Formal Modelling of (De)Pseudonymisation: A Case Study in Health Care Privacy

Meilof Veeningen, Benne de Weger, and Nicola Zannone

Eindhoven University of Technology, The Netherlands
{m.veeningen,b.m.m.d.weger,n.zannone}@tue.nl

Abstract. In recent years, a number of infrastructures have been proposed for the collection and distribution of medical data for research purposes. The design of such infrastructures is challenging: on the one hand, they should link patient data collected from different hospitals; on the other hand, they can only use anonymised data because of privacy regulations. In addition, they should allow data depseudonymisation in case research results provide information relevant for patients' health. The privacy analysis of such infrastructures can be seen as a problem of data minimisation. In this work, we introduce coalition graphs, a graphical representation of knowledge of personal information to study data minimisation. We show how this representation allows identification of privacy issues in existing infrastructures. To validate our approach, we use coalition graphs to formally analyse data minimisation in two (de)-pseudonymisation infrastructures proposed by the Parelsnoer initiative.

1 Introduction

The quality of medical research benefits from the collection of patient data from different health care organisations. By analysing data from different sources, researchers are able to study treatments from several angles, which can lead to new insights. To facilitate the collection and dissemination of medical data, several initiatives like the Dutch Parelsnoer initiative have developed data management infrastructures [11–13]. Such infrastructures store patient data collected from health care organisations into a central medical research database and then distribute such data to researchers. Besides providing data to researchers, they should also allow the sharing of findings about patients' conditions made by researchers to hospitals in order to provide treatments to patients.

When distributing patient data, these infrastructures should protect the patient's privacy by making sure that data are properly anonymised.In particular, researchers should not be able to link data to a particular patient, or data from different research projects to each other. However, it is not possible to just remove all identifiers from the data: the need to share findings with the patient's hospital implies that the data may need to be deanonymised. Thus, there is a need to supplement data management infrastructures for medical research with (de)pseudonymisation functionality. Depseudonymisation should be possible only following a rigurous process involving a coalition of several different

A. Jøsang, P. Samarati, and M. Petrocchi (Eds.): STM 2012, LNCS 7783, pp. 145–160, 2013.

parties; the infrastructure should technically ensure that other coalitions do not have anough knowledge to correlate patient data.

The problem of depseudonymisation exemplifies the broader concept of data minimisation [9], which is nowadays imposed by privacy regulations (e.g., EU Directive 95/46/EC, HIPAA). Data minimisation decreases both the risk of abuse by insiders [8] and the impact of information theft by outsiders. The concept is gaining relevance as the increase in personal information exchanged on-line has raised privacy issues not only about health care data, but also search data, data on shopping habits, etc. However, comprehensive analyses of data minimisation in these different settings are usually lacking. In discussions on (de)pseudonymisation of health care data and proposed infrastructures (e.g., [3, 10, 13]), the aim is usually to prevent knowledge of *particular* links by *particular* actors; however, a full analysis of data minimisation that considers *all* possible correlations allowed by a system and checks whether they are inherent to the setting or preventable, is usually missing.

In [19], we presented a formal model that, given the information exchanged between actors, analyses the knowledge of personal information that actors learn. This model makes it possible to verify privacy requirements by checking whether a particular coalition of actors can correlate particular pieces of information. However, it neither provides a general comparison of all knowledge of all coalitions in different infrastructures, nor discusses the concept of minimality.

In this work, we study data minimisation in the Parelsnoer infrastructure for health care data (de)pseudonymisation. We discuss general requirements for such infrastructures, but focus our analysis on Parelsnoer (we briefly discuss related proposals in Section 7). We introduce a novel formalism, called *coalition graphs*, to express the profiles of personal information that can be compiled by coalitions of actors within a system. We show how coalition graphs can be used to comprehensively model data minimisation, identify possible privacy improvements, and verify their effect. Specifically:

- We capture different privacy risks by considering actors that only store what they should store and actors that remember all information they observe;
- We show how coalition graphs can be derived automatically from a formal model describing actors' communication [19] ;
- By formalising requirements for distributing medical data for research, and privacy consequences of using a central database, we model the "optimal" situation in terms of data minimisation achievable by (de)pseudonymisation infrastructures;
- Using coalition graphs, we analyse two (de)pseudonymisation infrastructures proposed by the Parelsnoer initiative, and propose privacy improvements.

The paper is structured as follows. We first describe the setting of distributing patient data for research, and two infrastructures proposed by Parelsnoer (§2). We then introduce coalition graphs (§3). We derive an optimal coalition graph for the Parelsnoer setting (§4), use it to analyse data minimisation in the proposed infrastructures (§5), and analyse possible improvements (§6). Finally, we discuss related work and conclude by providing directions for future work (§7).

2 Pseudonymisation Infrastructure

In this section, we discuss the requirements for (de)pseudonymisation infrastructures for medical research databases. In particular, we consider the privacy requirements defined within the Dutch legal framework regarding the processing of medical data. We then present two (de)pseudonymisation infrastructures developed as a result of the Parelsnoer initiative (`http://www.parelsnoer.org/`): one based on hashing, and one based on a trusted "pseudonymisation service".

2.1 Setting

For medical research, data about a patient collected from different health care organisations have to be linked together into a single dataset. Data integration, however, is challenging because of the stringent constraints imposed by data protection regulations. We now discuss the functional **(FR)** and privacy **(PR)** requirements for the handling of medical data within the Dutch legal framework.

The Dutch legal framework constrains the identification of medical data. For treatment, health care organisations are obliged to use the "burgerservicenummer" (BSN: the Dutch social security number) **(FR1)**; for other purposes, they (and others) are forbidden to do so **(PR1)**. Medical data may be used for research **(FR2)** if anonymised so that association to the BSN or (indirectly) to the patient is impossible, and different projects' datasets cannot be linked **(PR2)**.

However, in certain circumstances, this anonymisation needs to be reversed. In case of a discovery beneficial for the patient (a so-called *coincidental finding*), the health care organisations which collected the data should be notified so they can provide treatment: *full depseudonymisation* **(FR3)**. Moreover, if additional patient data is needed for a certain research project, it should be possible to link together data about the same patient: *partial depseudonymisation* **(FR4)**.

To facilitate the provision of medical data to researchers, data collected from different health care organisations can be stored into a single database [11–13], hereafter called *Central Infrastructure* (CI). We describe the operation of systems with a CI by enumerating the main design decisions **(DD)** that cover the handling of personal information. The CI stores the data about one patient from different hospitals in one profile **(DD1)**. It obtains this data directly from the different hospitals **(DD2)**. When a researcher needs a dataset, the CI compiles it from its database and sends it to the researcher **(DD3)**, who is not otherwise involved in the pseudonymisation process. For extension of a dataset (i.e., partial depseudonymisation), the researcher contacts the CI, which then compiles the extended dataset without involving the original hospital **(DD4)**. Finally, depseudonymisation should be performed via a trusted third party to ensure that it is only possible under strictly defined conditions **(DD5)**.

2.2 Parelsnoer Initiative

This section presents two infrastructures for (de)pseudonymisation developed by the Parelsnoer initiative [11, 12]. This initiative is a collaboration between eight university medical centres in the Netherlands.

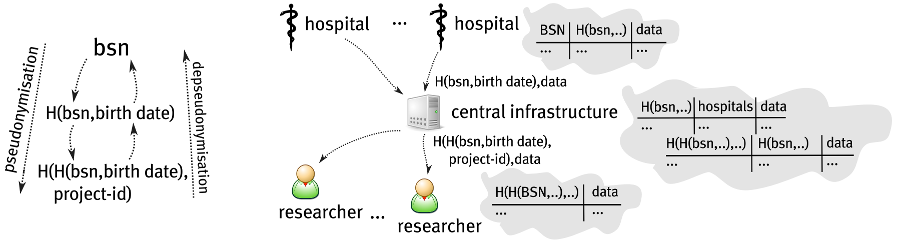


Fig. 1. Parelsnoer Hash-Based Pseudonymisation Infrastructure (H-PI): pseudonyms (left) and operation (right)

Hash-Based Pseudonymisation Infrastructure (H-PI). The first Parelsnoer proposal for (de)pseudonymisation [11] uses pseudonyms for the storage and transmission of medical data that are constructed using hash functions (Figure 1). In particular, when providing data to the CI, hospitals use a hash h_1 of a patient's BSN and birth date as pseudonym. This allows the CI to link data from different hospitals without learning the BSN. Each research project has a separate identifier; when the CI distributes data for a research project, the project identifier is hashed along with the pseudonym h_1 into a new pseudonym h_2. For partial depseudonymisation, the CI needs a table containing the links (h_1, h_2) for all distributed datasets. For full depseudonymisation, the CI additionally needs a table containing the identities of hospitals for all patient pseudonyms h_1. Each hospital stores a table containing the links (bsn, h_1) for its own patients.

One drawback of this approach is that an attacker who learns a pseudonym, can try to depseudonymise it using a dictionary attack: this is feasible because the entropy in the combination of BSN and birth date is at most 42 bits [11]. In addition, the fact that hospitals and CI need to keep pseudonym translation tables poses significant risks of data breaches. Note that H-PI does not use a TTP to control depseudonymisation; as shown later, this makes it non-optimal in terms of data minimisation.

Pseudonymisation Service Infrastructure (PS-PI). Parelsnoer's Pseudonymisation Service Infrastructure [12] addresses the limitations of the hash-based approach using a TTP called *pseudonymisation service* (PS). The pseudonyms used in the system are called "pseudocodes". These pseudocodes are unique given a BSN and a "domain" (i.e., the CI, hospitals, and research projects) in which patient data should be linked. The mapping between BSNs and pseudocodes and between pseudocodes from different domains is calculated using a domain-specific secret known only by the PS.

Figure 2 shows the translation steps (left) and the information that is exchanged and stored (right). First, the PS translates the BSN into a pseudocode in the hospital domain, which the hospital uses to send medical data to the CI. The CI requests the PS to re-translate the pseudocode to its own domain so it can link data from different hospitals together. When data are distributed to a researcher, the pseudocode is translated to the project domain. For depseudonymisation,

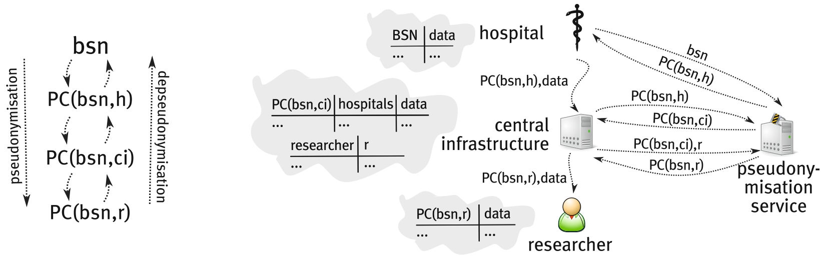


Fig. 2. Parelsnoer Infrastructure with Pseudonymisation Service (PS-PI): pseudonyms (left) and operation (right)

pseudocodes are translated back to the BSN in exactly the opposite order. For partial depseudonymisation, the researcher provides the research pseudocode to the CI, which requests the PS to translate it to its own domain. The CI remembers which research domain belongs to which researcher; it includes the research domain in the depseudonymisation request to the PS, which compares it to the actual domain within the pseudocode. For full depseudonymisation, the CI then requests the PS to translate the pseudocode from the CI domain to the hospital domain based on the list of hospitals that have provided data.

This infrastructure solves the drawbacks of the hash-based infrastructure. Since pseudocodes are calculated using a secret known only by the PS, this infrastructure is not subject to dictionary attacks. Moreover, the hospital and CI no longer store tables to translate pseudocodes to BSNs. Indeed, depseudonymisation is not possible without the PS, reducing privacy impact when data of hospitals and the CI are compromised.

2.3 Scenario

We introduce a scenario that is general enough to capture all aspects we are interested in yet small enough to allow a clear visualisation. We consider (coalitions of) six different actors: three hospitals umc_1, umc_2, and umc_3; one researcher r; and CI ci and TTP ttp (in PS-PI: the PS). These actors exchange information about a particular patient. Two of the three hospitals have medical data about the patient: umc_1 knows three pieces of information d_1, d_2, and d_3; umc_2 knows d_4, d_5, and d_6. The items d_i are non-identifying; i.e., they represent attributes for which different patients may have a common value. The hospitals identify their patient records by BSN bsn. The third hospital umc_3 does not know the patient. Researcher r needs data about the patient for two different research projects: d_1 and d_4 for one project, and d_2 and d_5 for a second project. By considering two hospitals with patient data and one without, we can consider correlation between these two types of hospitals and other actors, and between two different hospitals that both know the patient. Verifying privacy with respect to data of one single researcher from two different projects is sufficient: if she cannot link the data, then neither can two different researchers from two different projects.

Our scenario has three steps. First, umc_1 and umc_2 provide their patient data to ci. Second, r receives patient data from the ci in two different datasets for the two different projects. In both steps, the TTP may be involved. Third, as part of the investigation in the first research project, the researcher learns a coincidental finding d_7 that may be important for treatment of the patient. We consider the moment when the coincidental finding has been made, but depseudonymisation has not been performed yet. In particular, the hospitals do not know d_7 yet, so we can reason about coalitions that enable hospitals to link d_7 to the corresponding patient.

3 Coalition Graphs

In this section, we introduce coalition graphs as a graphical way of studying data minimisation. First, we phrase the data minimisation problem in terms of profiles derivable by coalitions of actors. We then introduce coalition graphs as a way to compare infrastructures. Finally, we show how to obtain coalition graphs automatically, and how to use them for data minimisation analysis.

3.1 Data Minimisation, Coalition Knowledge, and Forgetting

The data minimisation principle aims to prevent abuse of personal information by restricting the information an actor can collect to what is strictly necessary to carry out assigned duties. The adherence of actors to data minimisation can be analysed with respect to their behaviours. *Honest* actors store only the information the system allows them to store. However, actors may observe other information. We call actors who store all the information they observe *honest-but-curious.* As an example, the PS-PI architecture aims to ensure that depseudonymisation can only happen though the PS. However, this data minimisation goal can only be achieved when the other actors are honest: if hospitals and the CI are honest-but-curious, they can link data by remembering pseudocodes and thus bypass the PS. We analyse data minimisation with respect to arbitrary coalitions of honest and honest-but-curious actors, thus clarifying the assumptions under which privacy properties hold. Although honest and honest-but-curious actors differ in what they remember, they both only obtain the information that they are supposed to obtain. Privacy protection against actors who actively try to obtain information they should not know, is out of our scope (but could be captured by coalition graphs).

In [18], we proposed personal information models as a representation of actors' knowledge of personal information. A personal information model consists of items of interest (i.e., data items, identifiers, and entities) and linkability relations between them. *Data items* and *identifiers* are pieces of information that characterise an *entity.* Differently from data items, identifiers uniquely identify an entity (e.g., social security number). The set $\mathcal{O}$ denotes the items of interest, i.e., sensitive information to be protected. In our scenario, $\mathcal{O} = \{bsn, d_1, ..., d_7\}$, where bsn is an identifier and $d_1, \ldots, d_7$ are data items. This set does not contain pseudonyms because their knowledge in itself is not relevant; the fact that they can be used to link different pieces of data together is accounted for by the linkability relation. A *profile* is a set $O \subset \mathcal{O}$ of items of interest characterising the same entity; e.g., $\{bsn, d_1, d_2, d_3\}$ represents the patient's profile at umc_1.

Actors have partial knowledge of the personal information model. Set $\mathcal{A}$ denotes the set of actors involved in the system (in our scenario, $\mathcal{A} = \{umc_1, umc_2, umc_3, ci, ttp, r\}$). These actors are honest; we denote their honest-but-curious counterparts with a *, e.g., $ci^* \in \mathcal{A}^*$ is the honest-but-curious counterpart of ci. A *coalition* of actors is any subset of $\mathcal{A} \cup \mathcal{A}^*$ in which actors can be either honest or honest-but-curious. For instance, $\{umc_1, ci^*\}$ is a coalition formed by a honest umc_1 and a honest-but-curious ci. Coalition A can be *extended* to coalition B, denoted $A \sqsubseteq B$, if any honest actor in A is also in B (either honest-but-curious or honest), and any honest-but-curious actor in A is also honest-but-curious in B. For instance, $\{umc_1\} \sqsubseteq \{umc_1, ci\} \sqsubseteq \{umc_1, ci^*\}$ but $\{umc_1{}^*\} \not\sqsubseteq \{umc_1, ci^*\}$.

The knowledge of coalitions is captured by the *profile detectability* relation $\vDash$. Given a coalition A and a set $O \subset \mathcal{O}$ of pieces of information, $A \vDash O$ expresses that coalition A knows: 1) the items in O (detectability); and 2) the fact that the items in O are about one single person (linkability). For instance, $\{umc_1\} \vDash \{bsn, d_1, d_2, d_3\}$ indicates that umc_1 knows bsn, d_1, d_2, and d_3 and it knows that these items of interest belong to the same patient. Similarly, $\{r\} \nvDash \{d_1, d_2\}$ indicates that r is not able to link d_1 and d_2. Profile detectability $\vDash$ satisfies two properties: 1) if $A \vDash O$ and $A \sqsubseteq B$, then $B \vDash O$; and 2) if $A \vDash O$ and $P \subseteq O$, then $A \vDash P$. We say that $A \vDash O$ *implies* $B \vDash P$ if $A \sqsubseteq B$ and $P \subseteq O$.

3.2 Coalition Graphs, Comparison, and Reduction

The information known by coalitions of actors can be visually represented as a directed graph. Nodes are pairs (A, O) such that $A \vDash O$. Edges are defined by the partial order $\leq$ on nodes that combines the partial orders on coalitions and profiles. $(A_1, O_1) \leq (A_2, O_2)$ expresses that coalition A_2 is an extension of coalition A_1, and profile O_2 is a superset of profile O_1. Formally:

Definition 1. *The* coalition graph *for relation* $\vDash$ *is the graph* $(V, \leq)$ *with:*

- $V = \{(A, O) \mid A \sqsubseteq \mathcal{A}^*;\ O \subset \mathcal{O};\ A \vDash O\}$
- $(A_1, O_1) \leq (A_2, O_2)$ *iff* $A_1 \sqsubseteq A_2 \wedge O_1 \subseteq O_2$.

Infrastructures can be compared wrt data minimisation by means of their coalition graph. If in infrastructure X every coalition can derive at least the same information that it can in infrastructure Y, then the coalition graph of X includes all nodes of the coalition graph of Y. Moreover, $\leq$ is defined independently from $\vDash$, so if the two coalition graphs share two nodes and these nodes are connected in one graph, then they are also connected in the second graph. Based on these observations, we introduce the notion of achieving better privacy.

Definition 2. *Let X and Y be two infrastructures with coalition graphs G_X, G_Y, respectively. We say that X* achieves (strictly) better privacy *than Y if G_X is a (proper) subgraph of G_Y.*

For visualisation purposes, we introduce the *reduced coalition graph* $(V', \preceq)$ of a coalition graph $(V, \leq)$. In reduced graphs, redundant information from coalition graphs is eliminated: V' is the set of minimal nodes in V, i.e., nodes that are not implied by any other nodes in V; $\preceq$ is the non-reflexive, transitive reduction of $\leq$ on V'. The reduced coalition graph of an infrastructure can be determined automatically by enumerating the knowledge of all coalitions.

To verify whether infrastructure X achieves better privacy than infrastructure Y, we compare their reduced coalition graphs $(V_X, \preceq_X)$, $(V_Y, \preceq_Y)$. However, this comparison cannot be done by checking whether $(V_X, \preceq_X)$ is a subgraph of $(V_Y, \preceq_Y)$. This is because nodes that are minimal in one graph may not be minimal in the other graph. For instance, suppose $(\{a, b\}, \{d_1, d_2\}) \in V_X$ and $(\{a\}, \{d_1, d_2\}) \in V_Y$. In such a case, $(\{a, b\}, \{d_1, d_2\}) \notin V_Y$ because the node is not minimal in Y: it is implied by $(\{a\}, \{d_1, d_2\})$. Instead, in order to compare two infrastructure, we visualise their reduced coalition graphs in a single graph. The nodes of the new graph are $V_X \cup V_Y$; for each node, we indicate if it is implied in X, Y, or both. Edges are the non-reflexive, transitive reduction of $\leq$ on $V_X \cup V_Y$. Infrastructure X then satisfies better privacy than infrastructure Y if all nodes in $V_X \cup V_Y$ are implied by the nodes in V_Y.

3.3 Studying Data Minimisation by Coalition Graphs

Data minimisation analysis using coalition graphs is performed as follows. First, requirements and design decisions are formalised and represented in an "optimal" coalition graph. Then, an infrastructure is analysed through an iterative process: 1) determine the coalition graph of the infrastructure; 2) compare this graph to the optimal graph to detect design drawbacks; 3) propose enhancements.

The optimal graph is based on functional requirements and design decisions that specify the information that actors should know. They are modelled as profile detectability statements $A \vDash O$; the statement $A \vDash O$ and any statement it implies hold in all infrastructures providing the required functionality. Privacy requirements state that certain actors should *not* know certain information. These are formalised by profile undetectability statements $A \nvDash O$; the statement $A \vDash O$ and any statement implying it should not hold in well-designed infrastructures. Section 4.2 shows how to determine the optimal graph from $\vDash$.

The coalition graph of the system is computed using the formal analysis method in [19]. Given a description of initial knowledge and communication, the method determines which copies of items of interest (coalitions of) actors can detect, and which items they can link. Intuitively, actors can link items through identifiers (e.g., BSNs, pseudocodes, and session identifiers). Profile detectability $\vDash$ holds if there are detectable and mutually linkable copies of all items in the profile. We developed a tool (see `http://www.mobiman.me/downloads/`) that automatically generates the coalition graph by running the implementation of [19] on all coalitions, eliminating implied nodes, and visualising using GraphViz. The method in [19] only considers honest-but-curious actors. To represent honest actors, we have extended it by introducing a Store operation that describes what information actors should store. Intuitively, a data item is added to the

Table 1. Privacy consequences of functional **(FR)** and privacy **(PR)** requirements and design decisions **(DD)**

Requirement/Decision	**Privacy consequences**
(FR1) Hospitals store data using BSN	$\{umc_1\} \models \{bsn, d_1, d_2, d_3\}$, $\{umc_2\} \models \{bsn, d_4, d_5, d_6\}$
(FR2) Researchers obtain dataset	$\{r\} \models \{d_1, d_4, d_7\}$, $\{r\} \models \{d_2, d_5\}$
(FR3) Full depseudonymisation	$\{umc_1, ci, ttp, r\} \models \{bsn, d_7\}$, $\{umc_2, ci, ttp, r\} \models \{bsn, d_7\}$
(FR4) Partial depseudonymisation	$\{umc_1, ci, ttp, r\} \models \{d_1, d_2, d_3, d_7\}$, $\{umc_2, ci, ttp, r\} \models \{d_4, d_5, d_6, d_7\}$
(PR1) BSN not for research purposes	$\{r^*, ci^*, ttp^*\} \not\models \{bsn\}$
(PR2) Researcher cannot link datasets	$\{r^*\} \not\models \{d_1, d_2\}$, $\{r^*\} \not\models \{d_1, d_5\}$, $\{r^*\} \not\models \{d_2, d_4\}$, $\{r^*\} \not\models \{d_4, d_5\}$
(DD1) CI collects data	$\{ci\} \models \{d_1, d_2, d_3, d_4, d_5, d_6\}$ $\{umc_1, ci, ttp\} \models \{bsn, d_1, d_2, d_3, d_4, d_5, d_6\}$ $\{umc_2, ci, ttp\} \models \{bsn, d_1, d_2, d_3, d_4, d_5, d_6\}$
(DD2) Data transfer between UMC, CI	$\{{umc_1}^*, ci^*\} \models \{bsn, d_1, d_2, d_3, d_4, d_5, d_6\}$, $\{{umc_2}^*, ci^*\} \models \{bsn, d_1, d_2, d_3, d_4, d_5, d_6\}$
(DD3) Dataset from CI to researcher	$\{ci^*, r\} \models \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$
(DD4) Partial depseudo w/o hospital	$\{ci, ttp, r\} \models \{d_1, d_2, d_3, d_4, d_5, d_6, d_7\}$
(DD5) (De)pseudonymisation by TTP	(See consequences of **(FR)**, **(PR)**, **(DD)**)

knowledge base of an actor only if he is allowed to store it. For instance, in the model of PS-PI, Store does not store BSNs in the knowledge base of PS.

We compare the coalition graph to the optimal graph by visualising both in one picture. Non-optimal nodes highlight privacy drawbacks in the system design. The analysis of why these nodes exist may raise enhancements, which are then analysed to verify whether the drawbacks have been addressed.

4 Privacy-Optimal Graph

In this section, we analyse the optimal privacy achievable in (de)pseudonymisation infrastructures for medical research databases. An "optimal" coalition graph formalises the privacy consequences of functional requirements and design decisions. We also formalise privacy requirements defining the information a given actor should not know.

4.1 Formalisation of Requirements and Design Decisions

Table 1 formalises the privacy consequences of the functional requirements, privacy requirements, and design decisions described in Section 2.1. Actors' knowledge is taken after the CI has distributed the datasets to the researcher and she has made a coincidental finding, but before depseudonymisation has taken place.

Functional requirements **(FR1)** and **(FR2)** directly translate to the fact that hospitals and researchers know certain data about the patient. Functional requirements **(FR3)** and **(FR4)** state that full/partial depseudonymisation should be possible. In particular, a hospital, the TTP, the CI, and the researcher together should be able to perform full depseudonymisation, i.e., they should be able to link d_7 to bsn. Similarly, for partial depseudonymisation, they should be able to link d_7 to the patient data.

Privacy requirement **(PR1)** states that BSNs cannot be used for research purposes. Thus, even if the CI, TTP, and researcher are curious and combine their knowledge, they should not be able to derive the patient's BSN. For **(PR2)**, the researcher should not be able to link any information from his first dataset to any information from his second dataset, even if he is curious.

Introducing the medical research database CI has several privacy consequences. Design decision **(DD1)** states that the task of the CI is to collect and link the data from different hospitals; it has two consequences. First, the CI knows the medical data from the two hospitals in one profile. Second, if a hospital, CI and TTP combine their knowledge, they can link the BSN to the full patient record at the CI (by definition of the collection process). By design decision **(DD2)**, we consider systems where the CI and UMC communicate directly during the collection process. At the time of this communication, the hospital knew the BSN, and the CI knew the link to the full patient record. Therefore, if both have remembered some details of the communication such as the session identifier (i.e., they were curious), they can link the BSN to the full patient record without the PS. Design decision **(DD3)** states that the researcher is involved in (de)pseudonymisation merely as the passive recipient of the datasets. During the provision of such a dataset, the CI knew the link between records in the distributed dataset and the full patient records. If the CI is curious and remembers this link, and the researcher discovers an accidental finding related to some record, then together they can link the finding to the record. Design decision **(DD4)** states that hospitals are not involved in partial depseudonymisation; instead, it is performed by linking the incidental finding of the researcher to the patient record at the CI using the TTP. Finally, design decision **(DD5)** is the introduction of the TTP. This design decision is reflected by the fact that TTP is needed for data collection **(DD1)** and full **(FR3)** and partial **(FR4)**, **(DD4)** depseudonymisation, as well as by the fact that the TTP is introduced for research purposes and therefore should not know BSNs **(PR1)**.

4.2 Privacy-Optimal Graph

Figure 3 combines the privacy consequences in Table 1 into a coalition graph. Intuitively, it is the coalition graph of a hypothetical infrastructure O-PI which satisfies all requirements and design decisions, and whose design is optimal in terms of data minimisation. Nodes represent unavoidable disclosures.

The graph is obtained from Table 1 by considering which consequences apply to any particular coalition. Given a coalition A, we consider which profile detectability statements $A \vDash O$ are implied by the entries in the table. For instance, for coalition $A = \{umc_1\}$, the table implies detectability of profile $\{bsn, d_1, d_2, d_3\}$, which corresponds to a node in the graph. Coalition $A = \{r\}$ can detect two profiles $\{d_1, d_4, d_7\}$ and $\{d_2, d_5\}$ but it should not be able to

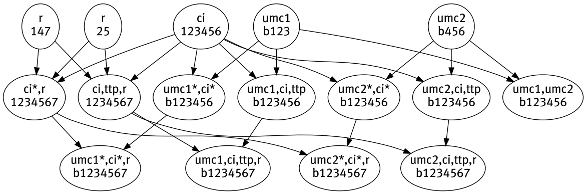


Fig. 3. Reduced coalition graph in optimal situation (O-PI). Node captions represent coalitions A and profiles O, respectively, with $A \vDash O$; 'b' means bsn, '1' means d_1, etc.

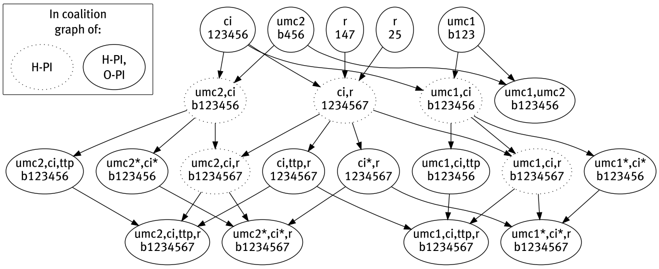


Fig. 4. Comparison between reduced coalition graphs of Parelsnoer hash-based pseudonymisation infrastructure (H-PI) and optimal situation (O-PI)

link them together, so the two profiles occur as two nodes in the graph. On the other hand, for coalition $A = \{umc_1, umc_2\}$, $A \vDash \{bsn, d_1, d_2, d_3\}$ is implied by $\{umc_1\} \vDash \{bsn, d_1, d_2, d_3\}$, and $A \vDash \{bsn, d_4, d_5, d_6\}$ is implied by $\{umc_2\} \vDash \{bsn, d_4, d_5, d_6\}$. These two profiles can be linked together because they both contain the BSN; therefore, they are represented by node $A \vDash \{bsn, d_1, \ldots, d_6\}$. Informally, coalitions of honest actors can link profiles if they have stored a shared identifier; coalitions of honest-but-curious actors can additionally link profiles if they have exchanged personal information from the profiles. These conclusions can be formally derived using the method in [19].

The reduced coalition graph of the optimal situation O-PI makes it possible to assess the extent to which existing infrastructures satisfy data minimisation. Namely, we can compare the reduced coalition graph of an existing infrastructure to the reduced coalition graph of O-PI, as described in Section 3. If the two graphs are the same, the infrastructure achieves optimal privacy. Otherwise, the privacy issues of the analysed infrastructure can be identified by analysing non-optimal nodes in the graph.

5 Coalition Graphs for Parelsnoer Infrastructures

In this section, we analyse data minimisation in the Parelsnoer infrastructures by comparing their reduced coalition graphs to the optimal one.

Hash-Based Infrastructure. Figure 4 compares the reduced coalition graph of the hash-based Parelsnoer infrastructure (H-PI) to the optimal situation (O-PI). The dotted nodes represent nodes that only occur in H-PI's coalition graph and thus point to violations of data minimisation. The solid nodes are also in O-PI's coalition graph and thus unavoidable. (H-PI does not use the TTP; it occurs in this graph because we compare it to the optimal situation, in which the TTP is needed for (de)pseudonymisation.)

The non-optimal nodes can be explained by the use of translation tables for depseudonymisation, as opposed to using the services of the TTP. Hospitals need to remember the pseudocode sent to the CI for full depseudonymisation, which implies $\{umc_i, ci\} \vDash \{bsn, d_1, \ldots, d_6\}$. The CI needs to remember the pseudocode sent to the researcher, implying $\{ci, r\} \vDash \{d_1, \ldots, d_7\}$. Combining the translation tables gives $\{umc_i, ci, r\} \vDash \{bsn, d_1, \ldots, d_7\}$.

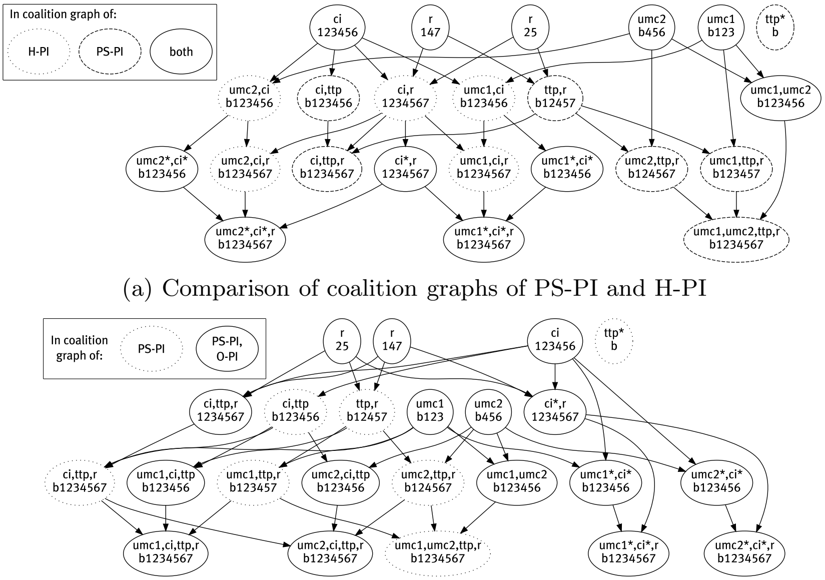
The privacy difference between using translation tables and using a TTP can be observed from the coalition extensions needed to make non-optimal nodes optimal. First, for any non-optimal node $A \vDash O$, node $A \cup \{ttp\} \vDash O$ is optimal. This expresses that actors A are in fact allowed to compile profile O; the problem is that H-PI does not ensure that this only happens through a rigorous process involving the TTP. Second, for any non-optimal node $A \vDash O$, node $A' \vDash O$ is optimal in which hospitals and CI in A are made curious. This expresses that these actors are allowed to store more data than is desirable. Such data are needed to overcome the absence of the TTP.

Finally, H-PI satisfies the privacy requirements from Table 1. Indeed, the BSN itself never leaves the hospitals; however, the model does not capture that the BSN can be determined from its hash using a dictionary attack.

Pseudonymisation Service. We now discuss privacy in the Pseudonymisation Service infrastructure. We compare it to the hash-based infrastructure (Figure 5(a)) and to the optimal situation (Figure 5(b)).

Figure 5(a) shows that all non-optimal nodes of H-PI (shown dotted) are eliminated in PS-PI; however, PS-PI introduces new non-optimal nodes (shown dashed) which reflect two new privacy problems. The first problem is that the PS *ttp* learns the patient's BSN in the pseudonymisation process, and can contribute this information to coalitions that should not know it. This is reflected by nodes $\{ttp^*\} \vDash \{bsn\}$, $\{ci, ttp\} \vDash \{bsn, d_1, \ldots, d_6\}$, and $\{ci, ttp, r\} \vDash \{bsn, d_1, \ldots, d_7\}$ (in H-PI, these actors know the same data, but without the BSN). The second problem is that the PS is able to link profiles of researchers and hospitals without involving the CI. This problem, combined with the first problem, is reflected by nodes $\{ttp, r\} \vDash \{bsn, d_1, d_2, d_4, d_5, d_7\}$ (linking profiles from different research projects); $\{umc_1, ttp, r\} \vDash \{bsn, d_1, d_2, d_3, d_4, d_5, d_7\}$ and $\{umc_2, ttp, r, \} \vDash \{bsn, d_1, d_2, d_4, d_5, d_6, d_7\}$ (linking profiles from researcher and hospital); and $\{umc_1, umc_2, ttp, r\} \vDash \{bsn, d_1, \ldots, d_7\}$ (combination of the two). As Figure 5(b) shows, these nodes, which all include the PS, are exactly PS-PI's non-optimal nodes.

The analysis shows how privacy protection in PS-PI crucially depends on the trustworthiness of the PS. If we assume that the PS is never involved in privacy



(a) Comparison of coalition graphs of PS-PI and H-PI

(b) Comparison of coalition graphs of PS-PI and O-PI

Fig. 5. Coalition graph comparison of the Pseudonymisation Service infrastructure (PS-PI) with the hash-based infrastructure (H-PI) and the optimal situation (O-PI)

breaches, then coalitions including the PS are not relevant; in this case, PS-PI is optimal. However, without this assumption, PS-PI provides worse privacy than H-PI by offering additional ways to establish links and find out the patient's BSN. In particular, the fact that a curious PS can find out the BSN violates privacy requirement **(PR1)**. To mitigate this, measures should be taken to make sure that the PS cannot use the BSN, e.g., by carrying out all computations on the BSN using trusted hardware (as done by Parelsnoer).

6 From Pseudonymisation Service to Optimal System

In the previous section, we have identified the privacy issues in the PS-PI infrastructure. We now discuss solutions, and then consider a hypothetical infrastructure incorporating these solutions and analyse it using coalition graphs.

The first privacy problem is that the PS learns the patient's BSN. Although it may be mitigated using trusted hardware, it is desirable to technically ensure that the BSN does not leave the hospitals, i.e., that requirement **(PR1)** is fully satisfied. The main challenge in achieving **(PR1)** is that the CI needs to link records from different hospitals. In particular, all hospitals should use the same pseudonym of a patient when communicating with the PS. Intuitively, all hospitals should use a shared secret to generate pseudonyms, or in case they do not share any secret, they should use the same procedure to generate pseudonyms, for instance hashing BSNs as in H-PI. The drawback of the first solution is that

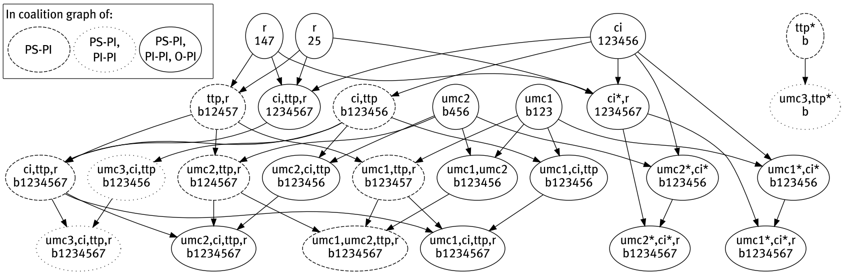


Fig. 6. Comparison of reduced coalition graphs of the improved PS infrastructure (PI-PI) with the original PS infrastructure (PS-PI) and the optimal situation (O-PI)

depseudonymisation can also be performed by hospitals that do not have a record of the patient (e.g., umc_3 in our scenario). On the other hand, if the pseudonyms are generated using one procedure, they may be vulnerable to dictionary attacks as in H-PI. We leave further analysis of this issue as future work.

The second problem is that the PS can help researchers link their data to hospitals or other researchers, bypassing the CI. To solve this, the PS should not be able to link pseudonymisation requests for different domains. This means that when the CI compiles a dataset for distribution, it should modify its linkable pseudocode before requesting the PS to repseudonymise it. The CI may either use the same secret for all datasets, or use different secrets for different datasets or records: both approaches seem possible.

To evaluate the privacy impact of the discussed solutions on PS-PI, we analyse an infrastructure PI-PI that incorporates the solutions in PS-PI. The order of the messages exchanged in PI-PI is as in Figure 2, but the information transmitted is changed. To make sure the BSN does not leave the hospitals, all hospitals share a symmetric key; instead of providing the BSN to the PS, they provide an encryption of the BSN under this key. To prevent linking of distributed datasets by the PS, the CI has a symmetric key for each research domain; when compiling a dataset for distribution, it sends to the PS not his pseudocode itself, but an encryption of the pseudocode under this symmetric key. Instead of retranslating the pseudocode from the CI's domain, the PS simply constructs a new pseudocode using this encryption as pseudonym.

Figure 6 compares PI-PI with the original infrastructure PS-PI and the optimal situation O-PI. As the figure shows, PI-PI indeed solves the privacy problems in PS-PI; however, one problem remains. Namely, besides umc_1 and umc_2, umc_3 can also help in depseudonymisation although it does not know the patient. Note that H-PI does not have this problem because umc_3 does not know the BSN and birth date of the patient. Hence, the privacy of H-PI and PI-PI is formally incomparable. In practice, we have a choice between depseudonymisation by any hospital knowing a secret (PI-PI), or by any third party able to perform a dictionary attack (H-PI).

7 Related Work and Conclusions

In this work, we formally analysed privacy by data minimisation in the setting of (de)pseudonymisation infrastructures for centralised medical research databases. We discussed the unavoidable privacy consequences of the requirements of this setting, and used them to analyse two infrastructures proposed by the Dutch Parelsnoer initiative. In the first, depseudonymisation is performed using tables, introducing privacy risks when data of hospitals or the CI are compromised. In addition, the use of hashes makes it vulnerable to a dictionary attack. The second solves these issues, but lets a TTP learn more information than necessary: it learns the BSN and can link distributed datasets. We discussed solutions to these issues and analysed a hypothetical infrastructure incorporating them.

Apart from Parelsnoer, there are several other proposals for (de)pseudonymisation of patient data for medical research. Serveral proposals [13] in the German legal framework are similar to H-PI and PS-PI, so we expect the findings of our analysis to also apply there. A model from Belgium [3] uses not central storage but a pseudonymisation service which also distributes the data (though encrypted so that the PS cannot read it). In such models, the pseudonymisation service can also be split into two parties [14] which separately do not learn any information. More general approaches for the exchange of medical data between health care providers [6, 15, 20] or pseudoynmised data in general [17] may also be adapted for pseudonymisation for research purposes. To our knowledge, there are no studies that analyse or compare privacy characteristics of these systems; this is an interesting direction for future work. Ultimately, the question is whether optimal data minimisation in this setting is (practically) achievable.

To formally analyse data minimisation, we introduced a novel representation of actor knowledge called coalition graphs. This graph shows which coalitions of actors in a system can compile which profiles of personal information. Honest and honest-but-curious actors capture different assumptions on their behaviour. An "optimal" coalition graph captures unavoidable knowledge; by comparing it to the coalition graph of an existing system, areas for privacy improvement can be identified. We have developed tools to automatically obtain a coalition graph from a formal model of communication.

Other formal methods, e.g. [1, 2, 4, 5, 7, 19], analyse knowledge of communicating actors. These methods verify that *particular* information cannot be derived by *particular* actors. In contrast, we express *all* relevant knowledge of *all* coalitions of actors in one single representation. These methods also do not usually distinguish between honest and honest-but-curious actors. In BAN-style [2] belief logics, a "Forget" operation has been proposed [16] usable for privacy analysis of honest actors [1]. Our work is more similar to state exploration techniques (e.g., [4, 5, 7, 19]). These only consider an (outside) attacker who may be passive or active, but always remembers everything he observes. The operation of honest actors could be simulated indirectly using these techniques; instead, our framework captures their operation explicitly.

Two issues not considered in our model are interesting for future work: first, linking medical data using statistical methods rather than by pseudonyms;

second, deriving implicit knowledge [19], for instance whether a researcher knows at which hospital the medical data in his dataset have been collected.

Acknowledgements. We are grateful for comments received from several people involved with Parelsnoer: Evert Jan Evers, Hans Maring, and Léon Haszing. This work is funded by the Dutch Sentinel Mobile IDM project (#10522), by the Dutch national program COMMIT through the THeCS project, and by the EU Seventh Framework Programme under grant agreement no. 295261 (MEALS).

References

1. Alcaide, A., Abdallah, A.E., González–Tablas, A.I., de Fuentes, J.M.: L–PEP: A logic to reason about privacy–enhancing cryptography protocols. In: Garcia-Alfaro, J., Navarro-Arribas, G., Cavalli, A., Leneutre, J. (eds.) DPM 2010 and SETOP 2010. LNCS, vol. 6514, pp. 108–122. Springer, Heidelberg (2011)
2. Burrows, M., Abadi, M., Needham, R.: A logic of authentication. ACM Trans. Comput. Syst. 8, 18–36 (1990)
3. Claerhout, B., DeMoor, G.J.E.: Privacy protection for clinical and genomic data: The use of privacy-enhancing techniques in medicine. IJMI 74(2-4), 257–265 (2005)
4. Dahl, M., Delaune, S., Steel, G.: Formal analysis of privacy for anonymous location based services. In: Mödersheim, S., Palamidessi, C. (eds.) TOSCA 2011. LNCS, vol. 6993, pp. 98–112. Springer, Heidelberg (2012)
5. Delaune, S., Kremer, S., Ryan, M.: Verifying privacy-type properties of electronic voting protocols. Comput. Secur. 17(4), 435–487 (2009)
6. Deng, M., Cock, D.D., Preneel, B.: Towards a cross-context identity management framework in e-health. Online Information Review 33(3), 422–442 (2009)
7. Dreier, J., Lafourcade, P., Lakhnech, Y.: A formal taxonomy of privacy in voting protocols. In: Proc. of SFCS 2012. IEEE (2012)
8. Fyffe, G.: Addressing the insider threat. Netw. Secur. 2008(3), 11–14 (2008)
9. Guarda, P., Zannone, N.: Towards the Development of Privacy-Aware Systems. Information and Software Technology 51(2), 337–350 (2009)
10. Lo Iacono, L.: Multi-centric universal pseudonymisation for secondary use of the EHR. Stud. Health Technol. Inform. 126, 239–247 (2007)
11. Parelsnoer Initiatief: Programma van eisen intstellingen. Tech Report v.1.2 (2008)
12. Parelsnoer Initiatief: Architecture Central Infrastructure. Tech Report v.1.0 (2009)
13. Pommerening, K., Reng, M.: Secondary use of the EHR via pseudonymisation. Stud. Health Technol. Inform. 103, 441–446 (2004)
14. Quantin, C., et al.: Medical record search engines, using pseudonymised patient identity: An alternative to centralised medical records. IJMI 80(2), 6–11 (2011)
15. Riedl, B., et al.: A secure architecture for the pseudonymization of medical data. In: Proc. of ARES 2007, pp. 318–324. IEEE (April 2007)
16. Rubin, A.D.: Nonmonotonic cryptographic protocols. In: Proceedings of Computer Security Foundations Workshop, pp. 100–116. IEEE (1994)
17. Teepe, W.: Integrity and dissemination control in administrative applications through information designators. Comput. Syst. Sci. Eng. 20(5) (2005)
18. Veeningen, M., de Weger, B., Zannone, N.: Modeling identity-related properties and their privacy strength. In: Degano, P., Etalle, S., Guttman, J. (eds.) FAST 2010. LNCS, vol. 6561, pp. 126–140. Springer, Heidelberg (2011)
19. Veeningen, M., de Weger, B., Zannone, N.: Formal privacy analysis of communication protocols for identity management. In: Jajodia, S., Mazumdar, C. (eds.) ICISS 2011. LNCS, vol. 7093, pp. 235–249. Springer, Heidelberg (2011)
20. Zhang, N., et al.: A linkable identity privacy algorithm for HealthGrid. In: Studies in Health Technology and Informatics, vol. 112, pp. 234–245. IOS Press (2005)