<http://www.diva-portal.org>

Preprint

This is the submitted version of a paper presented at *ARES ISPM 2016 Workshop*.

Citation for the original published paper:

Reuben, J., Martucci, L A., Fischer-Hübner, S., Packer, H., Hedbom, H. et al. (2016)
Privacy Impact Assessment Template for Provenance.

In: *The 11th International Conference on Availability, Reliability and Security (ARES 2016), August 31 – September 2, 2016, Salzburg* IEEE Press

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-43020>

Privacy Impact Assessment Template for Provenance

Jenni Reuben^{*}, Leonardo A. Martucci^{*}, Simone Fischer-Hübner^{*},
Heather S. Packer[†], Hans Hedbom^{*} and Luc Moreau[†]

^{*}Karlstad University

651 88 Karlstad, Sweden

Email: [firstname.lastname]@kau.se

[†]University of Southampton

Southampton SO17 1BJ, UK

Email: {hp3, l.moreau}@ecs.soton.ac.uk

Abstract—

Provenance data can be expressed as a graph with links informing who and which activities created, used and modified entities. The semantics of these links and domain specific reasoning can support the inference of additional information about the elements in the graph. If such elements include personal identifiers and/or personal identifiable information, then inferences may reveal unexpected links between elements, thus exposing personal data beyond an individual's intentions. Provenance graphs often entangle data relating to multiple individuals. It is therefore a challenge to protect personal data from unintended disclosure in provenance graphs. In this paper, we provide a Privacy Impact Assessment (PIA) template for identifying imminent privacy threats that arise from provenance graphs in an application-agnostic setting. The PIA template identifies privacy threats, lists potential countermeasures, helps to manage personal data protection risks, and maintains compliance with privacy data protection laws and regulations.

Keywords—Privacy Impact Assessment; provenance; privacy;

I. INTRODUCTION

Provenance is an enabler for transparency and, thus, for accountability [1]: it permits the origin of data and processes to be reasoned over. Transparency and accountability are basic privacy principles supporting the individuals' rights [2]–[4]. Specifically, individuals have the right to: (a) access their personal data and, (b) be informed about the data processing and its purposes. In addition, individuals have the right to know which stakeholders are responsible and liable for processing and safeguarding their personal data.

Systems that record provenance and handle people's information, such as authentication or health services, may capture personal data. The semantics and elements' identifiers captured in the provenance may support the inference of additional personal data about individuals, or groups of individuals, outside its intended context or purpose. It is thus a challenge when disclosing provenance graphs to protect personal data.

In order to mitigate privacy threats, developers must be privacy-aware when designing the provenance that a system captures, so that privacy requirements are addressed. Moreover, it is essential that privacy requirements are addressed at design time so that they can be adopted in open systems, such as the web. The first step towards supporting the designing of privacy-aware provenance is to identify and assess the privacy threats associated with provenance and the subsequent mitigation countermeasures. The second step is to design, implement and evaluate privacy-preserving technologies for provenance. In this paper, we provide the first step: a Privacy Impact Assessment (PIA) template for provenance in an application-agnostic setting.

A PIA is a systematic method for personal data risk assessment, a procedure that is well-known and established in the personal data protection community [5], and can be tailored to specific technologies and applications [6].¹ To the best of our knowledge, there is no general PIA template that addresses privacy threats in applications that record provenance. Our PIA template grounds the analysis of privacy risks in provenance and it is designed to help developers of applications that record provenance to: (a) reason about privacy threats arising from provenance; and (b) manage the identified threats and set countermeasures.

This paper is organized as follows. In Section III, we present an introductory example using PROV. In Section IV, we describe the PIA process for provenance and detail its component parts: (a) a list of main privacy targets in provenance systems and their privacy threats and (b) a list of proposed countermeasures and mitigation controls for the identified threats. In Section V, we show how the PIA template can be used by application developers. In Section VI, we summarize our main findings, provide our concluding remarks and discuss areas of future research on privacy controls.

This research was funded by SMARTSOCIETY, a research project of the Seventh Framework Programme for Research of the European Community under grant agreements no. 600854.

¹PIA is to be mandatory in Europe, after the ratification of the EU General Data Protection Regulation (GDPR) [3] (expected for Spring 2016).

II. RELATED WORK

The Privacy Impact Assessment Framework for RFID applications by the EU Commission [6] provides a PIA template for RFID applications, which has inspired our work. In contrast to their Framework, which is based on the privacy principles of the EU Data Protection Directive and focused on RFID, our work is based on the privacy principles of the upcoming EU General Data Protection Regulation (GDPR) and related to privacy threats in applications that record provenance. Other available PIA templates, such as the ones from the UK's ICO and the US Department of Homeland Security, are generic rather than application specific [7], [8].

Previous work on privacy assessment of provenance applications was carried out by Kifor et al. [9], who investigated the privacy implications of provenance in health care applications. They have identified two privacy risks (two different variations of threat T4 presented in Table IV-B) that arise due to the inclusion of provenance in health care information systems. Their approach to privacy risks analysis is however build on intuition, i.e., privacy threats are not systematically analyzed according to a list of applicable legal privacy principles, and thereafter the focus is shifted to the mitigation measure proposals for the identified risks. In contrast, we focus on systematic identification of privacy threats in accordance with legal principles of the GDPR. Further, we propose corresponding countermeasures for the privacy threats we identified. Provenance system architects and developers can use our PIA template to identify privacy threats in their systems systematically rather than intuitively.

Existing work on privacy and provenance are most often than not focuses on measures to mitigate privacy threats that arise when several provenance graphs are merged. They are individual discussed in Section IV-C along with the respective threats that they aim to mitigate.

III. EXAMPLE: PERSONAL DATA AND PROV

Provenance can be expressed using the W3C standard PROV [10]. PROV provides a standardized vocabulary and uses unique resource identifiers (URIs) for identifying who and what processes are responsible for data generation and its modification. The use of URIs allows elements mentioned in different provenance graphs to be linked for generating the data lineage. However, the use of URIs may also permit multiple graphs referring to an agent (involving natural persons) to be linked and, as result, personal data to be unintentionally exposed. Furthermore, the naming convention used by the URI leaks additional personal data.

For example, if the URI used for an agent called Anna is `http://computerscience/user/anna.soton.ac.uk`, then the generic top level domain exposes that Anna is affiliated to the University of Southampton (SOTON). It also exposes her name and it can be inferred that she is connected to the computer science department.

Provenance expresses the relationships that hold among the entities, activities and agents. The semantics of these

relationships allow different elements to be explicitly linked from multiple provenance graphs.

For example, consider Alice, a London-based journalist that is traveling to meet Bob, a whistle-blower, in Berlin. Alice pays for her travel and claims the costs. Her employer (a British newspaper) publishes the details of the claim because of its obligation towards data transparency (see Figure 1). When Alice tweets her arrival in Berlin, this action produces the provenance graph in Figure 2. Just before Alice's arrival in Berlin, Bob sends an (encrypted) email to Alice with the meeting details. Bob wants to meet at the *Café*, an international coffee serving venue belonging to a coffeehouse chain, that is known to Alice. Alice is a *Café*'s regular at one of its London branches and has a loyalty card to its shops (which she has registered using her URI). Alice and Bob meet at the *Café*, and discuss the wrongdoings that Bob's company / government / secret society is covering and he wants to reveal. Alice and Bob pay for their snacks and coffee and both use their *Café* loyalty cards (which offers a 10% discount on their expenses). The *Café*'s system records Alice and Bob's payments and link them to the time that the tab was closed and the table they were sitting (see Figure 3). The *Café* make its records public for its customers to track their coffee intake.

Bob did not explicitly state that he was meeting Alice nor where the meeting took place. The *Café* recorded the time of the transaction, the consumers involved and the table they were sitting, but not the branch and its location. However, from combining the provenance graphs in Figures 1–3, a semantic inference can be made that Bob was in Berlin (see Figure 4). Specifically, it can be inferred that Alice arrived in Berlin approximately one hour before meeting Bob, that the meeting was in Berlin and, by inference, that Bob was also there. This inference can be made because of the content and time of Alice's tweet, and the URI of activity “*Café*” in both the *Café*'s loyalty program system and in the employer's data (the British newspaper) linked to Alice's claim.

The conclusion that Alice and Bob met at the “*Café*” in Berlin, which would eventually identify Bob as the whistle-blower, is made from three public provenance records: from the newspaper's accounting system, from Twitter, and from the coffeehouse loyalty program.

IV. PRIVACY IMPACT ASSESSMENT TEMPLATE FOR PROVENANCE

In this section, we describe our PIA template for applications that record provenance. In particular, PIA templates use four steps, ours differ from others because it focus on privacy issues for provenance. It is based on the PIA template for RFID applications published by the EU Commission [6]. Concretely, a PIA template has the following four steps:

- 1) Description of the application (that records provenance, in our case), in terms of its objective, requirements, potential users, the stakeholders involved in data collection and processing, the application's architecture, and data flows.
- 2) Identification of the privacy threats with respect to the privacy targets embedded in the (EU) privacy legislation

Figure 1. From Alice’s employer.

Figure 2. From Alice’s tweet.

Figure 3. From the Café’s loyalty program.

Figure 4. The combined provenance graphs disclose, by inference, that Bob was in Berlin.

and regulation.

- 3) Proposal of technical and organizational controls for mitigating privacy threats.
- 4) Documentation of the PIA regarding the application being analyzed.

We describe steps 2 and 3 for applications that record provenance in the remainder of this section. Step 2 consists of a list of privacy targets, which were compiled by evaluating which of the targets from the upcoming EU GDPR [3]² were applicable to provenance (see Section IV-A) and a list of potential privacy threats to provenance data in Section IV-B. Then for step 3, we describe a non-exhaustive list of controls and countermeasures to mitigate the identified threats in Section IV-C.

A. Privacy Targets

In order to identify the privacy threats (Step 2), we extracted and listed the main privacy targets from the EU GDPR, which are presented in Table I.³ This list is general in terms of

personal data processing, and not specific to the concept of provenance.

B. Privacy Threats

In Section III we illustrated that Bob and his location being inferred (leaked) from provenance data. It exemplifies a case where personal information can be derived from provenance, which is a privacy threat to “*Compliance with the General Data Processing Principles*” target (in Table I). In this section, we build on this example by presenting a list of potential threats to the privacy targets detailed in Section IV-A. This list is non-exhaustive and was compiled by three specialists in personal data protection in a brainstorming sessions, and then reviewed by a fourth specialist.

The goal of Table II is to provide provenance system developers with a starting point to identify the relevant privacy threats, and help them to reason about threat exploitation, likelihood and impact.

C. Identification of Controls and Countermeasures

In this section, we list countermeasures to mitigate the threats T1–9 enumerated in Section IV-B. They include technical and administrative security and privacy controls. We describe the personal data protection goal and the (state-of-the-art) controls required to reach the goal. If no countermeasure for a threat exists, we highlight the problem and explain the research challenge involved.

²The EU GDPR was passed by the European Parliament in Dec. 2015 and is expected to be ratified during Spring 2016. The GDPR was chosen as our reference for many reasons: (a) it applies to data controllers or processors located in the EU, and to any organization processing personal data of EU residents, (b) it reflects the basic privacy principles of the OECD privacy guidelines and (c) of the US Federal Trade Commission’s (FTC) Fair Information Practice Principles (even going beyond them).

³This is an updated list of privacy target found in [6], which were initially taken from the EU Data Protection Directive 95/46/EC [2].

Table I
PRIVACY TARGETS DERIVED FROM THE EU GDPR.

Privacy Targets	Description
Compliance with General Data Processing Principles, data protection by default (Art. 5, 23)	The key privacy principles, including the following ones, are to be ensured, and must be enforced by default: Purpose binding: Personal data must be collected for specified and legitimate purposes and may later only be used for those purposes (Art. 6 1b). Data minimization: The amount of personal data and the extent to which they are collected and processed should be minimized (Art. 6 1c EU Directive), i.e., in particular if possible data should be anonymised or pseudonymised.
Lawfulness of personal data processing (Art. 6)	Lawfulness of processing to be ensured by an unambiguous informed consent, contract or legal obligation.
Lawfulness of processing special categories of data (Art. 9)	Lawfulness of the processing of “sensitive” personal data (such as data related to health, ethnicity, political opinions) must be ensured by explicit consent or special legal basis.
Compliance with the right to be informed (Art. 14)	A data subject, ¹ is to be provided with required privacy policy information including the identity of the data controller, ² and data processing purposes at the time when the data is collected from the data subject.
Compliance with transparency rights (Art. 15)	The data subject has the right to access their data (unless this adversely affects the privacy rights of others) and receive information about data processing purposes, data recipients or categories of recipients, the data retention period, the right to lodge a complaint with a supervisory authority as well as information about the logic involved on any decision making based solely on automatic processing including profiling, and the significance/envisaged consequences of such processing.
Compliance with rights to data portability, rectification, erasure and restricting data processing (Art. 16, 17, 17a, 18)	The data subject can exercise the right to export data, correct or delete their data, the right to restrict its processing, and the right to be forgotten in a timely manner.
Compliance with the right to object (Art. 19, 20)	It must be ensured that the data subject has the right to object to the processing of their data, especially in the case of automated individual decision making and profiling.
Security of processing (Art. 23, 30)	It must be ensured that suitable security measures, including data minimization and pseudonymization, are implemented by default.

¹ A data subject is a “person who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity” [2].

² A data controller is a “person, public authority, agency or any other body which . . . determines the purposes and means of the processing of personal data”. A data processor “processes personal data on behalf of the controller” [2].

T1 Collection of personal data beyond strictly needed.

Well-scoped specific purposes limit the collection of personal data that is strictly required for the identified purposes. Examples of how to define well-scoped purposes is presented in [11]. Effective countermeasures to T1 include administrative controls, such as privacy policies. Privacy policies define the scope and purposes of the personal data processing.

T2 User profiling. To prevent user profiling from provenance graphs, the purpose of a query for provenance should match the consent given to the personal data that is included in the resulting provenance graph. Provenance graphs should not be released for unauthorized purposes. Purpose-oriented disclosure of provenance graphs can be implemented with Hippocratic database systems [12]. Hippocratic database systems verifies whether the purpose of the query and the corresponding result is in compliance with user’s consent.

T3 Vague purposes and function creep. The mitigation of this threat involves well-defined purposes (see T1), and the enforcement of the purposes when the provenance data is processed. Machine readable privacy policies along with enforcement mechanisms can be implemented to mitigate or eliminate this threat. It requires a (formal) privacy policy language, e.g., EPAL [13] and PPL [14], and a policy enforcement mechanism, such as Garm, which identifies and enforces the policies that are applicable to data that is captured in provenance graphs. [15].

T4 Processing of inferred personal data without consent.

Suppression or obfuscation of personal data in the provenance graphs are solutions that prevent inference of additional (sensitive) personal data. The implementation of the countermeasure mostly depend on the provenance application setting. To prevent the disclosure of additional personal data from provenance graphs, query responses from provenance stores can be verified for personal data leakage. Several verification methods can be deployed, such as k -anonymity [16] and ℓ -diversity [17]. The use of such methods for controlling data release in provenance graphs was proposed in [18]. Alternatively, it is also possible to transform the provenance graphs to remove or redact personal data from them. A graph transformation technique that redacts graphs in accordance to privacy laws and regulations, e.g. HIPAA, is presented in [19]. And a graph transformation technique for PROV data model is described in [20]. It replaces sensitive personal data for anonymized data in the PROV graphs.

Furthermore, cryptographic based solutions that allow provenance graph properties to be proven for compliance to privacy policies, without revealing any details can be realized using zero-knowledge proofs of knowledge [21]. To the best of our knowledge, there is no implementation of a such a system specifically for provenance graphs, but instantiations of zero-knowledge proofs for graph as proposed in [22], could be applied in provenance context.

Table II
PRIVACY THREATS TO PROVENANCE

Privacy Threats	Description
T1 Collection of personal data beyond what is strictly needed (data minimization threat)	Abstract and vague purposes may lead to the collection of much more data than strictly needed. For example, refining a purpose of “accountability” to “accountability of user profiles” would reduce the amount of provenance data to be recorded. <i>Example: In our example above (see section III), the Café did not need to collect the information on what table the customers were sitting, for the purpose of letting them track their coffee intake. Removing that information would make the inference that Alice and Bob were there together much more uncertain.</i>
T2 Linkability and user profiling (data minimization threat)	Provenance makes it possible to link data, activities and agents. Such relationships constitute new types of personal information and can be used to create user profiles that include personal data beyond the initial purpose or need of the application and provenance data. <i>Example: Our example above demonstrates how by linking of the three provenance graphs, it can be derived that Bob is a whistleblower.</i>
T3 Vague purposes and function creep (purpose binding threat)	Vague purposes allow provenance to be used for purposes unintended at design time. For example, a purpose of “accountability” could be misinterpreted or extended to allow otherwise unauthorised parties to have access to the provenance data. <i>Example: In our example above, Alice’s employer publishes the spending details for the purpose of transparency. This is a very wide purpose that in the example leads the newspaper to (a) most likely grant unrestricted access to the information and (b) publish more information than needed.</i>
T4 Processing of inferred personal data without consent (threat to processing of special categories of data and to lawful processing of data)	The unforeseen links in provenance graphs could lead to derivation, inference and disclosure of potentially sensitive personal data without unambiguous or explicit consent of the data subjects concerned. <i>Example: In our example, several types of sensitive data can be inferred, e.g. by linking the Café customer data with the expense claim data from the newspaper, leading to the disclosure of Bob as the whistle-blower or political activities in the end.</i>
T5 Multiple data subjects: obstacles to deletion of data (threat to rights of rectification and to be forgotten)	Provenance describes relationships between different elements and, often, to multiple data subjects. Selectively deleting information about a given data subject without affecting others, who may have an interest that their data will not be deleted, might not always be possible. <i>Example: In the Smart Society project, we discussed for instance a health care scenario, in which sensors may capture data about when and for how long health care professionals and the patients have met. This implies that the sensor readings and related provenance data may reveal both personal information about health care professionals and the patients. While the data may initially be only intended for the health care professionals to organise their work and to check whether they spent enough time with each patient, the patient (or their relatives) may still have an interest to access data items that relate to them (e.g., to check whether a patient got the right treatment, especially after the occurrence of an incident). Hence, they may have strong interests that the data is not deleted.</i>
T6 Multiple data subjects: Obstacles to access data (transparency rights threat)	Provenance graphs can entangle personal data from multiple individuals. Access requests by data subjects to provenance graphs that contain their personal data might be denied, because it could disclose data about other data subjects that have an interest to keep their data confidential. <i>Example: In the health care example above, a nurse might not want others to have access to the provenance data related to the sensor readings, as this could be used to monitor her working performance.</i>
T7 Limited transparency to data subjects (transparency rights threat)	The consequences of processing personal data are often difficult to foresee. It is thus difficult for the data controller to inform the data subject about all possible consequences of personal data disclosures. <i>Example: For example, in the case with Alice and Bob, it would have been very hard for the Café to envision that the use of loyalty data would have such far-reaching consequences and therefore informing the users of the loyalty card of such consequences.</i>
T8 Unauthorized access to information (security of processing threat)	If the access control mechanism is not properly selected or configured, it may grant users access to more data than originally intended. This is a general security problem that is aggravated in the case of provenance because fine-grained access control rules are more difficult to be expressed. <i>Example: In the Alice and Bob example, one could argue whether the publication of the loyalty register is in accordance with the specified purpose. If the purpose is for customers to track their coffee intake, then only customers should have access to it and only to their own information. An example of granularity would be that in the uri http://computerscience/user/alice.soton.ac.uk every component of the uri might need its own access rights to be defined.</i>
T9 Obstacles to the implementation of technical measures, such as pseudonymization (security of processing threat)	The use of URIs in provenance enables multiple provenance graphs to be merged, and relationships between URIs established. This makes pseudonymization difficult in practice because for the graphs to be merged, they should remain linkable. This would require the use of the same set of (transaction) pseudonyms across all provenance graphs. <i>Example: In the example of Alice and Bob, there are no requirements for linkability between the expense claim data, the loyalty program data and twitter. Hence, using different pseudonyms (i.e., using so called transaction pseudonyms) for Alice and Bob in the different data sets would make linking harder. However, if there was a requirement for linkability of expense claim data and the loyalty data, e.g. for reimbursement purposes in a proper way, then only “traditional pseudonyms” such as aliases could be used, which would however only provide a weak protection against re-identification attacks.</i>

T5 Obstacles to deletion of data. The solution for this threat should allow the data subjects to delete or unlink their personal data from provenance graphs. However, the deletion of personal records may impact the integrity of provenance and impair its utility. Furthermore, in provenance graphs, personal data belonging to multiple data subjects is often entangled. This makes deletion of specific individual’s personal data difficult to implement because deleting personal data of an individual may impact the privacy rights of other individuals. This is a difficult problem with no clear solutions yet, which opens a new avenue for further research.

T6 Obstacles to access data. Mitigation measures should enable data subjects to access all their personal data recorded by provenance. However, provenance graphs can entangle personal data belonging to multiple individuals. Hence, access might be denied because it may adversely affect the privacy of other data subjects. There is no clear solution for this threat, but graph transformations (see T4) and pseudonyms could be used to obfuscate personal data of other involved individuals thus enhancing their privacy rights.

T7 Limited transparency. Mechanisms to predict privacy consequences of personal data processing either automatically or manually are, at best, limited. In addition, the consequences concerning the exploitation of personal data may evolve as new ways to exploit the data are discovered. It thus very difficult to predict all possible consequences of how personal data in provenance can be exploited. No solution exist for this threat.

T8 Unauthorized access. Access to provenance graphs should only be granted to authorized users. An access control model that takes authorization policies into consideration for determining users’ view⁴ is proposed and formalized in [23]. The security view (i.e., the permitted view for the user role) preserves the abstraction of provenance data structures. A simpler access control model that grant/deny access to nodes, edges and paths of provenance graphs is presented in [24]. In their solution the elements of the graph are grouped for efficiency. However, if the scope of the provenance application settings involve multiple stakeholders with complex access predicates, then a scalable access control model is needed. A scalable Attribute-Based Access Control (ABAC) model to express complex access predicates for provenance graphs is presented in [25]. A similar solution to securely manage access in a collaborative environment can be realized by means of attribute based encryption [26]. It allows access conditions to be defined for the encrypted provenance graphs, which are verified against the public attributes of the users attempting to decrypt the data.

T9 Obstacles to technical measures. Provenance graphs can be pseudonymized using transaction (one-time only) pseudonyms [27]. Transaction pseudonyms prevent activities and entities from different transactions to be linked to an individual in particular, and provide the highest level of anonymity. However, the use of transaction pseudonyms can undermine the utility of provenance graphs, which requires

a certain degree of linkability. The concept of multi-level linkable transaction pseudonyms [28] could be applied in provenance context to allow controlled pseudonym linkability and thus sustaining provenance utility. To best of our knowledge, the use of multi-level linkable transaction pseudonyms in provenance graphs is not yet studied, and it provides an interesting direction for future research.

V. HOW DOES THE PIA TEMPLATE HELP ALICE AND BOB?

In this section, we show how an application’s developers can use the PIA template presented in Section IV to reason about privacy threats to their users and identify countermeasures. We exemplify this using the scenario introduced in Section III as a starting point.

The scenario describes three PROV-enabled applications: Twitter, Alice’s employer’s accounting system and the *Café*’s loyalty system. It is assumed that Twitter publishes provenance graphs of all tweets, that the accounting system is required to disclose all elements in its graphs, and the *Café*’s loyalty system publishes provenance after every transaction. The purpose of the *Café*’s loyalty system to publish the recorded provenance is to allow its users to keep track of their coffee intake.

In order for the developers of *Café*’s loyalty system to assess the privacy impact of provenance that the system records and publishes, they would follow the four steps in the PIA template (see Section IV). In the **first step**, the developers would identify: the loyalty system’s objectives and its functional requirements; the parties involved in collecting and processing the loyalty system’s data; and the personal data recorded in the provenance. Specifically, the provenance that the loyalty system records, capture the identities of the participants and their visit to coffee house by means of URIs, and it also includes a description of the transaction.

A. Identifying Privacy Threats

In the **second step** of the PIA, the privacy threats are identified. The loyalty system’s developers would then use Table II to verify which threats apply to the provenance recorded by the application. The loyalty system may not collect more personal data than needed (T1), but the developers would need to verify if the purpose of the recorded provenance is vaguely defined and could be interpreted differently from intended. That would require the privacy policy included in the end-user license agreement to be precisely formulated, as part of the PIA third step. This countermeasure also preempts threats related to function creep (T3).

By recording the URIs of the elements in the provenance graph allows related provenance graphs from multiple sources to be linked (T2 and T4). The implausibility of foreseeing all possible consequences of personal data disclosure (T7) is applicable as well because of the personal identifiers captured in provenance graphs. From the example presented in III, the loyalty system developers cannot simply envision the use of Alice’s personal URI in different system. However this threat lacks privacy-preserving countermeasures.

⁴A view refers to the result set of a database query.

Threats concerning personal data from multiple data subjects (T5–6) in the provenance graph exist in the loyalty system, as events may refer to two or more individuals. These threats may hinder loyalty system users’ transparency rights and prevent deletion of data in the provenance graphs.

Data security threats (T8) are general to computer applications (they are not be further discussed in this paper). And the inability of implementing technical countermeasures (T9) can only be assessed after the countermeasures for all existing threats (in the loyalty system) are identified.

B. Setting Countermeasures

In the **third step** of the PIA, the countermeasures (technical and organizational controls) are identified. The loyalty system’s developers would then refer to list of countermeasures presented in Section IV-C and select the minimum set of controls that cover the majority or all identified threats.

In order to mitigate T2, T4 and T7, the loyalty system developers need to minimize or eliminate linking points between the provenance recorded by the loyalty system and other publicly available provenance graphs. To realize that, the URIs used in the provenance could be replaced by pseudonyms or encrypted data. The resulting provenance graph could still be made public, but could only be decrypted or verified by authorized parties, which is an effective counter to T8. The loyalty system developers could opt for a commitment scheme. These schema allow a prover, e.g. Bob, to commit to a value towards a set of receivers, e.g., the world, without revealing the value.⁵ That allows the provenance graphs to remain public but the provenance data is only verifiable with Bob’s consent.

However, commitments or data encryption do not provide a straightforward solution to T5, e.g., although Bob might request the loyalty system providers to delete records related to him, but depending on the way the commitment schema is implemented Bob’s personal data is still entangled with Alice’s personal data. A similar argument could be used to show that data encryption alone is not enough to counter T6 because the transparency rights of Alice are jeopardized. Furthermore the use of encryption data or commitment schema might undermine the linking of multiple yet related provenance graphs, which might be desirable in some cases (T9). As discussed in Section IV-C, threats T5–6, which relate to the entangled personal data, are subject to further research, and, as a consequence, so is T9.

In the Section III example, if the developers decided to set for the commitment scheme. This solution guarantees that Alice’s and Bob’s association would not be disclosed in the loyalty system’s provenance graph and also that Bob’s identity and location is not leaked from the merged graphs in Figures 1–3. The loyalty system developers document their solution and list the open threats as part of the **fourth step** of the PIA.

⁵Commitment schemes have two operations: $c \leftarrow \text{Commit}(m, r)$; and $\{true, false\} \leftarrow \text{Open}(c, m, r)$, where m is the message, c is the commitment, and r is a random value.

VI. CONCLUSIONS

In this paper, we present a (PIA) template for provenance in an application-agnostic setting. Privacy threats in provenance have so far been studied for specific provenance application settings, such as scientific workflows and healthcare information system. Identifying privacy threats and setting up countermeasures is of paramount importance when provenance is devised in open systems such as the web. This study provides a PIA template for provenance, which systematically guides the provenance system designers to mitigate privacy breaches arise in their applications. Furthermore, our study highlights the absence of the state-of-the-art counter measures for several privacy threats that arise in provenance.

Research in provenance and privacy is by large focused on securing the provenance information from unauthorized access and on unintentional disclosure of personal or sensitive information from a provenance store. In the provenance research the essence of the research challenges in terms of privacy is to maximize the utility of provenance information while preserving the privacy of the individuals involved. In most cases, the challenges entail the unlinking of personally identifiable information at varying degree of granularity depending on the application of provenance. Hence, the use of unlinkable pseudonyms is potentially an interesting privacy-preserving solution to investigate.

In our future work, we will study the application of unlinkable pseudonyms through a scenario-oriented approach that incorporates the threats introduced in this work. Further, we aim to formalize the linkability problem by defining the suitable attacker models and subsequently quantify the pseudonymity in terms of provenance utility.

REFERENCES

- [1] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, “Information accountability,” *Communications of the ACM*, vol. 51, no. 6, 2008.
- [2] “Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” Official Journal L No.281, 23 Nov 1995.
- [3] European Commission, “Proposal for a regulation of the European Parliament and of the Council on the protection of individual with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), Brussels, 15 December 2015,” Dec 2015.
- [4] T. Mendel, *Freedom of Information: A Comparative Legal Survey*. UN Educational and Scientific Cultural Organization, 2008.
- [5] R. Clarke, “Privacy impact assessment: Its origins and development,” *Computer law & security review*, vol. 25, no. 2, pp. 123–135, 2009.
- [6] European Commission, “Privacy and Data Protection Impact Assessment Framework for RFID Applications,” Jan 2012.
- [7] ICO UK, “Conducting privacy impact assessments code of practice, v. 1.0,” Information Commissioner’s Office (ICO), UK, Tech. Rep., Feb 2014.
- [8] DHS US, “Privacy impact assessments,” U.S. Department of Homeland Security, UK, Tech. Rep., Jun 2010.
- [9] T. Kifor, L. Z. Varga, S. Álvarez, J. Vázquez-salceda, and S. Willmott, “Privacy Issues of Provenance in Electronic Healthcare Record Systems,” in *Proceedings of the 1st Int. Workshop on Privacy and Security in Agent-based Collaborative Environments (PSACE)*, 2006.

- [10] L. Moreau, P. Missier (eds.), K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, S. Miles, J. Myers, S. Sahoo, and C. Tilmes, "PROV-DM: The PROV Data Model," W3C Recommendation, Tech. Rep. REC-prov-dm-20130430, Oct. 2013.
- [11] European Commission – Art. 29 Data Protection Working Party, "Opinion 03/2013 on purpose limitation," April 2013.
- [12] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Hippocratic databases," in *Proc. of 28th Int. Conf. on Very Large Data Bases VLDB*. Morgan Kaufmann, 2002.
- [13] P. Ashley, S. Hada, G. Karjoth, C. Powers, and M. Schunter, "Enterprise privacy authorization language (EPAL 1.2)," *Submission to W3C*, vol. 156, 2003.
- [14] PrimeLife, "PrimeLife – Privacy and Identity Management in Europe for Life: Policy Languages," Available at <http://primelife.ercim.eu/images/stories/primer/policylanguage-plb.pdf>, 2011.
- [15] B. Demsky, "Cross-application data provenance and policy enforcement," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 6, 2011.
- [16] L. Sweeney, " k -Anonymity: a Model for Protecting Privacy," *Int. J. on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, pp. 557–570, 2002.
- [17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, " ℓ -diversity: Privacy beyond k -anonymity," *ACM TKDD*, vol. 1, no. 1, p. 3, 2007.
- [18] S. B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen, "On provenance and privacy," in *ICDT'11*. ACM, 2011, pp. 3–10.
- [19] T. Cadenhead, V. Khadilkar, M. Kantarcioglu, and B. Thuraisingham, "Transforming Provenance Using Redaction," in *Proc. of the 16th ACM Symp. on Access Control Models and Technologies*, ser. SACMAT '11. ACM, 2011, pp. 93–102.
- [20] J. Hussein, L. Moreau, and V. Sassone, "Obscuring provenance confidential information via graph transformation," in *Trust Management IX*. Springer, 2015.
- [21] M. Blum, P. Feldman, and S. Micali, "Non-interactive zero-knowledge and its applications," in *20th Annual ACM Symp. on Theory of Computing*. ACM, 1988.
- [22] T. Groß, "Signatures and efficient proofs on committed graphs and np-statements," in *Financial Crypto 2015*, ser. LNCS, vol. 8975. Springer, 2015, pp. 293–314.
- [23] A. Chebotko, S. Chang, S. Lu, F. Fotouhi, and P. Yang, "Secure scientific workflow provenance querying with security views," in *9th Int. Conf. on Web-Age Information Management*, ser. WAIM, Jul 2008, pp. 349–356.
- [24] A. Syalim, Y. Hori, and K. Sakurai, "Grouping provenance information to improve efficiency of access control," in *Advances in Information Security and Assurance*, ser. LNCS. Springer Berlin Heidelberg, 2009, vol. 5576, pp. 51–59.
- [25] A. Rosenthal, L. Seligman, A. Chapman, and B. T. Blaustein, "Scalable Access Controls for Lineage," in *1st Workshop on the Theory and Practice of Provenance*, ser. TAPP. USENIX, 2009.
- [26] A. Sahai and B. Waters, "Fuzzy Identity-Based Encryption," in *Advances in Cryptology - EUROCRYPT 2005*, ser. LNCS, vol. 3494. Springer, 2005, pp. 457–473.
- [27] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management v.034," 10 Aug 2010, see <http://dud.inf.tu-dresden.de/literatur/>.
- [28] S. G. Weber, *Multilaterally secure pervasive cooperation*. Ios Press, 2012, vol. 9.