Problem-Based Consideration of Privacy-Relevant Domain Knowledge

Rene Meis

Problem-Based Consideration of Privacy-Relevant Domain Knowledge

Rene Meis

paluno - The Ruhr Institute for Software Technology – University of Duisburg-Essen
rene.meis@paluno.uni-due.de

Abstract. Especially for a privacy analysis, an adequate and accurate consideration of domain knowledge is needed. Domain knowledge is often only implicitly given and mainly stored in the minds of domain experts. It is important to make this implicit knowledge explicit and to use it in the privacy analysis of a software system. To our knowledge, no privacy-aware requirements engineering approach exists yet which explicitly considers the elicitation of privacy-relevant domain knowledge. This paper presents an extension of the problem-based privacy analysis (ProPAn) method. The extension consists of three parts. First, we elicit the relevant domain knowledge based on questionnaires which are derived from the stakeholder analysis literature. Second, we present generic patterns which can be instantiated to represent the elicited knowledge. Last, we extend the definitions of ProPAn's privacy graphs to take into account the domain knowledge.

1 Introduction

The quality of a privacy threat analysis strongly depends on the domain knowledge which is considered during the analysis. In general, the elicitation of domain knowledge for the development of a software system has a limited scope. Only those stakeholders and domains are identified that directly take part or are part of a functionality of the system-to-be. We will call these stakeholders *direct stakeholders*. But privacy threats in software systems can also stem from *indirect stakeholders* whose privacy is vulnerable or who possibly affect the privacy of other stakeholders. Another source of privacy threats is the missing domain knowledge about the behavior and further usages of legacy systems which are part of the system-to-be.

In this paper, we will consider privacy requirements, expressing that personal information of *stakeholders* shall not be accessible by *counterstakeholders*. In contrast to the term *attacker* a counterstakeholder may obtain personal data about the stakeholder involuntarily. We have a *privacy threat* for a privacy requirement in our system if there is an information flow from the stakeholder to a domain that is accessible by the counterstakeholder. Note that privacy has more facets than information flow and access control, such as transparency and intervenability [14].

This paper presents an extension of the Problem-based Privacy Analysis (ProPAn) method [4] considering the elicitation, modeling, and use of domain knowledge. ProPAn provides assistance for the initial steps of any given privacy analysis, which is to figure out those parts of the system where personal information of stakeholders can be disclosed by counterstakeholders. The focus of ProPAn, as we presented it in [4], is the

privacy analysis based on the functional requirements that have to be satisfied by the system-to-be. The extension of ProPAn that we present in this paper adds the consideration of domain knowledge to the privacy analysis. For a structured elicitation of the domain knowledge, we use questionnaires. Requirements engineers shall answer these questionnaires in cooperation with domain experts. On the basis of the answers of the questionnaires, we model the domain knowledge using the UML4PF-tool [5] on which ProPAn's tool-support is built. To assist the requirements engineer in the modeling process, we extended the ProPAn-tool[1] with wizards that generate the domain knowledge diagrams based on the answers of the questionnaires. Further, we extended the privacy threat graph generation of ProPAn such that the modeled domain knowledge is used.

The rest of the paper is structured as follows. First, we introduce the problem frames approach, UML4PF and ProPAn, as background of this paper in Section 2. In Section 3, we present the contribution of this paper. Section 4 presents the results of the empirical validation of our questionnaires. Then we discuss related work in Section 5. Finally, Section 6 concludes the paper and describes our future work.

2 Background

The problem frames approach, UML4PF, and ProPAn are described in this section.

Problem Frames Approach The problem frames approach is a requirements engineering approach proposed by Jackson [8]. The first step of the problem frames approach is to create a *context diagram*. A context diagram represents the environment (e.g. stakeholders, other software) in which the machine (i.e. software) shall be built. The context diagram consists of domains and connections between them. Jackson distinguishes the domain types causal domains that comply with some physical laws, lexical domains that are data representations, and biddable domains that are usually people. Connections between domains describe the phenomena they share. Both domains can observe the shared phenomena, but only one domain has the control over a phenomenon (denoted by a "!"). Then the problem of building the system-to-be is decomposed until subproblems are reached which fit into problem frames. Problem frames are patterns for frequently occurring problems. An instantiated problem frame is represented as a problem diagram which in addition to a context diagram also contains a requirement. A requirement can refer to and constrain phenomena of domains. Both relations are expressed by dependencies from the requirement to the respective domain annotated with the referred to or constrained phenomena. An example for a context diagram and a problem diagram is given in Fig. 1 in Section 3.

We use the UML4PF-framework to create problem frame models. UML4PF consists of a UML profile which comes with stereotypes that allow to represent problem frame diagrams as UML class diagrams. The UML4PF-tool stores all diagrams in one global UML model. Hence, we can perform analyses and consistency checks over multiple diagrams and artifacts of the software development process. A more detailed description can be found in [5].

[1] available at `http://www.uni-due.de/swe/propan.shtml`

ProPAn ProPAn extends the UML4PF-framework with a UML profile for privacy requirements and a reasoning technique. A privacy requirement in ProPAn consists of two domains of the system-to-be, namely a *stakeholder* and a *counterstakeholder*. It states that the counterstakeholder shall not be able to obtain personal information of the stakeholder using the system-to-be. The reasoning technique identifies the domains to which personal information of the *stakeholder* can flow. The information flow of the whole system is represented by the global information flow graph which is automatically generated by the ProPAn-tool from the requirements represented as problem diagrams. Formally, the global information flow graph $\mathcal{G}$ is a directed graph with domains as nodes and edges annotated with problem diagrams. An edge $(d_1, p, d_2) : \mathsf{Domain} \times \mathsf{ProblemDiagram} \times \mathsf{Domain}$ denotes that there is a possible information flow from domain d_1 to domain d_2 which stems from the requirement of problem diagram p. Due to the semantics of problem diagrams, there are possibly information flows from each domain in the problem diagram to the domains constrained by the requirement. In this paper, we annotate the edges in the graphical representation of the privacy graphs with the requirements of the problem diagrams. An example for a global information flow graph is the not dashed part of Fig. 3 in Section 3.

From $\mathcal{G}$ the stakeholder information flow graph $\mathcal{S}_s$ is generated, which is a subset of $\mathcal{G}$. $\mathcal{S}_s$ is generated for the stakeholder s of the privacy requirement, and it shows the possible information flows starting from the stakeholder s. Stakeholder information flow graphs are printed with thin lines and filled arrowheads.

Additionally, our technique identifies the domains to which the *counterstakeholder* c has access. This information is captured in the counterstakeholder graph $\mathcal{C}_c$. This graph is generated for the counterstakeholder c of the privacy requirement. $\mathcal{C}_c$ is of the same type as the information flow graphs, but its edges have a different semantics. An edge $(c, p, d) \in \mathcal{C}_c$ denotes that the counterstakeholder c may gain information from the domain d due to problem diagram p. Counterstakeholder graphs are represented with bold lines and empty arrowheads.

For a privacy requirement with stakeholder s and counterstakeholder c, $\mathcal{S}_s$ and $\mathcal{C}_c$ are combined to the privacy threat graph $\mathcal{T}_{s,c}$. This graph shows possible privacy threats of the system-to-be where the privacy of stakeholder s can be harmed by the counterstakeholder c. An example is shown in Fig. 4 in Section 3. Based on the problem diagrams and a given privacy requirement, all graphs are automatically generated by the ProPAn-tool. For more details see [4].

3 Domain Knowledge Extension of ProPAn

The extension of ProPAn for the consideration of domain knowledge consists of the steps *elicitation*, *modeling*, and *use*. Before these steps are presented in more detail, we introduce our running example that we use to illustrate our method.

Running Example We use a subsystem of 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)*[2] to il-

[2] `http://www.nessos-project.eu/`

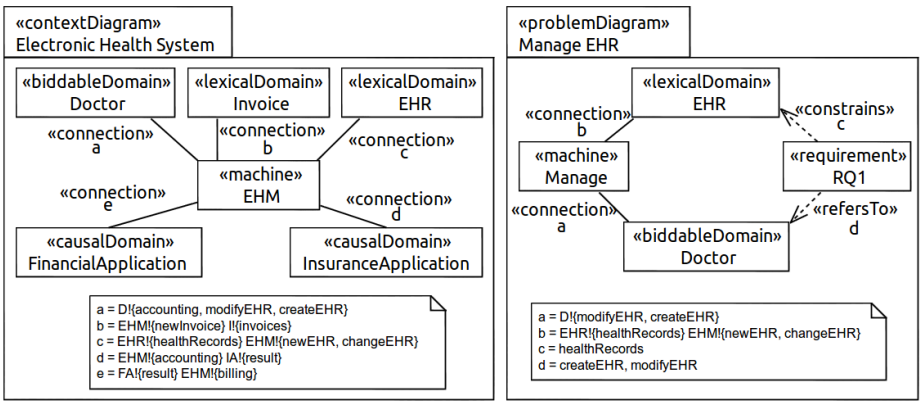

Fig. 1. Context diagram and problem diagrams of the EHS scenario

lustrate our method. This scenario is based on the German health care system which uses health insurance schemes for payment purposes. The context diagram of the EHS is shown in Fig. 1. The electronic health machine (EHM) is the machine to be built. It has to manage electronic health records (EHR) which are created and modified by doctors (functional requirement RQ1). Additionally, the EHS shall support the doctor to perform the accounting of patients. The accounting shall be based on the information stored in the health records. Using the insurance application it is possible to perform the accounting with the respective insurance of the patient. In the case that the insurance company does not bear the (complete) treating costs, the EHS shall create an invoice for the patient (RQ2). Patients shall be invoiced using a financial application (RQ3). The problem diagram for the functional requirement RQ1 is shown in Fig. 1. The not dashed part of Fig. 3 is the generated global information flow graph for the considered subsystem of the EHS.

The two privacy requirements that come into mind are that the doctor's privacy shall be protected against the financial and insurance applications. A privacy requirement about the patient cannot be expressed yet because the patient is not part of the problem frames model. We assume that the requirements engineer decided to leave out the patient because the patient does not directly interact with the machine.

Elicitation For the elicitation of the domain knowledge, we use questionnaires. All questions aim at the elicitation of indirect (counter)stakeholders or at the identification of hidden information flows in the considered system. Indirect (counter)stakeholders are (counter)stakeholders who are not considered yet because they are not directly part of the system-to-be. We distinguish two kinds of questions. Questions with the prefix 1 elicit counterstakeholders who can gain personal information from the domain. Questions with the prefix 2 elicit stakeholders of whom the domain provides personal information. We developed questionnaires for causal/lexical and biddable domains that refine these two question types. For the elicitation process, we have to consider all domains of the context diagram and to answer the corresponding questionnaires.

Table 1. Domain knowledge elicitation questionnaire for causal and lexical domains

No.	Question
1	**Elicitation of Counterstakeholders**
1.1	Is there a competitor that also uses the domain?
1.2	Could the domain be attacked by a hacker?
1.3	Does the domain provide information to legislators or law enforcement agencies?
1.4	Is the domain also used in other systems? State possible counterstakeholders that have access to the domain in these systems.
2	**Elicitation of Stakeholders**
2.1	Is the domain also used in other systems? State possible stakeholders of these systems from whom information is accessible through the domain.
2.2	Is initially personal information of stakeholders stored in the domain?
2.3	Does the domain store or process personal information of stakeholders directly, indirectly, or implicitly connected to it?

First, we consider all causal and lexical domains of the context diagram. The questionnaire for causal and lexical domains is shown in Table 1. We refined the first question type using the Volere stakeholder analysis template [2] which suggests the *negative stakeholders* competitor (question 1.1) and hacker (question 1.2). Furthermore, we refined question type 1 by asking for the *baseline stakeholder* legislator (question 1.3) suggested by [13]. Additionally, we added the possible counterstakeholder law enforcement agency to question 1.3. Competitors, hackers, legislators, and law enforcement agencies are all possible indirect counterstakeholders that are usually not considered as direct stakeholders of a software system. Questions 1.4 and 2.1 elicit (counter)stakeholders that can gain or provide personal information due to a re-use of a domain. Previous privacy analyses of the domain in other systems can be re-used to answer these questions. Especially lexical domains can already be filled with personal information, e.g. an existing database with contact information of customers. Question 2.2 elicits the stakeholders of this personal information. Systems may contain hidden information flows, i.e. storage or processing of personal information of stakeholders that are directly, indirectly, or implicitly connected to the domain. Question 2.3 elicits to which stakeholders a hidden information flow exists from the domain.

In the EHS example, we have to consider these questions for the domains EHR, insurance application, and financial application. Because of space limitations, we do not consider the insurance application for the domain knowledge elicitation in this paper. For the EHR, we have no indirect stakeholders that have access to the domain because this domain shall only be accessible using the EHS. But we have the indirect stakeholder *patient* because the EHRs may initially contain personal information about patients (question 2.2). The financial application (FA) is part of other systems and we identify its employees and customers as indirect (counter)stakeholders, due to questions 1.4 and 2.1. Furthermore, the application is considered to be a possible source of hacker attacks (question 1.2). The list of all (counter)stakeholders that we elicited for the FA using the questionnaire can be found in Table 4 in Section 4.

Second, we consider all biddable domains of the context diagram. The questionnaire for biddable domains is shown in Table 2. Question 1.1 aims at the trustworthiness of

Table 2. Domain knowledge elicitation questionnaire for biddable domains

No.	Question
1	**Elicitation of Counterstakeholders**
1.1	Is the domain vulnerable to social engineering attacks?
1.2	Does the biddable domain provide information to another biddable domain?
1.3	Does the biddable domain provide information to legislators or law enforcement agencies?
2	**Elicitation of Stakeholders**
2.1	Does the biddable domain get information of another biddable domain?
2.2	Does the biddable domain act on behalf of customers or wards (e.g. children)?

Table 3. Excerpt of the answer template for the EHS scenario

Dom ╲ Question	**1.2**	**1.4**	**2.1**	**2.2**
EHR	-	-	-	Patient
FA	Hacker	Employee, Customer	Employee, Customer	-
Doctor	Family of Patient	/	Family of Patient	Patient

a biddable domain in the system-to-be. With this question, we want to identify indirect stakeholders with whom the biddable domain possibly shares information that comes out of the system-to-be. Hence, question 1.1 elicits the source of so-called *social engineering attacks*. Questions 1.2 and 2.1 elicit implicit communications between biddable domains in the system. Question 1.3 is the same as question 1.3 from the previous questionnaire. Question 2.2 elicits those indirect stakeholders for whom a direct stakeholder acts on behalf of. These indirect stakeholders are of high relevance for the privacy analysis because personal information of those indirect stakeholders is stored and processed in the system-to-be in all likelihood.

In the EHS example, we only have the doctor as biddable domain. Hence, questions 1.2 and 2.1 are not relevant. For simplicity reasons, we do not consider question 1.1 because doctors are bound to professional discretion and hence we have no indirect counterstakeholders to which they provide information. It would be possible to consider corrupted doctors who break their professional discretion, and to elicit the stakeholders to whom they possibly provide information. All those stakeholders that we would consider as relevant are listed in Table 4 in Section 4. When we consider question 2.2, we identify the patients of the doctor, which can be seen as both customers and wards (i.e. persons who are being cared for by other persons). Doctors create health records on behalf of the patients whose personal information is stored in the records. The answers of the questionnaires are summarized in answer templates as shown in Table 3

It is reasonable to extend both questionnaires with questions specific to an application domain to give further assistance for the elicitation process. The questionnaires are easily extensible with questions aiming at the elicitation of privacy-relevant counterstakeholders (question type 1) and stakeholders (question type 2).

Modeling We use *domain knowledge diagrams* to model the domain knowledge, elicited using the questionnaires. Domain knowledge diagrams are already part of the

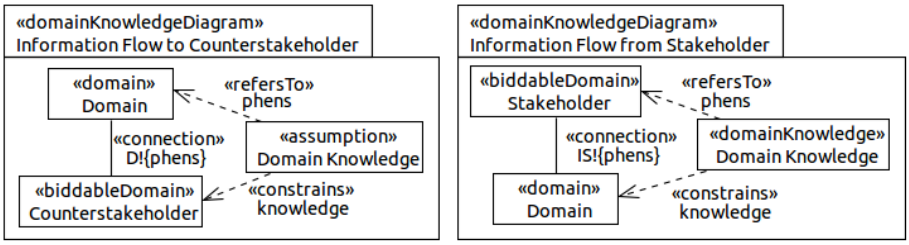

Fig. 2. General pattern for a domain knowledge diagram with an information flow from (left) / to (right) the indirect (counter)stakeholder

UML4PF-profile, for the representation of indicative statements. A domain knowledge diagram consists of an indicative statement, which is represented by the stereotype *DomainKnowledge*, and the domains referred to or constrained by the domain knowledge. The stereotype *DomainKnowledge* is specialized into the stereotypes *Fact* and *Assumption*. Facts are statements that are always true and assumptions only hold under specific circumstances. In general, we have presented two kinds of elicited domain knowledge. First, we elicit the indirect counterstakeholders who can gain information from the system-to-be (question type 1). Second, we elicit the indirect stakeholders of whom personal information is possibly stored and processed in the system-to-be (question type 2). The identified indirect (counter)stakeholders are modeled as biddable domains in all above cases. To represent the first kind of domain knowledge, we create a domain knowledge diagram by instantiating the pattern shown in Fig. 2 on the left-hand side. In the other case, we instantiate the pattern shown in Fig. 2 on the right-hand side. The domain *Domain* will be instantiated with the domain for which we answered the question, the *(Counter)Stakeholder* with the newly identified biddable domain, the title of the domain knowledge diagram and the domain knowledge with an appropriate name. Additionally, we have to decide whether the domain knowledge is a fact (a truth that always holds) or an assumption (a statement that could also be false under some circumstances). A dependency with the stereotype *refersTo* starting from the domain knowledge points to the source of the information flow and a dependency with the stereotype *constrains* starting from the domain knowledge points to the target of the information flow.

In the EHS example, we instantiate the patterns according to the answer template in Table 3. For example, the first variant is instantiated for the domain *Financial Application* with the indirect counterstakeholder *Hacker*. The second variant is instantiated for the domains *EHR* and *Doctor* with the indirect stakeholder *Patient*. Due to space limitations, we do not show the instantiated domain knowledge diagrams.

Use The domain knowledge can now be used for the generation of the information flow graphs and the counterstakeholder graph. In a domain knowledge diagram d, we have a possible information flow from the referred domain r to the constrained domain c, analogous to the possible information flows stemming from the problem diagrams [4]. Hence, we allow edges annotated with domain knowledge diagrams in the information

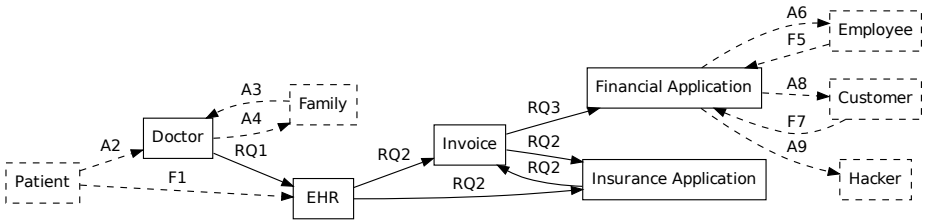

Fig. 3. Global information flow graph for the EHS with indirect stakeholders

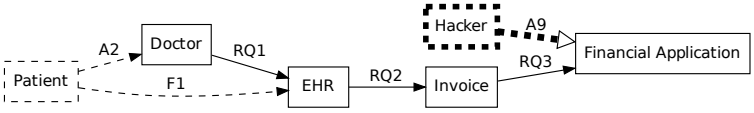

Fig. 4. Privacy threat graph for patient and hacker

flow graphs and add the edges (r, d, c) : Domain $\times$ ProblemDiagram $\times$ Domain for all domain knowledge diagrams to the global information flow graph. If a counterstakeholder c is constrained in a domain knowledge diagram d, then the counterstakeholder has possibly access to the information of the referred domain r. Hence, we add the edge (c, d, r) to c's counterstakeholder graph $\mathcal{C}_c$. With this additional rules for the graph generation, the privacy threat graphs now also provide possible threats that stem from indirect (counter)stakeholders.

In our example, it is now possible to consider the patient as a stakeholder whose privacy we want to protect. Additionally, hackers, as well as employees and customers of the financial application, can now be considered as counterstakeholders. From this small example, we can see that the elicitation of further domain knowledge is essential for a useful privacy analysis. The new global information flow graph is shown in Fig. 3 and the threat graph for the stakeholder patient and the counterstakeholder hacker in Fig. 4. The elements added by the proposed extension of ProPAn are drawn dashed.

Tool-Support We extended the ProPAn-tool[1] for the consideration of domain knowledge as described in this paper. The extension consists of a wizard that asks the user the questions of the questionnaires for all domains of the context diagram and directly creates the needed domain knowledge diagrams based on the domain knowledge patterns. Furthermore, the graph generation algorithms are extended such that they also consider the elicited (counter)stakeholders captured in domain knowledge diagrams.

4 Empirical Evaluation

We evaluated our questionnaires for the elicitation of indirect (counter)stakeholders during the presentation of this paper at the summer school. After the introduction of ProPAn

[1] available at `http://www.uni-due.de/swe/propan.shtml`

Table 4. Summarized results of the evaluation (privacy-relevant cells are printed in **bold** font)

	Doctor						Fin. App.								
Indirect Stakeholder	**S**			**C**			**S**			**C**			**Sum**		
Insurance companies	**0**	**4**	**4**	**8**	**9**	**17**	0	1	1	**1**	**6**	**7**	9	20	29
Patients	**5**	**8**	**13**	**1**	**6**	**7**	**0**	**5**	**5**	0	0	0	6	19	25
Other doctors	**4**	**5**	**9**	**5**	**6**	**11**	0	0	0	0	2	2	9	13	22
Nurses and staff	**3**	**3**	**6**	**5**	**6**	**11**	0	0	0	0	0	0	8	9	17
Pharmacy companies	**2**	**2**	**4**	**3**	**6**	**9**	0	1	1	**1**	**1**	**2**	6	10	16
Government and politicians	0	0	0	**5**	**2**	**7**	0	0	0	**1**	**6**	**7**	6	8	14
Family of patients	**4**	**2**	**6**	**0**	**4**	**4**	0	0	0	0	0	0	4	6	10
Hacker	0	0	0	0	0	0	0	0	0	**0**	**7**	**7**	0	7	7
Law enforcement agencies	0	1	1	**0**	**2**	**2**	0	0	0	**1**	**3**	**4**	1	6	7
Financial companies	0	2	2	**2**	**1**	**3**	**0**	**0**	**0**	**0**	**0**	**0**	2	3	5
Provider of financial app	0	0	0	0	0	0	**1**	**0**	**1**	**3**	**1**	**4**	4	1	5
Friends and family of doctor	0	0	0	**0**	**4**	**4**	0	0	0	0	0	0	0	4	4
Journalist	0	0	0	**1**	**2**	**3**	0	0	0	**0**	**1**	**1**	1	3	4
Researchers	0	1	1	**1**	**2**	**3**	0	0	0	0	0	0	1	3	4
Customers of financial app	0	0	0	0	0	0	**0**	**2**	**2**	**1**	**0**	**1**	1	2	3
Doctor	0	0	0	0	0	0	**0**	**2**	**2**	**0**	**1**	**1**	0	3	3
Employees of financial app	0	0	0	0	0	0	**0**	**3**	**3**	**0**	**0**	**0**	0	3	3
Social engineering attacker	0	0	0	**0**	**3**	**3**	0	0	0	0	0	0	0	3	3
Competitor of financial app	0	0	0	0	0	0	0	0	0	**0**	**2**	**2**	0	2	2
Employers	0	0	0	**1**	**1**	**2**	0	0	0	**0**	**0**	**0**	1	1	2

and the running example, the audience of the presentation was randomly split into two groups. Both groups had 10 minutes time to identify indirect (counter)stakeholders for the doctor and the financial application of the running example. One group had to guess indirect stakeholders without assistance and the other group used the developed questionnaires for the elicitation. The questionnaires used in the experiment can be found in the Appendix. There were 12 participants in the control group (without assistance) and 15 in the questionnaire group.

We consider twenty indirect (counter)stakeholder from the overall amount of thirty indirect (counter)stakeholders identified by the participants of the experiment as relevant. These are listed in the first column of Table 4. The following columns contain three numbers and show how often the indirect stakeholder was identified as stakeholder (S) or counterstakeholder (C) for the doctor and the financial application, respectively. The first number in these columns indicates how often the indirect stakeholder was considered by the control group, the second how often by the questionnaire group and the third gives the total amount of considerations. We printed a cell of the table in bold font if we consider the indirect stakeholder of the row as a relevant stakeholder (S) or counterstakeholder (C) for the domain in the column in a privacy analysis.

Based on the relationships that we consider as relevant, we computed the average precision, specificity, accuracy, and recall of both groups shown in Table 5. The precision and specificity of both groups is above 90%. The questionnaire group identified a few more unexpected indirect (counter)stakeholder relationships (false positives) than the control group, which leads to a smaller precision and specificity. The ques-

Table 5. Precision, specificity, accuracy, and recall of both groups

Group	Precision	Specificity	Accuracy	Recall
control group	93,39%	97,91%	41,67%	12,82%
questionnaire group	90,65%	93,33%	44,97%	20,17%

tionnaire group has a slightly larger accuracy than the control group. The recall of both groups lies below 20%, which is surely caused by the limited time of 10 minutes the participants had for the elicitation. Nevertheless, the questionnaire group identified 1,5 times more correct indirect (counter)stakeholder relationships. In summary, the questionnaires seem to help to increase the number of correct identified indirect (counter)stakeholders and their relationships to the domains of the context diagram significantly. The trade-off of the questionnaires is that the precision and specificity is slightly decreased. But this is reasonable because our main focus is the elicitation of all relevant indirect (counter)stakeholders for the privacy analysis.

5 Related Work

In this section, we discuss privacy-aware requirements engineering and stakeholder analysis methods that are related to this work.

Privacy-Aware Requirements Engineering The LINDDUN-framework proposed by Deng et al. [6] is an extension of Microsoft's security analysis framework STRIDE [7]. LINDDUN adds the seven privacy threats linkability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and policy/consent noncompliance to STRIDE. In contrast to ProPAn, the system to be analyzed is modeled as a data flow diagram (DFD), which has to be set up carefully for the analysis. ProPAn is based on a problem frames model which is assumed to be already existing and which can systematically be created using the problem frames approach [8]. Additionally, LINDDUN has to be carried out manually.

The PriS method introduced by Kalloniatis et al. [9] considers privacy requirements as organizational goals. The impact of the privacy requirements on the other organizational goals and their related business processes is analyzed. The authors use privacy process patterns to suggest a set of privacy enhancing technologies (PETs) to realize the privacy requirements. PriS is a goal-based approach, whereas ProPAn is problem-based. In addition, the PriS method has to be carried out manually.

Liu et al. [10] propose a security and privacy requirements analysis based on the goal and agent-based requirements engineering approach i^* [15]. The authors integrate the security and privacy analysis into the elicitation process of i^*. Already elicited actors from i^* are considered as attackers. Additional skills and malicious intents of the attackers are combined with the capabilities and interests of the actors. Then the vulnerabilities implied by the identified attackers and their malicious intentions are investigated in the i^* model. In contrast to our work, the approach of Liu et al. is goal based and it does not elicit additional privacy-relevant stakeholders for the analysis.

Stakeholder Analysis Stakeholder analysis originates from information systems research [11]. We describe the research of requirements engineers on this field.

Sharp et al. [13] present a method for the identification of stakeholders for requirements engineering. The authors distinguish four groups of *baseline stakeholders*, namely users, developers, legislators, and decision-makers. For each baseline role, the method identifies *supplier stakeholders* who provide information, *client stakeholders* who process or inspect the products, and *satellite stakeholders* who interact or support the baseline stakeholders and vice versa.

Alexander and Robertson [2] recommend a combination of two methods. The first method is the onion model [1] for the identification of stakeholders. The model arranges different generic stakeholder roles around the product, which is the center of the onion. The distance of a stakeholder to the product expresses how directly the stakeholder interacts with the product. The second method is the usage of the Volere stakeholder analysis template [12]. This template suggests 72 stakeholder roles that are divided into 14 stakeholder classes which again are divided into 4 categories of stakeholder classes. The template shall elicit stakeholders that hold relevant knowledge for the project.

The stakeholder analysis approaches all aim at the identification of stakeholders that are relevant to successfully complete a project. In contrast, we are interested in those stakeholders whose privacy is affected or those counterstakeholders that can harm the privacy of stakeholders in the system-to-be and not at the time of development.

6 Conclusion and Future Work

In this paper, we extended ProPAn with a structured method for the consideration of privacy-relevant domain knowledge. Three steps are necessary for the extension of ProPAn. First, we elicit the relevant domain knowledge based on questionnaires. Second, we introduce two generic patterns that can be instantiated to represent the elicited domain knowledge in the UML model. Third, we extend the definition of the global information flow graph and the counterstakeholder graph such that the domain knowledge is also considered in the privacy threat graphs. To support our method, we extended the ProPAn-tool[1] with the presented questionnaires from which the corresponding domain knowledge diagrams are generated.

The proposed questionnaires can easily be extended to provide better support for the elicitation of privacy-relevant indirect (counter)stakeholders. Our extension improves the expressiveness of the ProPAn-method because ProPAn can now consider both indicative and optative statements for the privacy analysis. Despite the fact that the amount of 27 participants of the empirical evaluation at the summer school is not representative and that the participants were no experts in the health care domain, it yields promising results. The evaluation shows that the questionnaires help to significantly increase the amount of identified privacy-relevant indirect (counter)stakeholders, while the amount of those who are not relevant is slightly increased. The questionnaires itself will generally not lead to a comprehensive and correct list of indirect (counter)stakeholders, but they give guidance for the elicitation process that has to be performed by requirements engineers in cooperation with domain experts.

[1] available at `http://www.uni-due.de/swe/propan.shtml`

As future work, we want to investigate how to prioritize privacy threats by the risk they cause. This prioritization can then be used to narrow down the amount of threats that has to be considered for development. Furthermore, we plan to extend ProPAn with specific analyses for privacy requirements such as unlinkability, transparency, and intervenability. Furthermore, ProPAn shall be extended to bridge the gap between the problem and the solution space. Therefore, we want to suggest PETs that can be chosen to implement a specific privacy requirement. The work of Deng et al. [6], Kalloniatis et al. [9], and Antón et al. [3] will serve as a starting point for this work. The application of ProPAn and the extension presented in this paper to an industrial-size case study and further empirical evaluations are also part of our future work.

Acknowledgment We thank Maritta Heisel, Azadeh Alebrahim, Kristian Beckers, Stephan Faßbender, Denis Hatebur, Marit Hansen and the anonymous reviewers for their constructive and valuable comments on earlier versions of this paper.

References

1. I. F. Alexander. A taxonomy of stakeholders: Human roles in system development. *IJTHI*, 1(1):23–59, 2005.
2. I. F. Alexander and S. Robertson. Understanding project sociology by modeling stakeholders. *IEEE Software*, 21(1):23–27, 2004.
3. A. I. Antón, J. B. Earp, and A. Reese. Analyzing website privacy requirements using a privacy goal taxonomy. In *RE*, pages 23–31. IEEE Computer Society, 2002.
4. K. Beckers, S. Faßbender, M. Heisel, and R. Meis. A problem-based approach for computer aided privacy threat identification. In *APF 2012*, LNCS. Springer, 2012.
5. I. Côté, D. Hatebur, M. Heisel, and H. Schmidt. UML4PF – a tool for problem-oriented requirements analysis. In *Proceedings of RE*, pages 349–350. IEEE Computer Society, 2011.
6. M. Deng, K. Wuyts, R. Scandariato, B. Preneel, and W. Joosen. A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. *RE*, 2011.
7. M. Howard and S. Lipner. *The Security Development Lifecycle*. Microsoft Press, Redmond, WA, USA, 2006.
8. M. Jackson. *Problem Frames. Analyzing and structuring software development problems*. Addison-Wesley, 2001.
9. C. Kalloniatis, E. Kavakli, and S. Gritzalis. Addressing privacy requirements in system design: the PriS method. *Requir. Eng.*, 13:241–255, August 2008.
10. L. Liu, E. Yu, and J. Mylopoulos. Security and privacy requirements analysis within a social setting. In *Requirements Engineering Conference, 2003. Proceedings. 11th IEEE International*, pages 151–161, 2003.
11. A. Pouloudi. Aspects of the stakeholder concept and their implications for information systems development. In *HICSS*, 1999.
12. S. Robertson and J. Robertson. *Mastering the Requirements Process (2nd Edition)*. Addison-Wesley Professional, 2006.
13. H. Sharp, A. Finkelstein, and G. Galal. Stakeholder identification in the requirements engineering process. In *DEXA Workshop*, pages 387–391, 1999.
14. A. F. Westin. *Privacy and Freedom*. Atheneum, New York, 1967.
15. E. Yu. Towards modeling and reasoning support for early-phase requirements engineering. In *Proceedings of the 3rd IEEE International Symposium on Requirements Engineering*, RE '97, pages 226–235, Washington, DC, USA, 1997. IEEE Computer Society.

Appendix: Questionnaires of the Experiment

As mentioned in Section 4, we split the audience of the presentation of this paper during the summer school randomly into two groups. One group got the developed questionnaire shown in Fig. 5 and 6. The other group was the control group and got the questionnaire shown in Fig. 7.

June 20, 2013 — **Elicitation of Privacy-Relevant Domain Knowledge** — UNIVERSITÄT DUISBURG ESSEN *Open-Minded*

What is your expertise in requirements engineering?

☐ expert ☐ high ☐ medium ☐ low ☐ none

Questionnaire for the Doctor

1.1 Is the **doctor** vulnerable to social engineering attacks? State possible indirect counterstakeholders, who could perform a social engineering attack on the **doctor**.

1.2 Does the **doctor** provide information to another biddable domain of the system? State possible connections between the **doctor** and other biddable domains.

2.1 Does the **doctor** get information from another biddable domain of the system? State possible connections between the **doctor** and other biddable domains.

2.2 Does the **doctor** act on behalf of other people (e.g. customers, children)? State indirect stakeholders the **doctor** acts on behalf of.

Fig. 5. First page of the developed questionnaire

Questionnaire for the Financial Application

1.1 Is there a competitor that also uses the **financial application**? State possible indirect counterstakeholders, who also use the **financial application**.

1.2 Could the **financial application** be attacked by a hacker? State possible hackers that could attack the **financial application**.

1.3 Provides the **financial application** information to legislators? State possible legislators to which the **financial application** provides information.

1.4 Is the **financial application** also used in other systems? State possible counterstakeholders that have access to the **financial application** in these systems.

2.1 Is the **financial application** also used in other systems? State possible stakeholders of these systems from whom information is accessible through the **financial application**.

2.2 Is initially personal information of stakeholders stored in the **financial application**? State stakeholders from whom the **financial application** possibly provides personal information.

Fig. 6. Second page of the developed questionnaire

What is your expertise in requirements engineering?

☐ expert ☐ high ☐ medium ☐ low ☐ none

Identify indirect (counter)stakeholders and state their relation to the Doctor.

Identify indirect (counter)stakeholders and state their relation to the Financial Application.

Fig. 7. Control group questionnaire