Towards Systematic Privacy and Operability (PRIOP) Studies

Rene Meis, Maritta Heisel

Towards Systematic Privacy and Operability (PRIOP) Studies

Rene Meis (http://orcid.org/0000-0001-5274-0324) and Maritta Heisel

paluno - The Ruhr Institute for Software Technology, University of Duisburg-Essen, Duisburg, Germany
{rene.meis, maritta.heisel}@paluno.uni-due.de

Abstract. The assessment of privacy properties of software systems gains more and more importance nowadays. This is, on the one hand because of increasing privacy concerns of end-users due to numerous reported privacy breaches, and on the other hand due to stricter data protection regulations, e.g., the EU General Data Protection Regulation that prescribes an assessment of the privacy implications that a project possibly has. The lack of systematic methods to assist a comprehensive and detailed privacy analysis makes it hard for analysts to address the end-users' and legal requirements. In this paper, we adopt the principles of the hazard and operability (HAZOP) studies, which have successfully been used for safety analyses, to privacy to provide a systematic method to identify the relevant privacy threats for a software to be developed. We propose a method called privacy and operability (PRIOP) studies that allows to systematically analyze the potential privacy issues that a software to be developed might raise, based on the software's functionality at the requirements level.

1 Introduction

Privacy is a software quality that gains more and more attention these days. On the one hand end-users are more concerned about privacy and call for more transparency on how their personal information[1] (PI) is processed [1]. On the other hand different legislators prescribe that data protection/privacy impact assessments ((D)PIAs) are performed, e.g., the European Union in the new EU General Data Protection Regulation. A (D)PIA has to be performed for all kinds of projects that involve the processing of PI. Its goal is to assess the implications of the project on the data subjects' privacy.

A central element of a (D)PIA is the identification and evaluation of privacy threats to estimate the privacy risks implied by the considered project. In this paper, we focus on software projects and want to assist analysts to identify and evaluate the privacy threats of a software project as early as possible during the development process, namely in the requirements engineering phase. To

[1] We consider any information that is related to a natural person as personal information. We call this natural person *data subject*.

do so, we adopt the Hazard and Operability (HAZOP) [2] studies, which have successfully been used to assess the safety implications of a system, to a systematic methodology called Privacy and Operability (PRIOP) studies. We illustrate how PRIOP can be applied based on artifacts produced by the Problem-based Privacy Analysis (ProPAn) method [3].

The rest of the paper is structured as follows. Section 2 introduces a small eHealth scenario as running example, and HAZOP and ProPAn as background of this work. PRIOP is introduced in Section 3. Section 4 discusses related work and Section 5 concludes the paper.

2 Background

Running Example We illustrate how a PRIOP study is performed using an electronic health system (EHS) scenario provided by the industrial partners of the EU project *Network of Excellence (NoE) on Engineering Secure Future Internet Software Services and Systems* (*NESSoS*). This scenario is based on the German health care system which uses health insurance schemes for the accounting of treatments. The functionalities of the considered system cover the management of electronic health records (EHRs) (functional requirements R1 and R2), the interaction with mobile devices of patients (R5 and R6), the accounting and billing of patients (R3 and R4), and providing anonymized medical data for clinical research (R7).

In this paper, we focus on the functional requirement R3. R3 is concerned with the problem that doctors shall be able to perform the accounting of treatments that patients received from them. For this, the treatments, diagnoses, and insurance number of the patient are passed to an external insurance application that provides the connection to the patient's insurance company. This insurance application then returns the information which treatments are beared by the patient's insurance contract and the software-to-be shall create an invoice for the treatments that are not covered by the patient's insurance contract. For this, the doctor additionally enters the costs for the treatments.

Hazard and Operability Studies The international standard IEC 61882 [2] defines what a Hazard and Operability (HAZOP) study is and a process to perform a HAZOP study. HAZOP aims at identifying potential hazards and operability problems. A hazard is defined as the potential source of *"physical injury or damage to the health of people or damage to property or the environment"* [2] and an operability problem is any *deviation* from the intended behavior of the system that leads to non-conformance with its (functional) requirements. During a HAZOP study small parts of a system are analyzed in isolation. To systematically identify the potential hazards or operability problems of these parts, HAZOP proposes the eleven guide words NO, MORE, LESS, AS WELL AS, PART OF, REVERSE, OTHER THAN, EARLY, LATE, BEFORE, and AFTER. These guide words are interpreted in the context of the behavioral characteristics of the part under consideration and lead to deviations of the intended behavior. The derived deviations for a part are documented together with

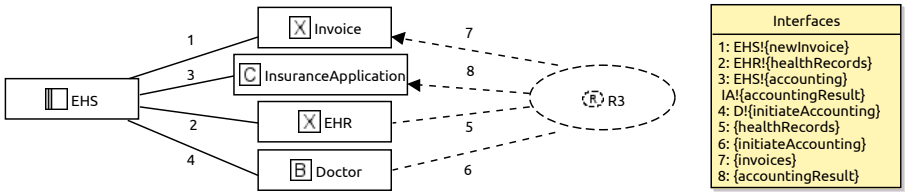

Fig. 1. Problem diagram for functional requirement R3

the *possible causes* of the described situation, its *consequences*, and *safeguards* that shall prevent the occurrence of this situation, or reduce the consequences the deviation may have in a template.

In this paper, we adapt HAZOP to be used in the context of a privacy threat analysis. Next, we introduce the Problem-based Privacy Analysis (ProPAn) method that can be used as a starting point for a PRIOP study.

Problem-based Privacy Analysis To perform a privacy threat analysis, first, the *system*, consisting of the *machine* (software to be developed) and the *environment* it shall be integrated in (cf. [4]), has to be analyzed. To be more precise, it has to be known 1) which PI of which data subjects is processed by the machine, 2) how is this PI collected by the machine, 3) where and how is the PI stored, and 4) to which other entities the machine provides the PI it processes.

In this paper, we demonstrate PRIOP based on inputs provided by the Problem-based Privacy Analysis (ProPAn). ProPAn is a systematic and tool-supported[2] method to perform a privacy analysis starting with a set of functional requirements. The functional requirements represent a decomposition of the overall problem of building the machine and they have to be modeled as problem diagrams following Jackson's problem frame approach [4]. Figure 1 shows the problem diagram for requirement R3 of the EHS example.

It shows on the left the machine EHS (box on the left) and its *interfaces* (lines between the boxes) to the environment (boxes in the middle). The environment of the machine consists of *domains*. Jackson distinguishes three types of domains. Biddable domains (B) are usually people, lexical domains (X) are physical representations of data, and causal domains (C) are objects that behave according to a given specification. The relevant environment for R3 consists of the lexical domains EHR (representing the electronic health records) and Invoice (representing the invoices for treatments that the patient's insurance contracts do not cover), the causal domain InsuranceApplication (which is the interface to the patients' insurances to perform the accounting of treatments patients received), and the biddable domain Doctor (who initiates the accounting).

On the right, the problem diagram shows the *functional requirement* R3 (dashed oval on the right) and its references to the environment. Jackson distinguishes two kinds of references from functional requirements to domains. First, a requirement can *refer to* (dashed line) an event, action, or state of a domain due

[2] `http://www.uml4pf.org/ext-propan`

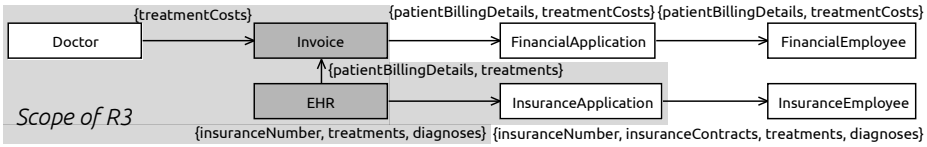

Fig. 2. Excerpt of a graph that visualizes how PI of patients is processed

to which the environment shall behave in the desired way. This desired behavior of the environment is expressed using the second kind of reference. That is, a requirement can *constrain* (dashed line with filled arrow head) events, actions, or states of a domain. R3 refers to the event that the Doctor initiates the accounting and to the EHRs of the involved patients, which contain the PI treatments, diagnoses, insurance number, and the patients billing information. Additionally, R3 constrains that the InsuranceApplication provides the feedback which treatments are covered by the patients' insurance contracts and that a corresponding invoice is created for treatments not beared by the patient's insurance.

ProPAn helps an analysis team that incorporates expertise in requirements engineering, privacy, and the application domain to 1) elicit privacy-relevant domain knowledge [5], 2) identify the PI processed by the system and how it flows through it [6], and 3) derive the relevant privacy requirements for the machine [3]. Figure 2 shows an excerpt of a graph that visualizes the flow of patient's PI due to the functional requirements. This graph is a result of ProPAn's steps to identify the PI that is processed by the system and how it is processed by the system. The gray highlighted part of the graph shows the information flows that were elicited due to R3. The flows outside of the gray part originate from other functional requirements or domain knowledge. The gray printed domains represent designed domains, and the white domains represent given domains. According to Jackson, designed domains are part of the machine and hence, part of the development problem. In contrast, given domains are the parts of the machine's environment that have to be considered as they are, i.e., their specified behavior is not under the control of the development team and cannot be changed. The graph shows that due to R3 treatmentCosts are collected from Doctors (flow from a given domain to a designed domain) and stored in an Invoice together with patientBillingDetails and treatments (flows to a designed domain). In other words, during the privacy analysis with ProPAn, we identified and documented that an invoice contains the previously mentioned PI. Furthermore, insuranceNumber, treatments, and diagnoses flow to the InsuranceApplication (flow from a designed domain to a given domain), due to the machine because of R3. Due to requirement R4 and elicited privacy-relevant domain knowledge, the information provided to the InsuranceApplication flows further to InsuranceEmployees and the patientBillingDetails and treatmentCosts are sent to a FinancialApplication (R4) and further to its employees to perform the billing.

The privacy requirements considered by ProPAn are based on the six protection goals for privacy proposed by Hansen et al. [7]. These consists of the classical

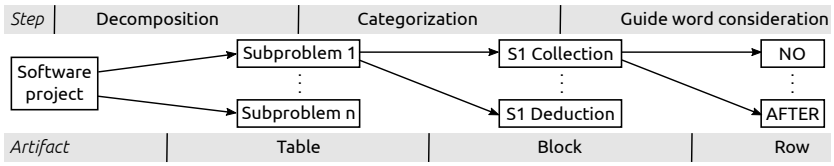

Fig. 3. Steps and artifacts of a PRIOP study

security requirements confidentiality (SC), integrity (SI), and availability (SA) and the privacy goals unlinkability, transparency, and intervenability. We refined the privacy goal unlinkability based on the work of Pfitzmann and Hansen [8] to the privacy requirements anonymity (UA), pseudonymity (UP), undetectability (UU), and data unlinkability (UD). The privacy goal transparency was refined by us in [9] into information requirements for the collection (TC), storage (TS), and flow (TF) of PI, and informing about exceptional cases (TE) concerning the processing of PI. In [10] we refined the privacy goal intervenability into intervention requirements for data subjects (ID) and authorities (IA).

In this paper, we consider the artifacts produced by ProPAn as input for PRIOP, but any method supporting points 1)-3) (mentioned above) can be used.

3 Privacy and Operability Studies

PRIOP aims at a systematic privacy and operability analysis of a software project. Figure 3 visualizes the central steps (arrows) of a PRIOP study and the created artifacts (boxes). First, the software project has to be *decomposed* into subproblems. PRIOP does not prescribe how the decomposition is achieved. For example, Jackson's problem frame approach can be used to derive the project's subproblems. For each of these subproblems, we create a *table* for further analysis. This table should contain a short summary of the subproblem that is considered and should mention who is involved in the PRIOP study of the subproblem. Then each subproblem is *categorized* based on its functionality, as discussed later in this section. For each identified category of the subproblem a *block* is added to the subproblem's table. Finally, the PRIOP guide words have to be *considered* for every combination of subproblem and category. The consideration of a guide word results in a *row* in the block of the considered category in the table of the considered subproblem. In the following, we provide more details on the categorization of subproblems and the consideration of guide words.

PRIOP Operation Categories During the analysis of the identified subproblems, we distinguish four categories of how PI can be processed by the machine. These categories are *collection*, *storage*, *flow*, and *deduction* of PI. An operation is in the category *collection*, if it describes that information is collected by the machine from a given domain. Operations in the category *storage* are concerned with the storage of PI at designed domains. If an operation causes a flow of PI from the machine to a given domain, then it is in the category

flow. Operations in the category *deduction* are concerned with the deduction or computation of PI based on other information.

An operation can be in none (i.e., it does not process PI) or multiple of these categories, depending on the characteristics of the operation. This differentiation of operation categories helps to systematically assess the characteristics of a subproblem in order to identify privacy threats that it possibly causes. To refer to all of these categories simultaneously, we will use the term *processing*.

If the PRIOP study is performed based on ProPAn, then we consider the functional requirements as subproblems and can perform their categorization automatically based on the artifacts created by this method. For requirement R3 of the EHS example, we can see from Figure 2 that it is concerned with the collection of PI from Doctors, the storage of PI at the domain Invoice, and sending PI to the InsuranceApplication. Figure 2 also shows which PI is collected, stored, and flows. Furthermore, the ProPAn model documents (not shown in this paper) that the PI treatmentCosts is derived from the PI treatments, diagnoses, and insuranceContracts by doctors based on the feedback provided by the insurance application. Hence, R3 belongs to all operation categories.

PRIOP Guide Words We consider all HAZOP guide words as useful to identify privacy threats, because these guide words describe in general the deviations that can occur in all kinds of operations a subproblem may be concerned with. We add one additional guide word, namely INCORRECT. This guide word shall cover the cases in which operations are performed incorrectly or with incorrect information as an input. Table 1 shows all PRIOP guide words and our deviation patterns for the four previously introduced operation categories. If these deviation patterns are used for a concrete subproblem, then the terms in angle brackets (< >) have to be instantiated for the subproblem (cf. column deviations in Table 2). The term `<PI>` is instantiated with the PI that is collected/stored/flown/deduced due to the subproblem. If a subproblem is in the category *flow*, then the term `<target>` has to be instantiated with the given domains to which the PI flows. Furthermore, there are some terms in *italics*. While the other terms can be instantiated based on the combination of operation category and subproblem, the italic terms have to be instantiated under consideration of the concrete deviations the guide words imply. The terms *<other PI>* and *<additional PI>* have to be instantiated with the PI that is considered to be unintendedly collected/stored/flown/deduced. *<other domain>*, *<other target>*, and *<additional target>* have to be instantiated with the domains to which information flows unintendedly.

The PRIOP Template The previously introduced guide words shall help to identify deviations of the intended behavior of the operations a subproblem is concerned with. These deviations can lead to violations of privacy requirements and the subproblem's operability. In the case that such an identified deviation leads to a violation of a privacy requirement, the deviation is a privacy threat. We developed a template that is based on the templates proposed to be used in HAZOP studies in [2], but enhanced with additional fields to allow to elicit and

Table 1. Deviation patterns for the all combinations of proposed guide words and operation categories

Guide word	Deviation patterns for operation category			
	Collection	**Storage**	**Flow**	**Deduction**
NO	<PI> is not collected	<PI> is not stored	<PI> does not flow to <target>.	<PI> is not deduced
MORE	More <PI> is collected than intended, including collection of <PI> with additional methods, with higher linkability, in higher amount, or with higher availability.	More <PI> is stored than intended, including storage of <PI> with higher linkability, in higher amount, with higher availability, or with longer duration.	More <PI> flows to <target> than intended, including flow of <PI> with higher linkability, in higher amount, or with higher availability.	More <PI> is deduced than necessary, including deduction of <PI> with higher linkability, in higher amount, or with higher availability.
LESS	Less <PI> is collected than intended, including collection of <PI> with less methods, with lower linkability, in lower amount, or with lower availability.	Less <PI> is stored than intended, including storage of <PI> with lower linkability, in lower amount, with lower availability, or with shorter duration.	Less <PI> flows to <target> than intended, including flow of <PI> with lower linkability, in lower amount, or with lower availability.	Less <PI> is deduced than necessary, including deduction of <PI> with lower linkability, in lower amount, or with lower availability.
AS WELL AS	In addition to <PI> <*additional PI*> is collected or in addition to the software-to-be <*other domains*> collect the <PI>.	In addition to <PI> <*additional PI*> is stored.	In addition to <PI> <*additional PI*> flows to <target>, or<PI> flows to an <*additional target*>.	In addition to <PI> <*additional PI*> is deduced.
PART OF	Only a part of <PI> is collected.	Only a part of <PI> is stored.	Only a part of <PI> flows to <target>, or <PI> flows to fewer targets.	Only a part of <PI> is deduced.
INCORRECT	The collected <PI> is incorrect.	The stored <PI> is incorrect.	The <PI> flowing to <target> is incorrect.	The deduced <PI> is incorrect.
REVERSE	<PI> flows from machine to source of collection.	<PI> is deleted.	<PI> or <*other PI*> flows from <target> to the machine.	<original PI> is or can be deduced from <PI>.
OTHER THAN	<*other PI*> is collected instead of <PI>.	<*other PI*> is stored instead of <PI>.	<*other PI*> flows to <target> instead of <PI> or <PI> flows to <*other target*>.	<*other PI*> is or can be deduced instead of <PI>.
EARLY	<PI> is collected earlier than intended relative to clock time.	<PI> is stored earlier than intended relative to clock time.	<PI> flows earlier than intended to <target> relative to clock time.	<PI> is deduced earlier than intended relative to clock time.
LATE	<PI> is collected later than intended relative to clock time.	<PI> is stored later than intended relative to clock time.	<PI> flows later than intended to <target> relative to clock time.	<PI> is deduced later than intended relative to clock time.
BEFORE	<PI> is collected before another prior operation. E.g., collection before gaining consent.	<PI> is stored before another prior subsequent operation. E.g., storing before gaining consent, or before anonymization.	<PI> flows before another prior operation to <target>. E.g., sending before gaining consent, or before anonymization.	<PI> is or can be deduced before another prior operation. E.g., deduction before gaining consent, or before anonymization.
AFTER	<PI> is collected after another subsequent operation. E.g., collection of up-to-date information after it was needed or the data subject withdrew consent.	<PI> is stored after another subsequent operation. E.g., storage of after another operation would have needed <PI>, or data subject has withdrawn consent.	<PI> flows after another subsequent operation to <target>. E.g., operation on <target> is performed before the up-to-date <PI> was provided, or data subject has withdrawn consent.	<PI> is or can be deduced after another subsequent operation. E.g., deduction after another operation would have needed <PI>, or data subject has withdrawn consent.

Table 2. Excerpt of the instantiated template for functional requirement R3

Guide word	Deviations	Possible causes	Likelihood	Consequences	Harmed Privacy Requirements	Impact
Operation category: Collection		**from:** Doctor		**PI:** treatmentCosts		
NO	treatmentCosts are not collected	Doctor forgets to enter the costs of his/her treatments or does not save the changes made.	Unlikely	Doctor will not get paid and Patients not billed.	-	-
MORE	Higher treatmentCosts are collected than intended.	Doctors incidentially enter costs for treatments not performed or too high treatment costs	Unlikely	Patients will get too high bills or are billed for treatments that they did not receive.	1) SI, 2) TE, 3) ID, and 4) IA for treatmentCosts	1) Major 2), 3), and 4) Moderate
Operation category: Storage		**at:** Invoice		**PI:** treatments, patientBillingDetails, treatmentCosts		
BEFORE	treatments, patientBillingDetails, and treatmentCosts are stored in Invoice before it is known which treatments are beared by the Patient's insurance contract.	a) Doctor explicitly initiates the creation of an invoice without knowing whether the concerned treatments are beared by the Patient's insurance contract b) A software error causes the creation of the invoice without before having the necessary information	Unlikely	Patients will get too high or too low bills.	1) SI, 2) TS, 3) TE, 4) ID, and 5) IA for treatments, patientBillingDetails, treatmentCosts	Moderate
Operation category: Flow		**to:** InsuranceApplication		**PI:** insuranceNumber, diagnoses, treatments		
LESS	Less diagnoses and treatments flow than the patients received, or it is not possible for the insurance application to link the diagnoses and treatments to the Patient's insuranceNumber.	a) Software error b) Insurance Application is temporary unreachable	a) Unlikely b) Possible	i) If too few diagnoses are transmitted to the InsuranceApplication or if diagnoses and insuranceNumber are not linkable to the treatments, then the Insurance Application is not able to perform the accounting. This will result in higher bills. ii) If treatments that are not beared by the Patient's insurance contract are not transmitted to the InsuranceApplication, then no invoice will be created for them.	i) 1) SI, 2) TF, 3) TE, 4) ID, and 5) IA for treatments, diagnoses and insuranceNumber ii) -	i) Major
INCOR-RECT	The insuranceNumber, diagnoses, or treatments flowing to InsuranceApplication are incorrect.	a) Software error b) Incorrect data in EHR	a) Unlikely b) Rare	Patients will get too high bills or are billed for treatments that they did not receive.	i) 1) SI, 2) TF, 3) TE, 4) ID, and 5) IA for treatments, diagnoses and insuranceNumber	Major
Operation category: Deduction		**by: Doct**or and InsuranceApplication		**PI:** treatmentCosts		
RE-VERSE	healthInsurances, diagnoses, and treatments can be deduced from treatmentCosts.	TreatmentCosts may allow to deduce the treatments performed, diagnoses, or insuranceContracts if observed over a longer time especially if to some extent the date of deduction is known.	Rare	Financial employees that are able to observe the treatment costs over a longer time might be able to deduce the treatments, diagnoses, or insuranceContracts of specific Patients.	1) SC against Financial Employees, 2) TF, 3) TE, 4) ID, and 5) IA for treatments, diagnoses, and insuranceContracts.	Major

document attributes that are needed for a later risk evaluation of the identified privacy threats.

An excerpt of the PRIOP template instance for R3 of the EHS example is shown in Table 2. We omit the general information about the subproblem and the people involved in the PRIOP study. In Table 2, we see for each operation category a block (introduced with a row with black background). We selected one or two guide words for each operation category block to illustrate how the proposed template could be filled. For each operation category it is documented from, at, to, or by which domain which PI is collected, stored, flows, or is deduced, respectively. The columns are separated into three areas.

The first two columns show the considered guide word for the row and the deviations it can lead to for the operation category and the considered subproblem. Our deviation patterns shown in Table 1 can be used as starting point for the derivation of the deviations implied by a guide word. The terms `<PI>` and `<target>` can be instantiated with the corresponding information provided in the operation category block. Nevertheless, the deviation pattern instances need to be modified to fit into the context of the subproblem. The deviations possibly represent privacy threats or operability issues.

The second area consists of the third and fourth column. In this area, the analysis team has to document the identified causes that possibly lead to the deviations. Additionally, the likelihood of each cause shall be documented. The analysis team should agree on a common likelihood scale, be it qualitative or quantitative. A common scale will make it easier to homogeneously evaluate the risks implied by the identified privacy threats.

The third area consists of the last three columns. This area is concerned with the consequences the deviations may have on the privacy requirements or the operability of the subproblem. The consequences are first documented as free text, then the harmed privacy requirements are explicitly listed, and it is documented to which degree the described consequences impact the listed privacy requirements. Similar to the likelihood scale, the analysis team has also to agree on a consequence scale.

If the analysis team identified possible causes for a guide word and consequences that harm privacy requirements, then the deviation represents a privacy threat. Whether and how this threat has to be further assessed is in most cases determined using a risk matrix that defines which combinations of likelihood and consequence of a threat are acceptable and which are not. Our template already provides this information such that a risk matrix can easily be filled based on an instantiated template.

We only discuss the last row of the template instance for R3 in Table 2. The row is concerned with the deduction of the treatment costs for treatments not covered by the patients' insurance contracts which is performed by doctors based on the result of the accounting provided by the insurance application. The deviation that is derived for the guide word REVERSE is that the patient's PI healthInsurances, diagnoses, and treatments can be deduced from treatmentCosts. This deviation could be possible if the treatment costs are observed over

a longer time, e.g., because specific diagnosed illnesses could imply a series of treatments that lead to specific treatment costs allowing to conclude from the treatment costs the diagnosed illness and received treatments. The analysis team decided that this is rarely possible. As a consequence the financial employees who are able to observe the treatment costs over a longer time might be able to deduce the treatments, diagnoses, or insuranceContracts of specific patients. The analysis team identified that this consequence harms a confidentiality requirement saying that the deducable PI shall not be disclosed to financial employees. Furthermore, the transparency requirements that are concerned with informing the patient about the flow of and exceptional cases for the deducable PI and the related intervenability requirements for the patient and authorities are harmed. From the documented consequence a major impact on all listed privacy requirements is expected.

The shown template can be enriched with further columns. For example, it can be helpful to provide additional columns to document rationales, e.g., why a specific likelihood was selected for a possible cause, why a possible cause has a documented consequence, or why a consequence impacts the stated privacy requirements in the defined way. Furthermore, already existing safeguards or possible treatments could be documented that shall either reduce the likelihood of a possible cause or the consequence on a privacy requirement.

Relation of Guide Words to Privacy Requirements If the taxonomy of privacy requirements used by ProPAn (see Section 2) is used, we can provide additional support to instantiate the template. Based on the deviation patterns (see Table 1), we identified the privacy requirements that are expected to be harmed by a deviation. Table 3 shows the relations that we identified. An "X" in the table means that a deviation implied by the guide word for the operation category, could harm the respective privacy requirement. If a cell is empty or a privacy requirement is not mentioned, then we do not expect a violation of this privacy requirement for deviations implied by the respective guide word and operation category. In Table 3, we use the abbreviations for the privacy requirements that were introduced in Section 2 and we introduce three groups (Gn) of privacy requirements that share the combinations of guide words and operation categories for which they are relevant. This mapping of combinations of guide words and operation categories to privacy requirements shall help to identify the privacy requirements that are harmed by an identified deviation, but it could also serve as a starting point to elicit scenarios that violate the privacy requirements under consideration of the guide word and operation category.

We identified that for all combinations of guide words and operation categories the privacy requirements integrity (SI), availability (SA), exceptional information (TE), data subject intervention (ID), and authority intervention (IA), which all belong to group G1, might be harmed. This is, because every change in the behavior of an operation could damage the integrity and availability of the processed information, and every change of the way that the PI is processed by the machine could lead to exceptional cases about which the data subject has to be informed and that could violate intervention options the data subject or

Table 3. Privacy requirements that might be harmed by guide words' deviations

Guide Words	Collection			Storage		Flow			Deduction	
	G1, TC	G3	TF	G1, TS	G3	G1, TF	G3	TC	G1, G2	G3
NO, LESS, PART OF, INCORRECT	X			X		X			X	
MORE, AS WELL AS, EARLY, LATE, BEFORE, AFTER	X	X		X	X	X	X		X	X
REVERSE	X	X	X	X		X	X	X	X	X
OTHER THAN	X	X	X	X	X	X	X		X	X

G1 = {SI, SA, TE, ID, IA}, G2 = {TC, TF, TS}, G3 = {SC, UU, UA, UD, UP}

authorities have. Additionally, all modifications of how PI is collected, stored, or flows can lead to a violation of the transparency requirements collection, storage, and flow information, respectively. A change in the deduction of information might affect collection, storage, and flow information requirements (G2).

Group G3 consists of the privacy requirements confidentiality (SC), undetectability (UU), anonymity (UA), data unlinkability (UD), and pseudonymity (UP). These requirements might be relevant for the guide words MORE, AS WELL AS, OTHER THAN, EARLY, LATE, BEFORE, and AFTER in all operation categories, because the guide words imply either that more, additional or other information is processed by the machine, or in a different order, earlier, or later as expected, which could lead to a violation of these requirements. Note that for the guide words MORE, EARLY, LATE, BEFORE, and AFTER, the requirements about the PI that is processed are affected. In contrast, for the guide words AS WELL AS and OTHER THAN, the requirements about the additional or other PI that is processed in addition to or instead of the PI that originally should be processed could be harmed. The guide words NO, LESS, PART OF, and INCORRECT are not implying a violation of the privacy requirements in group G3, because they only concern that fewer or incorrect PI is processed, which does not harm the privacy requirements contained in G3. The guide word REVERSE is interpreted differently depending on the operation category (cf. Table 1). Hence, for the categories collection, flow, and deduction it might harm the requirements in G3, but for the category *storage* it does not.

The transparency requirement flow information (TF) might be harmed by deviations for the guide words REVERSE and OTHER THAN in the operation category *collection*, because they possibly imply a flow from the machine to another domain that is not intended. Similarly, collection information requirements (TC) might be harmed by deviations for the guide word REVERSE in the operation category *flow*. This is, because these scenarios would consider that instead of sending information to other domains, the machine would receive (collect) this or even other information which might be unintended.

Discussion The procedure described in [2] to perform a HAZOP study stresses that for an analysis, the team has to carefully select the guide words that are considered for the system under consideration. Similarly, it can be the case that only a subset of the proposed guide words is relevant for a PRIOP study of a specific software project and that even additional guide words are identified as

important. Hence, we do not claim that our selection of guide words represents a complete set of guide words relevant for the identification of privacy threats of a software project, but expect that it provides a good foundation.

Similarly, the operation categories could be extended. For example, Gürses [11] mentions that information can be collected, used, processed, distributed, or deleted. Collection and distribution (flow to other domains) are covered by our proposed categories. Usage contains from our point-of-view deduction and storage, but other kinds of usage might be identified for a concrete system as additional operation categories. Processing is considered by us as a high-level term describing that something is done with the PI, be that collection, storage, etc. Deletion is an additional category that is worth to analyze in future work, because it is only partly covered by PRIOP. The HAZOP standard does not categorize operations in a way that we propose in this work, but we think that making these operation categories explicit can help analysts to identify scenarios that lead to a harm of privacy requirements. Nevertheless, it can also be valuable to consider the guide words for a given subproblem without considering the operation categories, because this could prevent that the scope of the considered deviations is unnecessarily limited to the operation categories.

Anyway, no method for the identification of any kind of threats can guarantee to elicit a complete set of relevant threats [12]. Nevertheless, we think that our proposed systematic analysis will help analysts to identify, evaluate, and document the privacy threats relevant for their software projects.

An important point that always needs to be assessed critically is the scaleability of a proposed analysis method. If we perform a PRIOP study, then we have to fill in a template for every subproblem. For each operation category a subproblem is assigned to, we have to consider the 12 guide words. That means that in the worst case, we have to fill in 48 rows of the proposed template for each subproblem. Our observation is that this maximum is rarely reached. If it is reached, this is an indicator that the subproblem could be further decomposed into simpler subproblems, because it includes collection, storage, flow, and deduction of PI. Overall, we expect that the effort that has to be spent to perform a PRIOP study scales linearly with the complexity of the software project. The central attributes describing the complexity of the software project for a PRIOP study are the number of subproblems, data subjects, and PI that shall be processed by the machine. For the EHS example, we filled out 168 rows in total. This took us 28 hours in total and 10 minutes per row in average.

Limitations of PRIOP are that 1) the analysis of the subproblems in isolation may not be sufficient if threats arise from the combination of different functionalities, 2) the analysis is limited to the documented subproblems and hence, PRIOP will not help in identifying privacy threats if subproblems are missing or lack important details, and 3) the success of a PRIOP study depends on the analysis team. To address limitation 3), we encourage that the analysis team has to incorporate expertise in requirements engineering, privacy, and the application domain.

4 Related Work

Deng et al. [13] propose a privacy threat analysis framework called LINDDUN. LINDDUN considers the high-level privacy threats linkability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and policy/consent noncompliance, which are negations of popular privacy goals. For the considered system, a DFD (data flow diagram) is created. For each combination of privacy threat and DFD element kind, a threat graph is provided. These are used to derive the possible concrete privacy threats that have to be handled. Based on the high-level privacy threats, the authors also suggest PETs (privacy enhancing technologies) that shall help to mitigate the concrete threats. In comparison to our work, the threat graphs of LINDDUN provide more detailed information that may help to identify whether a high-level privacy threat is relevant or not. But it is possible that the usage of these threat graphs unnecessarily limits the scope of the privacy threat analysis. In future work, we want to elaborate how LINDDUN and PRIOP could be combined to provide better support for the identification of privacy threats.

Several authors investigated the needs of (D)PIAs and methodologies that can be followed in order to perform a (D)PIA. Wright [14] gives an overview of the state of the art in PIA. Oetzel and Spiekermann [15] describe a methodology to support a complete process for a PIA, and Bieker et al. [16] describe a methodology for a DPIA under the EU General Data Protection Regulation. The proposed methodologies describe which steps have to be performed in which order to perform a (D)PIA, but they do not describe concrete techniques that can be used to systematically identify privacy threats. PRIOP can be used to realize the threat identification and risk evaluation steps of the proposed methods. Alnemr et al. [17] propose a DPIA methodology for clouds. They support the identification of privacy threats based on an exhaustive questionnaire. This questionnaire is complementary to PRIOP, and we want to investigate in future work how the questionnaire can be integrated into PRIOP.

5 Conclusions

In this paper, we present with PRIOP a systematic method to identify and document privacy threats and operability issues of software projects. During a PRIOP study, possible deviations of the software project's subproblems are examined under consideration of the four operation categories collection, storage, flow, and deduction. The deviations of a subproblem in the context of the relevant operation categories are derived using the twelve proposed PRIOP guide words. Deviation patterns are provided by PRIOP for all combinations of guide words and operation categories to support an analysis team. The identified deviations for the guide words then have to be further analyzed for possible causes and consequences they might have on the privacy requirements of the software project or the operability of the subproblem. To further support the execution of a PRIOP study, we provide a mapping that shows which privacy requirements

could be harmed by a deviation for a combination of guide word and operation category. The documentation created using PRIOP can be used to further assess the risks implied by the identified privacy threats. We illustrated PRIOP using an EHS example and artifacts produced with the ProPAn method.

In future research, we will integrate PRIOP into the ProPAn tool to benefit from the artifacts created using the ProPAn method. Furthermore, we want to investigate how generic threats, e.g., in the form of threat patterns as introduced by Uzunov and Fernandez [18] for security, can be related to the operation categories and guide words to further assist the identification of privacy threats. The evaluation of PRIOP using a real case study is also future work.

References

1. GSMA: MOBILE PRIVACY: Consumer research insights and considerations for policymakers. `http://goo.gl/pAcvAm` (accessed on 1 March 2017) (February 2014)
2. IEC: IEC 61882:2001 Hazard and Operability Studies (HAZOP Studies) – Application Guide (2001)
3. Meis, R., Heisel, M.: Computer-aided identification and validation of privacy requirements. Information **7**(2) (2016)
4. Jackson, M.: Problem Frames. Analyzing and structuring software development problems. Addison-Wesley (2001)
5. Meis, R.: Problem-Based Consideration of Privacy-Relevant Domain Knowledge. In: Privacy and Identity Management for Emerging Services and Technologies. IFIP AICT 421. Springer (2014)
6. Meis, R., Heisel, M.: Supporting privacy impact assessments using problem-based privacy analysis. In: Software Technologies - 10th International Joint Conference, ICSOFT 2015, Revised Selected Papers. CCIS 586, Springer (2016) 79–98
7. Hansen, M., Jensen, M., Rost, M.: Protection goals for privacy engineering. In: IEEE Symposium on Security and Privacy Workshops, SPW, IEEE Computer Society (2015) 159–166
8. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management (August 2010) v0.34.
9. Meis, R., Heisel, M., Wirtz, R.: A taxonomy of requirements for the privacy goal transparency. In: Trust, Privacy, and Security in Digital Business. LNCS 9264, Springer (2015) 195–209
10. Meis, R., Heisel, M.: Understanding the privacy goal intervenability. In: Trust, Privacy, and Security in Digital Business. LNCS 9830, Springer (2016) 79–94
11. Gürses, F.S.: Multilateral Privacy Requirements Analysis in Online Social Network Services. PhD thesis, Katholieke Universiteit Leuven (2010)
12. Young, W., Leveson, N.G.: An integrated approach to safety and security based on systems theory. Commun. ACM **57**(2) (February 2014) 31–35
13. Deng, M., Wuyts, K., Scandariato, R., Preneel, B., Joosen, W.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requir. Eng. **16** (March 2011) 3–32
14. Wright, D.: The state of the art in privacy impact assessment. Computer Law & Security Review **28**(1) (2012) 54 – 61

15. Oetzel, M., Spiekermann, S.: A systematic methodology for privacy impact assessments: A design science approach. European Journal of Information Systems **23**(2) (2014) 126–150
16. Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., Rost, M.: A process for data protection impact assessment under the european general data protection regulation. In: Privacy Technologies and Policy - 4th Annual Privacy Forum, APF. Volume 9857 of LNCS., Springer (2016) 21–37
17. Alnemr, R., Cayirci, E., Corte, L.D., Garaga, A., Leenes, R., Mhungu, R., Pearson, S., Reed, C., de Oliveira, A.S., Stefanatou, D., Tetrimida, K., Vranaki, A.: A data protection impact assessment methodology for cloud. In: Privacy Technologies and Policy - 3rd Annual Privacy Forum. LNCS 9484, Cham, Springer (2015) 60–92
18. Uzunov, A.V., Fernandez, E.B.: An extensible pattern-based library and taxonomy of security threats for distributed systems. Computer Standards & Interfaces **36**(4) (2014) 734 – 747