebID analytics for multiple SPDs.

4 Transparent and Accountable Propagation Control

We should build a highly transparent and accountable policy enforcement platform for WebID propagation control. This platform has been established in the distributed Apache open source Hadoop ecosystems, such as Cloudera's or Hortonworks Sandbox's distribution of Hadoop, to provide and simulate socially aware WebID management services. Data owners and data analysts can assure that the WebID management services, including acquisition, recording, anonymizing, sharing, integration,

dissemination, disclosure, and analytics, are all following the privacy-preserving WebID analytics principles. A data owner first configures the appropriate WebID protection rules at a pre-selected, trusted data controller for its retention, dissemination, disclosure, and usage. Then, a data owner can track or be notified about his/her WebID dissemination, disclosure, and usage through a privacy-aware notification system.

In [32], the author proposes a " framework of contextual integrity" for privacy protection techniques, and describes the underlying philosophy of contextual integrity: "a right to privacy is neither a right to secrecy nor a right to control but a right to appropriate flow of personal information."

In the present study, we enact appropriate WebID propagation control services at the (super) data controllers. A data controller is accountable and responsible for providing transparent WebID management services to the WebID owners and users.

In [46], the authors propose new strategies to enforce privacy protection policies and to model transparent, accountable data mining. Similarly, in the present study *transparency* indicates that the history of WebID management control inferences is maintained and can be examined by related stakeholders, such as WebID owners and users. Furthermore, *accountability* implies that WebID owners and users can check whether the propagation control services that govern WebID management control inferences do in fact adhere to the types of previously declared privacy protection principles at the data controllers.

On the one hand, a data owner can trace WebID provenance for analytics, and can negotiate with a data controller for maximum usage utility. On the other hand, a data user (or analyst) becomes aware of the potential WebID dataset utility once the requesting sanitized datasets are available from the (super) data controllers.

The super data controller selects the SDC methods and parameter values based on the feedback information from a data analyst, and forwards these methods and parameters to various data controllers to ensure a balance between privacy protection and WebID utility. Moreover, the super data controller could postulate any possible WebID disclosure threats, especially in the consecutive query scenario, by using auditing logs to see whether his/her disclosure protection strategies proved to be sufficient and effective.

Transparent and accountable features are also applied to propagation path control for WebID provenance. Similar to the open data provenance model (OPM) [30], the propagation path control model should provide the metadata of WebID sources, destinations, and propagation paths created by the data controllers so that the authorized data users and data owners may examine and verify them. We attempt to provide an origin source of WebIDs and allow WebID propagation paths to be documented as semantic metadata in a coherent manner to assure the trustworthiness of each WebID source and propagation path control on the information value chain.

4.1 Appropriate WebID Propagation Control

Online social media is one of the major data sources for analytics. However, the metadata that describe data provenance and dissemination are not yet available in the current OSNs [5]. This situation limits the tracing and sharing of data sources across social media sites for effective integral data analytics. To address the issues presented in Section 3, we need to restructure social networks from their centralized architecture into a decentralized Social Web. Moreover, we should provide appropriate WebID propagation control to prevent cross-site data sharing and integration from unexpected private data disclosure.

In [32], the privacy rights of each individual data owner not only were confined to a data owner's preferences and control but also were expanded into the *appropriate flow* of personal information. Here, the *appropriateness* is associated with a *context* and *information norm*, so an "appropriate flow" is defined as information flow in accordance with information norms. The key concepts of information norms are context, norm, actor, information type, and transmission principle. Contexts are recognized as "abstract representations of social structures experienced in daily life," while "information norms" are usually embedded in the data protection law.

How do we enact unambiguous information norms that are software executable without too much human intervention? The answer is using software defined semantics-enabled policy for appropriate WebID propagation control. The purpose is to ensure that the actions of WebID manipulation for acquisition, recording, anonymizing, sharing, integration, aggregation, and analytics all satisfy the information norms. These actions are automatically triggered when the conditions of data manipulation satisfy the rules specified for the information norms.

How do we accomplish appropriate propagation control services that satisfy the information norms at various stages of the big data analytics processes? In [38], Solove proposes four kinds of operation activities that might result in harmful effects on privacy: (1) information collection, (2) information processing, (3) information dissemination, and (4) invasion. Here, the information norms, represented as RDF(S) ontologies and enforced as SPARQL queries, are established at the (super) data controllers.

4.2 RDF(S) and SPARQL for Information Norm

The well-known Semantic Web layered architecture[2] has undergone several revisions and has evolved into a static state. The semantics-enabled policy is a software defined specification of an information norm based on the Semantic Web technologies. When policies are formulated and treated as knowledge bases, including ontologies and rules [17], many operations can be automated, thereby reducing ad-hoc program coding to a minimum and enabling automated documentation. Moreover,

[2] `http://www.w3.org/2007/03/layerCake.svg`

the context of policy itself is described in a machine understandable way. A policy's explicit representation in terms of ontologies or rules depends on what the underlying logic foundation of the policy language is.

Policy languages vary considerably, ranging from Description Logic (DL)-based policy language, such as KAoS and Rei to the Logic Program (LP)-based policy language, such as EPAL [21]. This leads to different stances w.r.t unique name assumption (UNA) and the closed world assumption (CWA) [33]. If policies are created from DL-based policy language, then they are shown as $\mathscr{T}$ $\mathscr{B}$ox ontology schema and $\mathscr{A}$ $\mathscr{B}$ox instances. Otherwise, when policies are created from LP-based policy language, they are a set of rules with unary and binary predicate variables and facts. A Datalog rule is a restricted LP, which enables a machine to process security and privacy protection verification for information norms [9].

However, in this study, the concept of information norm is described as the RDF(S) ontology language, and its execution is enacted by the SPARQL query language. First, unifying DL with LP within the first-order logic (FOL) is a considerable challenge. DL uses an open world assumption (OWA) with no UNA, whereas LP uses a CWA with UNA.

Second, although DL-based OWL has more expressive power than graph-based RDF(S), RDF(S) graphs are gaining wide popularity, driven by efforts such as the Linked Data Initiative. In fact, RDF(S) datasets are becoming available in several ways, making the current amount of available RDF(S) data substantial. We have WebID ontologies that are only described as RDF(S)-based FOAF with an additional description of abstract privacy protection concepts for the decentralized Social Web. More details about the incentives to use WebID linked data appear in Section 3.2.

Third, the Datalog rules can be represented as SPARQL query language, because SPARQL aligns with Datalog rules in several ways [8]: the rule language is compatible with the use of SPARQL as a language for querying RDF datasets; SPARQL queries can be represented as Datalog rules; and SPARQL's CONSTRUCT queries can be viewed as deductive rules that create new RDF triples from RDF datasets; .

Propagation control services are executable processes for WebID manipulation actions triggered by a SPARQL query at the (super) data controllers. In the Event-Condition-Action (ECA) rules, e.g., `On Event If Conditions Do Actions`, an incoming request event causes a rule's conditions to be verified, and the rule further derives a conclusion to trigger actions to call for various WebID propagation control services, such as data manipulation and output perturbation, which affects the WebID datasets and the final outputs. The SPARQL query with its host language can provide an equivalent expression power of the ECA rule on the corresponding event, conditions, actions, and effects operations.

5 Privacy-Aware Propagation Control

We can acquire a WebID data owner's privacy preferences at a data controller. A data controller should guarantee that the privacy rights of a WebID owner will be

respected once the privacy preferences are collected. Instead of using natural language to declare a WebID owner's privacy statements, previous P3P was used for the machine-readable privacy preference statement declaration in a Web server to present the privacy policy of an information norm. However, the P3P privacy statements offer the data owner only a yes or no option without any negotiation. Moreover, a Web server might use another enterprise privacy authorization language, EPAL or XACML to specify an enterprise server's internal data access control authorization policy [2] [26]. These issues will be our challenges when we offer a privacy-preserving WebID data analytics service on the decentralized Social Web.

Because a Web user's privacy preferences are hard to describe and capture with the P3P language, we considered an alternative approach, e.g., accountable WebID propagation control services, to assure that the machine-executable data protection and usage utility criteria satisfy the information norms. A data owner hand-picks a trusted data controller who is responsible for his/her WebID manipulation. The WebID propagation control services are machine-executable processes that can be called for by the semantics-enabled policies established at a (super) data controller.

5.1 Semantics-enabled Policy

In [46], the authors use discretionary rule-based policy-aware techniques for a Web server's resource access verification. In this policy-aware Web, a user is authenticated by the rule-based access control policy without having to register with the Website. Similarly, we need a policy-aware Social Web to support privacy protection while providing data analytics services to the data analysts. The research objective of privacy protection data disclosure for analysts is different from the other research objectives of privacy-aware access control systems for regular subject-based query [4] [10]. On the one hand, a Web user can act as a data owner who requests Web services from an online social networking site. On the other hand, another Web user can act as a data analyst who discovers new insights into massive datasets through effective data analytics.

The primary objective of this study is to design a policy-aware Social Web architecture that incorporates semantics-enabled policy that is represented as a combination of ontologies and queries.

5.2 Call for WebID Propagation Control Services

Semantics-enabled policy enforcement that calls for WebID propagation control services should be accountable, which means any stakeholders should be able to examine and verify the semantics-enabled policy enforcement at data controllers to assure that the propagation control services are trusted and compliant with the WebID in-

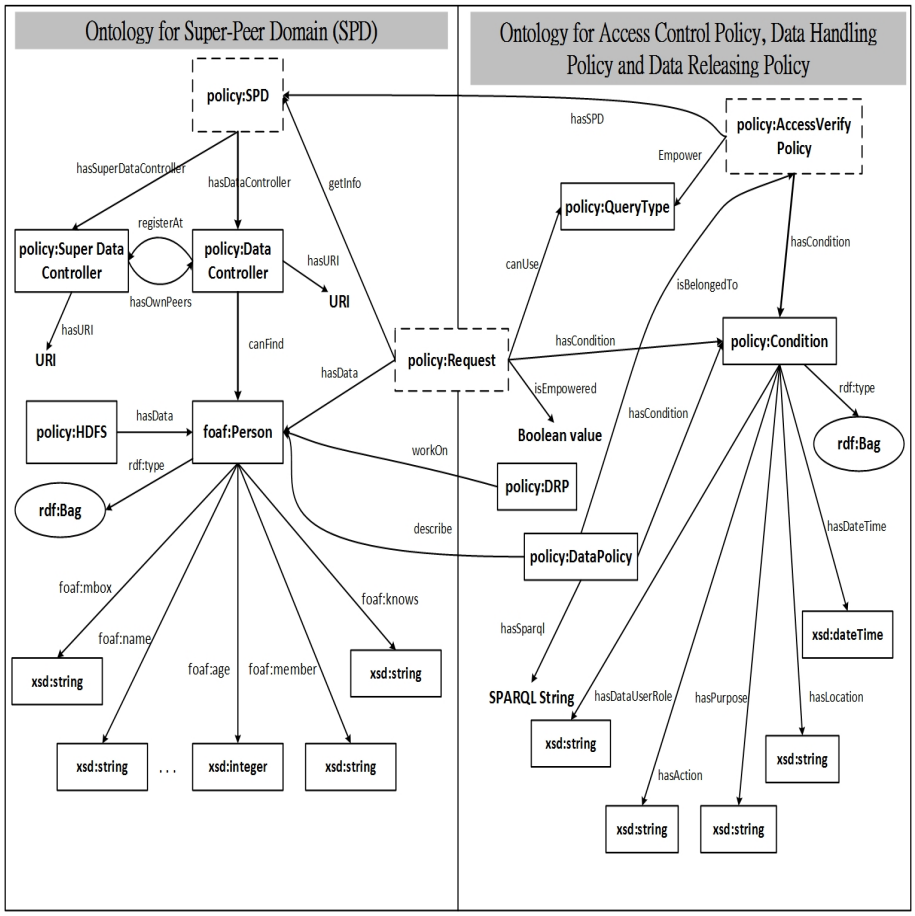

Fig. 3 The semantics of a super-peer data cloud are described as the policy ontology, which includes two modular concepts: (**a**) SPD and (**b**) three types of data policy for access control, data handling, and data releasing operations.

formation norm's context, including actor role, access conditions, usage purpose, and WebID de-identifiable criteria.

However, the types of semantics-enabled policy are originally proposed and owned by a data controller, so assuring the accountability and trust-worthiness of policy enforcement for an outsider, e.g., a data owner or a data user, is a considerable challenge. One possible solution is disclosing a policy's auditing logs under certain conditions upon request for its related stakeholders to examine and verify its accountability. Nevertheless, we still do not know under what conditions to provide auditing log disclosure or how to provide auditing log disclosure, which can follow the need-to-know principle to disseminate to the stakeholders without violating the privacy protection principles of the data owners. We need a comprehensive solution

to deal with accountable propagation control to satisfy the information norms. This problem still needs further study.

We propose an ontology to describe a super-peer domain (SPD), including various data controllers and the super data controller, for WebID propagation control services (see the left-side of Figure 3). Three types of semantics-enabled policies, shown as a combination of ontologies and queries, call for propagation control services to enable access control, data handling, and data disclosure operations. A WebID is described as a FOAF ontology and the WebID access control, manipulation, and disclosure are described as other ontologies. Then the actions of WebID propagation control services are triggered by the SPARQL queries (see the right-side of Figure 3). More details are given as follows:

1. Access Control Policy (ACP)

   An ACP is used for data request verifications. The super data controller uses an ACP to decide whether a data request from a data analyst is permitted. The concept is represented as an ACP ontology, and is enforced as a SPARQL query. Let's image that a data analyst named Peter submits a data request in `PeterRequest.rdf` shown as the following set of triples:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix policy: <http://nccu.edu.tw/policy>.

policy:QueryType rdf:type rdf:Class
policy:PBQ rdf:type policy:QueryType
policy:Condition [
policy:hasDataUserRole "DataAnalyst";
policy:hasPurpose "Analytics";
policy:hasAction "Read";
policy:hasLocation "Taipei";
policy:hasDateTime "2013:12:25:15:00" ].
```

   The super data controller initiates a *restricted query service* after accepting the following SPARQL's `Ask` Boolean query, and then proceed with a return value of yes (or no) to indicate whether the data query from Peter is permitted or (not permitted). If the answer is yes, then more anonymized datasets will be available for later disclosure. Otherwise, the reasons for rejection will be sent to the related stakeholders.

```
Ask ?permit
From <PeterRequest.rdf>
Where {?r policy:isEmpowered ?permit.
?r [ ?qt rdf:type policy:QueryType;
policy:hasCondition ?c [
policy:hasDataUserRole ?role;
policy:hasPurpose ?purpose;
```

```
policy:hasAction ?action;
policy:hasLocation ?location;
policy:hasDateTime ?time ] ].}
```

2. Data Handling Policy (DHP)

   The anonymizing principle for data handling is to decide which attribute is protected by which SDC technique [23]. The categories of an attribute combination are (1) an identifier attribute that is completely de-identified; (2) quasi-identifier attributes that are selectively revealed with the applicable SDC methods for categorical or continuous attribute types; and (3) confidential attributes that are disclosed only as they are coupled with the de-anonymized (quasi-)identifiable attributes that satisfy at least the k-anonymity criteria, etc.
   A data controller uses a DHP's ontology to describe which SDC techniques can be applied to anonymize a WebID Profile and sensitive social relationship variables. This method provides selective data revelation from its own anonymized WebID dataset from a data controller to the super data controller.
   After a data controller initiates an *anonymizing service*, anonymized WebID profile attributes with one hop of friendships are represented as a RDF(S)-based Turtle file shown as:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix policy: <policy>.

<http://nccu.edu.tw/j/foaf.rdf> a
foaf:PersonalProfileDocument.
<http://nccu.edu.tw/j/foaf.rdf> foaf:maker :me.
<http://nccu.edu.tw/j/foaf.rdf> foaf:primaryTopic :me.
:me a foaf:Person.
/*De-identification*/
:me [ foaf:name "Yuh-Jong Hu";
foaf:homepage <http://nccu.edu.tw/j>;
foaf:mbox <mailto:j@cs.nccu.edu.tw>;
/*Generalization*/
foaf:phone <tel:+886-2-29387620>;
.....
foaf:knows [ a foaf:Person;
/*De-identification*/
foaf:name "Kua-Ping Cheng";
rdfs:seeAlso
/*enhanced microdata protection techniques*/
<http://nccu.edu.tw/k/foaf.rdf> ].
foaf:knows [ a foaf:Person;
/*De-identification*/
foaf:name "Ya-Ling Huang";
```

```
rdfs:seeAlso
/*enhanced microdata protection techniques*/
<http://nccu.edu.tw/y/foaf.rdf> ].
..... ]
```

3. Data Releasing Policy (DRP)

   A DRP governs what conditions are acceptable for anonymized WebID dataset disclosure and which anonymized WebID attributes are available for analytics so that the WebID dataset disclosure does not violate the privacy protection principles after verifying the consecutive query auditing logs. Even when WebID attributes disclosure does not violate the privacy protection principles of each single query, the consecutive queries do not necessarily satisfy the pre-setting rules. We need a comprehensive solution to deal with this problem. The super data controller can further initiate an *output perturbation service* with an algorithm, such as differential privacy, to add certain noises to the originally anonymized WebID profile attributes while providing SPARQL queries.

The following is a simple SPARQL query to access anonymized profile attributes with one hop of Social links:

```
Select ?graph ?gender ?age ?member ?interest
From <http://nccu.edu.tw/j/foaf.rdf>
From named graph???
Where{<http://nccu.edu.tw/j/foaf.rdf#me>
foaf:knows ?X.
{ ?X rdfs:seeAlso ?graph.
graph ?graph {[ a foaf:Person.
/*De-identification*/
foaf:mbox ?mbox;
foaf:name ?name;
foaf:gender ?gender;
.....;
/*Generalization*/
foaf:phone ?phone;
/*GlobalRecording*/
foaf:age ?age;
foaf:member ?member;
foaf:interest ?interest;
foaf:knows [ ?graph ]. ]}}}
```

5.3 R and Hadoop for WebID Analytics

Each personal WebID anonymized Profile attribute is defined as a key-value set of RDF(S) triples so that the entire dataset of WebID Profile attribute is a set of key-value RDF(S)-based triples. These WebID datasets of key-value Profile attributes can be applied to various R and Hadoop analytics as follows.

1. In *lightweight data analytics*, the service provides simple analytics for unstructured key-value triples with small mathematical operations, such as sum, average, sort, and median, in a MapReduce of the Hadoop distributed framework.
   MapReduce is a programming model and an associated implementation for processing and generating large datasets. The computation processes a set of input key-values pairs to produce a set of output key-value pairs [13]. The user-written *mapper function* takes an input pair and produces a set of intermediate key-value pairs. The *reducer function*, also written by the user, accepts an intermediate key and a set of values for that key. It merges these values together to form a possible smaller set of values. While Hadoop is an excellent platform for managing large and complex datasets, its capability for complex analytics is rather limited. In fact, Hadoop's "native" complex analytics module, Apache Mahout, is only suitable for highly trained developers with strong backgrounds in Java and MapReduce.
2. In *heavyweight data analytics*, it provides complex analytics of structured data with complex mathematical operations for machine learning, clustering, and trend detection with statistical computational software, such as R.
   R is an open source available language and environment for statistical computing and graphics, and it provides a wide variety of statistical and graphical techniques for heavyweight data analytics, including linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.
3. In *hybrid data analytics* [3] , it provides a combination of a lightweight and heavyweight data analytics to leverage the power of both data analytics services.

RHadoop is a collection of the `plyrmr`,`rmr`,`rhdfs`, and `rhbase` packages that allow users to manage and analyze data with Hadoop. We can use RHadoop as a testing environment to verify the concepts of semantics-enabled policy to call for WebID propagation control services. A data analyst enables hybrid data analytics for WebID datasets by combining the R's heavyweight and the MapReduce's lightweight data analytics.

On the one hand, standard R originally worked only for heavyweight analytics in small memory datasets analytics of a standalone computer, However, R can be extended to the Hadoop cluster computing environment that allows the distributed processing of large datasets. On the other hand, the Hadoop framework with a MapReduce programming paradigm only for lightweight data analytics can incorporate

[3] The hybrid data analytics services of RHadoop platform have been established by Revolution Analytics. The packages have been implemented and tested in Cloudera's or Hortonworks Sandbox's distribution of Hadoop.

heavyweight data analytics through the packages available in the Comprehensive R Archive Network (CRAN). In summary, the purpose of integrating R and Hadoop, i.e., RHadoop, is to bring the distributed (or parallel) MapReduce processing capability of Hadoop to the heavyweight data analytics of R.

6 Related Work

Propagation control for WebID is an access control problem for the decentralized Social Web. This access control problem can be modeled and solved as a policy-aware access control system architecture. Access control is the process of mediating every request for the resources of a system, and it plays an important role in overall system security. In [43], the authors present a comprehensive introduction to access control models and languages to protect data and resources against unauthorized disclosure and usage. Access control services can be defined as a RDF(S) linked data ontology in the decentralized online social network so each resource request will be verified through this access control ontology module [48]. Applying transparent and accountable policies to assure privacy-preserving data mining is a new direction for this research [45].

Decentralized information flow control (DIFC) is a mandatory access control paradigm that allows users to specify how data can propagate through a system [31]. DIFC uses *labels* to denote the sensitivity of data while the *capabilities* allow the principals to acquire or drop labels. The secrecy label is used to prevent data leakage while the integrity label is used to prevent data corruption.

Two Assured Information Sharing (AIS) projects are related to this study: the project AISL and the project Presidio[4]. In the AISL project, the notion of policy-driven assurance of an information sharing lifecycle was proposed for the information value chain. AISL consists of three major phases: (1) information discovery and advertising, (2) information acquisition, release, and integration, and (3) information usage and control.

The Airavat of the project Presidio [34] integrates DIFC and differential privacy to carry out the privacy-preserving computations in the MapReduce framework. The DIFC ensures that the system is free of unauthorized data leaking from untrusted mapper computations, and its various privacy techniques guarantee that "too much" information will not be revealed about any of its inputs.

7 Conclusion and Future Work

A WebID is a single sign-on token for a user's authentication to access multiple social network sites without preparing numerous user names and passwords. This

[4] AISL stands for Assured Information Sharing Lifecycle, and Presidio for Collaborative Policies and Assured Information Sharing.

breaks the walled garden data silo problem that exists in the current centralized online social network. In this study, we demonstrate how to restructure the centralized online social network architecture into the decentralized Social Web for wide-scale WebID capturing, recording, anonymizing, sharing, integration, modeling, and analyzing.

We presented a WebID analytics process to investigate the privacy-preserving WebID disclosure problem by using transparent and accountable propagation control services at the data controllers of an information value chain.

We show how to apply transparent and accountable propagation control services at the data controllers to assure WebID protection for the data owner and to ensure the analytics utility for the data user. In addition, we demonstrate how to provide WebID dataset protection and utility through types of semantics-enabled policy enforcement to call for WebID propagation control services at the data controllers of an information value chain. In our future work, we will investigate a comprehensive solution for the problem inherent in transparent accountable propagation control services. Furthermore, types of semantics-enabled policy for hybrid RHadoop analytics for the WebID datasets will be fully implemented on the decentralized privacy-aware Social Web.

Acknowledgements

This research was partially supported by the NSC Taiwan under Grant NSC 102-2221-E-004-014.

References

1. Adam, R.N., Worthmann, C.J.: Security-control methods for statistical databases: A comparative study. ACM Computing Survey **21**(4), 515–556 (1989)
2. Anderson, A.H.: A comparison of two privacy policy languages: EPAL and XACML. In: Proceedings of the 3rd ACM Workshop on Secure Web Services (SWS'06), pp. 53–60. ACM (2006)
3. Appelquist, D., et al.: A standard-based, open and privacy-aware social web. Tech. rep., W3C Incubator Group Report (2010)
4. Ardagna, A.C., et al.: A privacy-aware access control system. Journal of Computer Security **16** (2008)
5. Barbier, G., et al.: Provenance Data in Social Media. Morgan & Claypoole Publishers (2013)
6. Beckett, D., et al.: Turtle: Terse RDF triple language. Tech. rep., W3C Candidate Recommendation (2013)
7. Berners-Lee, T.: Socially aware cloud storage (2011)
8. Boley, H., et al.: Rule interchange on the web. In: Reasoning Web 2007, Third International Summer School, LNCS 4636. Springer, Dresden, Germany (2007)
9. Bonatti, A.P.: Datalog for security, privacy and trust. In: Datalog Reloaded, LNCS 6702, pp. 21–36. Springer (2011)

10. Carminati, B., Ferrari, E.: Privacy-aware access control in social networks: Issues and solutions. In: J. Nin, J. Herranz (eds.) Privacy and Anonymity in Information Management Systems, pp. 181–195. Springer (2010)
11. Ciriani, V., et al.: Microdata protection. In: T. Yu, S. Jajodia (eds.) Secure Data Management in Decentralized Systems, pp. 291–321. Springer (2007)
12. Cox, H.L., Karr, F.A., Kinney, K.S.: Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act. International Statistical Review **79**(2), 160–183 (2011)
13. Dean, J., Ghemawat, S.: Mapreduce: Simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
14. Dwork, C.: Differential privacy. In: Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), LNCS 4052, pp. 1–12 (2006)
15. Dwork, C.: A firm foundation for private data analysis. Communications of the ACM **54**(1), 86–95 (2011)
16. Editors: Data, data everywhere: a special report on managing information. Tech. rep., The Economist (2010)
17. Eiter, T., et al.: Rules and Ontologies for the Semantic Web. Springer (2008)
18. Foresti, S.: Preserving Privacy in Data Outsourcing. Springer (2011)
19. Gentry, C.: Computing arbitrary functions of encrypted data. Communications of the ACM **53**(3), 97–105 (2010)
20. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers (2011)
21. Hu, Y.J., Boley, H.: SemPIF: A semantic meta-policy interchange format for multiple web policies. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 302–307 (2010)
22. Hu, Y.J., et al.: Crafting a balance between big data utility and protection in the semantic data cloud. In: International Conference on Web Intelligence, Mining and Semantics (WIMS'13). ACM Press (2013)
23. Hundepool, A., et al.: Statistical Disclosure Control. Wiley Series in Survey Methodology (2012)
24. Inkster, T., Story, H., Harbulot, B.: WebID-TLS: WebID authentication over TLS, W3C editor's draft. Tech. rep., W3C (2013)
25. Jung, J.J.: Integrating social networks for context fusion in mobile service platforms. Journal of Universal Computer Science **16**(15), 2099–2110 (2010)
26. Karjoth, G., Schunter, M., Herreweghen, E.V.: Translating privacy practices into privacy promises - how to promise what you can keep. In: POLICY'03. IEEE (2003)
27. Krohn, M., et al.: A world wide web without walls. In: 6th ACM Workshop on Hot Topics in Networking (Hotnets). ACM (2007)
28. Labrinidis, A., et al.: Challenges and opportunities with big data. Tech. rep., Computing Research Consortium (CSR) (2012)
29. Manyika, J., et al.: Big data the next frontier for innovation, competition, and productivity. Tech. rep., McKinsey Global Institute (2011)
30. Moreau, L.: The foundations for provenance for the web. Foundations and Trends in Web Science **2**(2-3), 99–241 (2010)
31. Myers, A.C., Liskov, B.: Protecting privacy using the decentralized label model. ACM Transactions on Computer System **9**(4), 410–442 (2000)
32. Nissenbaum, H.: Privacy Context: Technology, Policy, and the Integrity of Social Life. Standford University Press (2010)
33. Patel-Schneider, F.P., Horrocks, I.: A comparison of two modelling paradigms in the semantic web. Journal of Web Semantics pp. 240–250 (2007)
34. Roy, I., et al.: Airavat: Security and privacy for mapreduce. In: Proceedings of the 7th USENIX Conference on Networked System Design and Implementation (NSDI'10) (2010)
35. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transaction on Knowledge and Data Engineering **13**(6), 1010–1027 (2001)
36. Sambra, A., Story, H., Berners-Lee, T.: WebID 1.0: Web identity and discovery, W3C editor's draft. Tech. rep., W3C (2013)

37. Schwab, K., et al.: Personal data: The emergence of a new asset class. Tech. rep., World Economic Forum (2011)
38. Solove, J.D.: A taxonomy of privacy. University of Pennsylvania Law Review **154**(3) (2006)
39. Spomy, M., et al.: JSON-LD 1.0. Tech. rep., W3C Proposed Recommendation (2013)
40. Stonebraker, M.: What does 'big data mean. BLOG@CACM (2012)
41. Sweeney, L.: K-annonumity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowedge Based Systems **10**(5), 557–570 (2002)
42. Tene, O., Polonetsky, J.: Privacy in the age of big data: A time for big decisions. 64 Stanford Law Review Online 63 (2012)
43. Vimercati, S.D.C.d., et al.: Access control policies and languages in open environments. In: T. Yu, S. Jajodia (eds.) Secure Data Management in Decentralized Systems, pp. 21–58. Springer (2007)
44. Weaver, J., Tarjan, P.: Facebook linked data via the graph API. Semantic Web - Interoperability, Usability, Applicability (2012)
45. Weitzner, J.D., et al.: Creating a policy-aware web: Discretionary, rule-based access for the world wide web. In: E. Ferrari, B. Thuraisingham (eds.) Web and Information Security, pp. 1–31. IGI (2006)
46. Weitzner, J.D., et al.: Transparent accountable data mining: New strategies for privacy protection. Tech. Rep. MIT-CSAIL-TR-2006-007, MIT CSAIL (2006)
47. Wood, D., et al.: Linked Data: Structured Data on the Web. Manning (2014)
48. Yeung A., C., et al.: Decentralization: The future of online social networking. In: W3C Workshop on the Future of Social Networking. W3C (2009)
49. Zheleva, E., Terizi, E., Getoor, L.: Privacy in Social Networks. Morgan&Claypool (2012)