HAL open science

Privacy Patterns for Pseudonymity

Alexander Gabel, Ina Schiering

To cite this version:

Alexander Gabel, Ina Schiering. Privacy Patterns for Pseudonymity. Eleni Kosta; Jo Pierson; Daniel Slamanig; Simone Fischer-Hübner; Stephan Krenn. Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data : 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers, AICT-547, Springer International Publishing, pp.155-172, 2019, IFIP Advances in Information and Communication Technology, 978-3-030-16743-1. 10.1007/978-3-030-16744-8_11 . hal-02271669

HAL Id: hal-02271669

https://inria.hal.science/hal-02271669v1

Submitted on 27 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Privacy Patterns for Pseudonymity

Alexander Gabel[1] and Ina Schiering[2]

[1] Ostfalia University of Applied Sciences
Wolfenbüttel, Germany
ale.gabel@ostfalia.de
[2] i.schiering@ostfalia.de

Abstract. To implement the principle of Privacy by Design mentioned in the European General Data Protection Regulation one important measurement stated there is pseudonymisation. Pseudonymous data is widely used in medical applications and is investigated e.g. for vehicular ad-hoc networks and Smart Grid. The concepts used there address a broad range of important aspects and are therefore often specific and complex. Some privacy patterns are already addressing pseudonymity, but they are mostly abstract or rather very specific. This paper proposes privacy patterns for the development of pseudonymity concepts based on the analysis of pseudonymity solutions in use cases.

Keywords: privacy by design, privacy patterns, pseudonymity, anonymity

1 Introduction

The use of pseudonymisation is proposed in the European General Data Protection Regulation (GDPR) [1] as an important measurement for implementing Privacy by Design and to enhance the security of processing. It would be preferable to render data anonynomous such that the data subject is no longer identifiable, but this has been proven hard in some applications by Naranyan without considerable data utility loss [23]. Pseudonymisation of data is already widely used for the processing of patient data in medical studies or in the context of e-health applications [14]. Other application areas where pseudonymity concepts are investigated include Smart Grid applications [32], vehicular ad-hoc networks (VANETs) [20] where location privacy is in the focus, billing [8] and RFID applications [13].

Compared to this considerable amount of pseudonymity approaches for specific use cases, privacy patterns collections [6] and a review of privacy pattern research by Lenhard et al. [18] mention relatively few pseudonymity patterns in their spreadsheet. These patterns are mainly very abstract as e.g. *Pseudonymous Identity, Pseudonymous Messaging* and few are rather complex e.g. *Attribute-based Credentials* [6] or *Pseudonym Broker Pattern* proposed by Hillen [15].

The aim of this paper is to analyse pseudonymity solutions for use cases in various domains, identify important elements of these solutions and propose additional pseudonymity patterns based on these elements. These patterns are integrated with existing patterns in the context of a pattern language.

2 Related Work

Pseudonymity patterns were already proposed by Hafiz [12]. In his pattern language for privacy enhancing technologies he integrated the rather general pattern *Pseudonymous Identity*. This pattern is described as "hid[ing] anonymity targets under a pseudonym" [12]. It is recommended to hide an identity using a "random pseudonym that does not relate to the original". Hafiz lists important use cases and related work regarding pseudonymisation technologies. The pattern itself is however very generic. Important issues (Insider attacks, Reversibility of the pseudonym mapping, ephemeral pseudonyms etc.) are already mentioned but not addressed in detail.

The pattern *Pseudonymous Messaging* [6] has a focus on a specific use case. The idea is to exchange the communication partners' addresses with pseudonyms known by a trusted third party, which preserves the pseudonymity of users but itself is able to re-identify the pseudonyms. This pattern is also known as *Pseudonymous E-Mail* proposed by Schumacher in 2003 [33]. Pseudonymity may also be implemented using *Attribute-based Credentials* [6], which provide a rather complex but full-fledged identity management solution. Privacy Enhancing Technologies (PETs), such as IBM Identity Mixer allow a user to generate unlinkable pseudonyms, while allowing zero-knowledge attribute verifications [17]. Pseudonyms may also be bound to a certain context (domain pseudonyms), to allow linking multiple visits of the same person. Attribute-based Credentials also provide an Inspector Authority for identity recovery. The *Pseudonym Broker Pattern* was proposed by Hillen [15] and is based on a Trusted Third Party (TTP), which generates pseudonyms from the combination of a subject ID, a partner cloud ID and a time-frame. The pseudonyms are therefore relationship-based and time-limited.

Beside single patterns and pattern languages there are several privacy pattern catalogues. The website privacypatterns.eu consists of 26 patterns, with additional eight dark patterns on the subdomain dark.privacypatterns.eu. Another catalogue is privacypatterns.org covering 50 patterns (duplicates removed) without any dark patterns. In general this catalogue is a superset of privacypatterns.eu. Drozd suggests a catalogue where 38 patterns are classified according to ISO/IEC 29100:2011 (E), to integrate privacy patterns into the software development process [7]. Furthermore Lenhard et al. collected and categorized a large list of 148 (not necessarily unique) patterns from different publications as part of their literature study [18]. Caiza et al. created a taxonomy of types of relationships of patterns [3]. These relationships are also employed here to investigate the connections between pseudonym patterns.

3 Background

As a basis for the following analysis pseudonymisation approaches from different use cases are reviewed. Also the comprehensive investigation of general aspects of pseudonymity in the terminology paper by Pfitzmann et al. [28] are considered.

Pseudonyms are mostly prevalent in the health sector and are particularly used for the pseudonymisation of patients data used in medical research. The data usage can be divided into primary or secondary usage. Often data for medical research projects is derived from data collected during the treatment of diseases etc. and as such described as secondary use. Primary usage however may be present, for example when a new medication is tested, without actual treatment in the first place.

Other uses of pseudonyms may occur, for example when only partial records are transmitted to a third party for evaluation (such as blood samples being sent to a laboratory). Furthermore e-health (Electronic Health Record (EHR), German Electronic Health Card (eGK)) approaches often employ pseudonyms to prevent linkability between multiple health organizations, and as such follow the principle of data separation. Modern approaches for pseudonym-based privacy in e-health are usually data owner centric and protect against attackers from the inside (e.g. administrators).

Riedl et al. created PIPE [25], a privacy-preserving EHR system, which employs layer-based security in combination with pseudonymised data fragments to provide unlinkability between a patient's data and their identity, as well as unlinkability between different health record fragments of the same patient. They furthermore employ a thresholded secret sharing scheme as a mechanism to recover access keys in case of destroyed or lost smart cards [31]. However their approach is patented and therefore there are usage restrictions. Heurix et al. also proposed PERiMETER, which extends their previous work to also include privacy-preserving metadata queries [14].

Caumanns describes an architecture developed for the German electronic health insurance card [5], which was developed at the Fraunhofer Institute for Software and Systems Engineering. The approach uses a ticket-based (challenge-response) method to authenticate users, while keeping links between data fragments hidden using pseudonyms. Stingl and Slamanig also proposed an approach based on unlinkable data fragments in 2007 [35]. Other pseudonymisation systems, which are not data owner centric, often employ trusted third parties (TTPs) [29,26] for organizational separated pseudonymisation (often required by law). The TTPs store pseudonym tables or cryptographic secrets necessary to perform pseudonymisation, and often furthermore allow the inverted mapping: re-identification of pseudonyms. Neubauer and Kolb compare different pseudonymisation methods for medical data with a focus on legal aspects [24].

Another area where pseudonyms are investigated are Smart Grid solutions. Data owners in this scenario are typically inhabitants. Detailed data about their energy consumption is collected by smart meters. Low-frequency data is collected for billing purposes, while high-frequency data may be used for fast demand response and to improve the grid efficiency. Furthermore there may be advanced use cases, such as incentive-based demand response schemes [10]. While low-frequency data was more or less collected previously in combination with the customers identity, high-frequency data may have an high impact on the privacy of inhabitants. Therefore to prevent misuse of the data, many approaches use

pseudonyms to establish unlinkability between the customers identity and the collected power consumption data. Furthermore temporal unlinkability (established through changing pseudonyms) between sequentially recorded profiles is used to reduce the traceability and therefore the risk of re-identification. Rottondi et al. [32] deploy so-called privacy-preserving nodes (PPNs) together with a secret sharing scheme, to separate the pseudonymisation process from the assigned data and to unlink the network address of the smart meter from the pseudonymised data. Finster and Baumgart combine blind signatures, a lightweight one-way peer-to-peer anonymisation network and a bloom filter to realize pseudonymised data collection without a trusted third party, while preserving the unlinkability between network addresses and customer data [9].

Another interesting area to consider is the incorporation of electric vehicles into the smart grid via vehicle-to-grid (V2G) networks. There challenges arise, such as location privacy, when the vehicle is authenticating to the grid in many different places, as needed for online electric vehicles [16], as well as information about the battery level which may be used for further tracking [19].

In the field of vehicular ad-hoc networks (VANETs) privacy-preserving solutions are investigated mainly with a focus on location privacy [37,20], often by using changing pseudonyms. Mano et al. express the need for pseudonymisation of datasets of location trajectories for analysis of mobility patterns. They claim that anonymised datasets (e.g. using k-anonymity) typically do not provide enough information about those patterns, when compared against pseudonymised per-user trajectories [21]. To protect them concerning re-identification, they propose to exchange the pseudonym at hub locations and introduce metrics and a verification algorithm to check whether the pseudonym exchange can be effective for all users based on plausible paths.

In the area of billing, pseudonyms are used to separate the process of payment (which typically but not always [22] requires the identity of the user) from the actual usage of a particular service [38,8]. Furthermore pseudonyms may be applied to create transactions, which are not linkable in different contexts [34]. Gudymenko proposes a privacy-preserving e-ticketing system for fine-granular billing, by separating pseudonymised tracing of travel records and end user billing using a trusted third party [11]. Falletta et al. propose a distributed billing system, which requires the interaction of multiple entities to disclose the user's identity, therefore avoiding a single trusted third party [8].

In RFID systems, regularly changing pseudonyms (often based on cryptographic algorithms) are used to prevent tracking of RFID tags for unauthenticated readers. Henrici et al. apply the concept of onion routing in an RFID tag pseudonymisation infrastructure to prevent unwanted tracking of RFID tags [13].

Biskup and Flegel use transaction-based pseudonyms and apply a thresholded secret sharing scheme in an intrusion detection system to allow re-identification of a particular user only when a certain threshold of policy violations has been exceeded [2].

4 Analysis of Pseudonymity Approaches

To dissect the different pseudonym systems in the use cases summarized in Section 3 as a starting point for the analysis, the following central areas are investigated to identify the basic building blocks of pseudonym systems. First pseudonym generation is investigated and second additional functionality is considered which is necessary for the pseudonym system to fulfil its purpose.

When a pseudonym is used to protect privacy, its purpose is usually to foster the unlinkability between an individual and its pseudonyms. Therefore an important question is the scope a pseudonym. As described by Pfitzmann and Hansen [28], there are different types of pseudonyms, depending on the scope/context[3] of their usage (e.g. role pseudonym, relationship pseudonym, transaction pseudonym etc.). We extend this concept to a general scope, which may be defined by a combination of many factors that limit the usage/validity of a pseudonym. For example a pseudonym may be time-limited (i.e. only valid for one week), as well as relationship-based (i.e. differs for each party interacting with the pseudonym). The idea of the *Minimal Pseudonym Scope* pattern we propose, is to limit this scope to the smallest possible one for the purpose of data processing.

Another component of pseudonym generation is how the actual pseudonym is created. Hafiz suggests in the *Pseudonymous Identity* pattern, that "a random pseudonym [should be adopted], that does not relate to the original" [12]. However, it is not always the case that a pseudonym is really random, since typically pseudonyms are generated. For example cryptographic techniques may be used to generate a pseudonym, e.g. by encrypting or hashing certain information. This is especially useful when pseudonyms should be re-identifiable by a trusted third party. Furthermore, techniques such as *Attribute-based Credentials* also allow the creation of pseudonyms, which may need to fulfil certain cryptographic properties, e.g. in the case of a domain pseudonym. The actual method of generation often depends on other properties of the pseudonym system, therefore no additional pattern is proposed in this area.

In many cases pseudonyms are generated by a trusted third party, which in most cases also allows this trusted third party to re-identify pseudonyms, i.e. to link them back to the original (hidden) identity. However, particularly research regarding pseudonymous e-health systems noticed an inherent risk of re-identification by insiders (e.g. administrators) or due to database leaks. Therefore the abstract pattern *Data-owner based Pseudonymisation* is proposed, which switches roles and allows the data owner to create pseudonyms. The idea is to decrease the risk of re-identification and unwanted linkability in comparison to a trusted third party for pseudonym creation. This however does not mean that re-identification (e.g. in the case of misuse) is always completely impossible. For example in the case of *Attribute-based Credentials* with the presence of an inspector authority, it is still possible to recover the identity behind a pseudonym, while the separation of entities (issuer, verifier, inspector authority and user)

[3] In this paper the notion "scope" is used in order to prevent confusion with the context of patterns.

separates powers. Furthermore in some cases it may be sufficient to let the user prove that she/he is or is not the holder of a pseudonym, e.g. in legal disputes, without a trusted party being able to recover an identity behind a pseudonym.

For the second category regarding additional functionality in pseudonym systems, two main strategies were identified: Protection against re-identification and its counterpart Selective Linkability if needed for a specific service. To protect against re-identification, especially the use of *Anonymisation Networks* is common. The existing pattern *Onion Routing* is not always used, instead proxies or lightweight anonymisation networks are employed, especially in Internet of Things use cases such as Smart Grid or RFID systems. This may indicate, that a more abstract pattern *Anonymisation networks* is necessary to capture the diverse requirements and approaches of such systems.

Furthermore with additional data which may being linked to a pseudonym, the risk of re-identification due to inference attacks increases. To cope with this risk, strategies such as data de-identification/de-sensitization can be used. However, this may also decrease the utility of the data and therefore the quality of the service. Another approach, especially present in e-health use cases is to separate the data into small fragments, which are unlinkable by default but may be linked by the data owner. This prevents trivial linkability for insiders as well as in the case of a database breach, while keeping utility to authorized parties. Hence the pattern *Data fragments* is proposed especially for the use in the e-health context, were sensitive (i.e. medical) data is processed.

To allow users to share the information which pseudonyms for data fragments are connected to the same individual in a selective way, the *Encrypted Link* pattern as a kind of data owner based authorization system is proposed, in contrast to standard access control systems. While the *Data fragments* pattern can be seen as primarily establishing unlinkability and preventing re-identification, the *Encrypted Link* pattern selectively establishes linkability in a secure way without leaking unnecessary information to unauthorized entities.

When the pattern *Minimal Pseudonym Scope* is applied, but selective linkability is necessary to exchange data across different scopes, the *Pseudonym Converter* pattern can be applied. When party A wants to send data regarding a pseudonymous subject S to party B, but A and B have their own distinct pseudonyms referring to S, a pseudonym converter may translate between the parties without directly establishing the link between two pseudonyms. On the other side, given a pseudonym system based on a trusted third party responsible for pseudonym generation, a good practice for data minimization is to apply the *Data hidden from Pseudonymiser* pattern, such that the pseudonymiser is only responsible for translating between identities and pseudonyms without access to related data, not necessary for that sole purpose. We found different methods for hiding the data, such as de-identification, encryption and secret splitting.

Finally it may be necessary to recover the identity behind a pseudonym for reasons such as handling of misuse. This functionality was required in many systems and handled in different ways, therefore the pattern *Recoverable Identity* is proposed to capture this important concept and ways to implement it.

5 Pseudonymity Patterns

In this section we present our patterns for pseudonymity, as well as relations between those pattern and existing patterns (see Fig. 1). We created eight patterns, however due to page limitations, we present only a subset of them[4]. An overview of the remaining can be seen in Table 1.

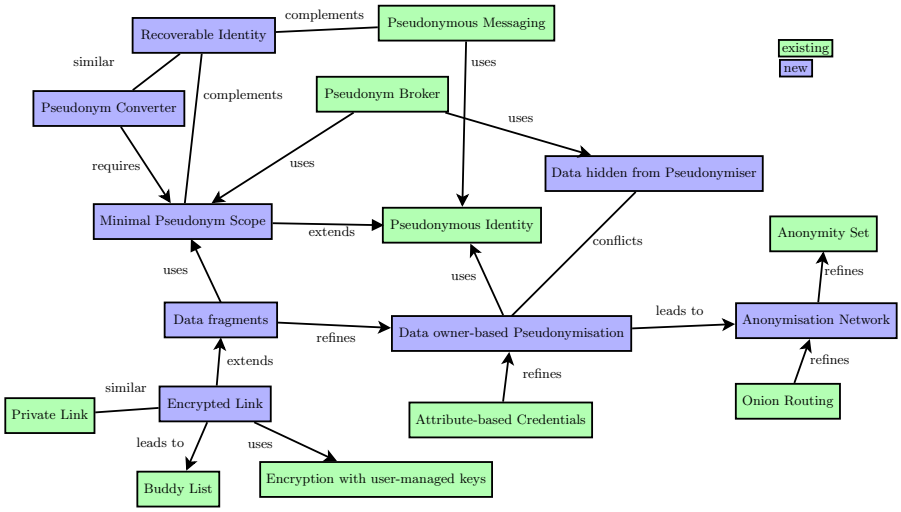

Fig. 1. Pattern language for pseudonymity patterns

5.1 Minimal Pseudonym Scope

Summary: Restrict the linkability of a pseudonym by limiting the usage to the smallest possible scope for the purpose of data processing (data minimization). **Context:** It is often not necessary for a pseudonym to have a very broad scope in the general case. Even if linkability across different scopes is necessary, usually not every party (e.g. an attacker) should be able to link pseudonyms trivially. **Problem:** Pseudonyms are usually used to protect an identity from being disclosed. However when using only a single unique pseudonym for an identity, it becomes increasingly traceable and it may be linked across several databases and scopes. With more information about an identity, re-identification of a pseudonym becomes increasingly likely. Also in case of a data breach, datasets with potentially different information about an identity, which refer to the same pseudonym become linkable for attackers.

[4] The full pattern catalogue can be retrieved from `https://github.com/a-gabel/pseudonym-privacy-patterns`

Forces/Concerns: Controllers may want linkability across different scopes for some services. Users may prefer not to be tracked across multiple scopes.
Solution: To prevent linkability across different scopes using a pseudonym, one may limit the use of a pseudonym to a small scope. For different scopes, different pseudonyms are used, which cannot be linked without additional information. A scope may be depend on a role (e.g. shopping or video on demand), relationship (Company A or B), location, time frame or transaction (one-time use). Furthermore combinations may be useful (e.g. role-relationship), depending on the use case. The controller needs to balance the purpose of the service and privacy of users. The scope has to be chosen according to the principle of data minimization. Selective Linkability can also be established via *Recoverable Identity* or *Pseudonym Converter*, which might decrease the risk in case of a data breach.
Benefits: In case of a data breach, pseudonyms across different scopes may not be linked trivially. Pseudonyms only refer to (small) partial identities, which cannot be linked trivially.
Liabilities: Additional complexity may be necessary, if linkability of pseudonyms in different scopes is necessary under certain conditions (e.g. by applying a *Pseudonym Converter*).
Examples: A user may use different **relationship-pseudonyms** [28], to limit linkability across different organizations. For example a user may want to use a different pseudonym for a dating website and for their business profile. Furthermore the pseudonym of a car in a car-to-x network may change depending on **location** and **time-frame**.
[Known Uses]: Pommerening and Reng use a different pseudonym for each secondary use project of electronic health record (EHR) data [29]. Mano et. al exchange pseudonyms of users when they meet at the same hub and propose a privacy verification algorithm [21]. Rottondi et al. use a time-limited pseudonym to prevent linkability of smart meters over a longer time window in a smart grid system [32]. Industrial uses include the GSM standard with the Temporary Mobile Subscriber Identity (TMSI; time- and location-limited scope), or tokenization which is recommended by the Payment Card Industry Data Security Standard (PCI DSS), resulting in different pseudonyms per party (relationship-based) and time-frame (validity of the tokenization key) [27].
[Related Patterns]: Extends *Pseudonymous Identity*, as it improves the existing solution of protecting identities behind a pseudonym by giving it a small scope, thus making it more difficult to re-identify. Complements *Recoverable Identity*, as the small scope leads to less data being linked to the real identity in case of re-identification. It is complemented by *Recoverable Identity*, as it may help to prevent misuse when many pseudonyms make it hard to track/block a user. Used by *Pseudonym Broker*, as pseudonyms are different for each organization and time frame, as well as by *Data Fragments* and *Data owner-based Pseudonymisation*. Required by *Pseudonym Converter*.

5.2 Recoverable Identity

Summary: The identity behind a pseudonym is recoverable under certain conditions.
Context: Pseudonym system are usually designed such that the re-identification of a pseudonym (i.e. determining the identity behind a pseudonym) is reasonably hard. In many cases it is sufficient to be able to link different transactions via pseudonyms. However in some cases, it might be necessary to recover the identity behind a pseudonym, for example in case of misuse of the system. Then e.g. only a trusted party/combination of multiple trusted parties should be able to recover the identity behind a pseudonym.
Problem: If identity recovery is necessary, it should usually only be possible in very specific and constrained cases.
Forces/Concerns: Users may fear, that their identity is recovered in cases where it is not necessary (e.g. the user did not misuse the system), resulting in compromise of their privacy. Therefore the trusted party which is able to recover pseudonyms should transparently show and enforce their policies. The party should be trusted by both the controller and the users. The controller may want to identify users, e.g. for legal or payment purposes.
Solution: Restrict the ability of identity recovery via organizational and technical constraints.
[**Implementation:**] One option can be to use a Trusted Third Party for Identity Recovery. The pseudonym mapping may be stored in a table or encrypted inside the pseudonym. Another option is to use secret sharing to allow identity recovery only with $n > t$ operators, or with enough evidence (in case of misuse). Furthermore anonymous credentials / *Attribute-based Credentials* with a trusted inspector authority may be used.
Benefits: The identity behind a pseudonym is only recoverable in very specific, constrained cases. Misuse of the system by pseudonymous users may be limited, as users are informed about the possibility of identity recovery in such cases.
Liabilities: Users may have less trust in the system, if the policy for identity recovery or the technological barriers are too lax.
Examples: In a Pseudonymous Messaging system where users are communicating via email, pseudonymous users may be re-identified by a trusted third party, if they abuse the system, e.g. for illegal purposes. The pseudonymiser (the entity which translates real email addresses to pseudonymous ones) encrypts the original identity inside the pseudonymous e-mail address and is therefore the only entity which is able to recover an identity from a pseudonym only. Another example: In a smart grid system, each smart meter uses a pseudonym, which is generated by encrypting identifiable information (e.g. an ID known to the grid operator) using the public key of a trusted third party (TTP). The TTP may recover identities behind pseudonyms in case of misuse using its private key.
[**Known Uses]:** Hussain et al. use a secret sharing scheme to allow only the combination of all revocation authorities to recover the identity behind the pseudonym of an online electric vehicle (OLEV) in case of a legal need, such as refusing to pay after electricity consumption [16]. Rottondi et al. allow the Configurator,

a trusted party of a Smart Grid system, the recovery of identities by decrypting the identity as part of the pseudonym using its private key [32]. Biskup and Flegel use a secret sharing scheme to allow re-identification of pseudonyms in an intrusion detection system only when there is enough evidence (i.e. enough events from a certain identity within a time-frame) [2]. Attribute-based Credential Systems allow re-identification of users via a separate Inspector authority.
[Related Patterns]: Complements *Pseudonymous Messaging*, as it may help to prevent misuse of the messaging service. Complements *Minimal Pseudonym Scope*, as it helps to re-identify users in case of misuse. Similar to *Pseudonym Converter*, as both patterns allow a trusted third party (TTP) to selectively link a pseudonym. In case of the *Pseudonym Converter*, a TTP can link pseudonyms, while in *Recoverable Identity* the TTP can link a pseudonym to an identity.

5.3 Data hidden from Pseudonymiser

Summary: Data being pseudonymised is not readable by the Pseudonymiser (entity which assigns pseudonyms).
Context: The pseudonymiser (i.e. the entity which creates pseudonyms and assigns them to identities) is usually only responsible for assigning pseudonyms, but does not need to have access to additional data. For example a pseudonymisation entity for medical data may not need access to the assigned medical reports etc.. Additionally, pseudonyms may be generated based on unique IDs instead of identifiable information (e.g. name).
Problem: When assigning a pseudonym to an identity the pseudonymiser might learn additional information, which may be unwanted and unnecessary.
Forces/Concerns: The pseudonymiser needs some kind of reference to the original identity. However, information about the person (such as the name or further information) may not be necessary. A secure channel between a data source and the party which receives pseudonymised data might be needed.
Solution: Hide data assigned to an identity by e.g. applying cryptographic measures before pseudonymisation.
[Implementation]: Encryption of the data: Before sending an identity and data to a pseudonymiser, encrypt the assigned data using public key cryptography. The pseudonymiser will receive a tuple $(ID, Enc(data))$ from the data source as pseudonymisation request and will send a tuple $(Pseudonym, Enc(data))$ to a party from which the real identity should be hidden. The receiving party is able to decrypt the hidden data using its private key. Secret Sharing: Use a secret sharing scheme to split the assigned data into parts, which are then pseudonymised by multiple distinct pseudonymisers. The receiving party is able to reconstruct the data if all parts are received, but each pseudonymiser on its own is unable to do so. De-identification: If the pseudonymiser for a specific reason needs to have access to the assigned data, the additional use of de-identification methods to remove identifiable data (e.g. name, ID card number, birth data, ...) is strongly recommended.
Benefits: The pseudonymiser does not learn additional information about an identity. Identities may be referred to as unique random identifiers, such that

other identifiable data (such as a person's name) is also not available to the pseudonymiser.
Liabilities: Additional complexity of the system may arise depending on how the hiding mechanism is implemented.
Examples: A medical clinic may need to pseudonymise patients' medical data to be used in a research project. Instead of sending complete patient records with identifiable data (name, birth date etc.) to a pseudonymiser, only a list of randomly generated unique IDs is sent to the pseudonymiser. The pseudonymiser then converts each ID to a unique pseudonym and sends the resulting list (with the same order as the original list) to the research organization. Furthermore the clinic sends de-identified medical records (same order) to the research organization. The research organisation may then refer to a patient using the pseudonym from the list, while the pseudonymiser does not have any access to the medical data. Instead of sending the medical data separately, a clinic may also encrypt it for the research party and send it encrypted to the pseudonymiser, who is unable to read the encrypted data.
[**Known Uses]:** Pommerening and Reng hide associated medical data for the pseudonymiser by encrypting it for the receiving research organization [29]. Noumeir et al. perform de-identification of radiology data before sending it to a pseudonymisation system to reduce the risk of identification [26]. Rottondi et al. use a secret splitting scheme in a smart meter system to let several pseudonymisation nodes pseudonymise shares of a smart meter (producer) reading, ensuring that these nodes cannot read the data, while the receiving node (consumer) can do so, when receiving all secret shares [32]. Rahim et al. perform pre-pseudonymisation of patient identifiers in addition to encryption of the assigned medical data to completely hide identifiable information from the pseudonymisation server [30].
[**Related Patterns]:** Used by *Pseudonym Broker*, as the data assigned to a pseudonym is sent to a database or to a portal without any interaction with the Trusted Third Party, which acts as the pseudonymiser. Conflicts with *Data owner-based Pseudonymisation*, because the data owner (i.e. the pseudonymiser) already has knowledge of the data and it is not useful to hide that data. Complements *Pseudonymous Messaging*, as it hides the message content from the party which performs the pseudonymisation of the messages, providing additional privacy.

5.4 Data fragments

Summary: Split data of a single identity into small fragments and assign each fragment its own pseudonym. Only authorized entities are given the knowledge of which pseudonyms belong together.
Context: Whenever a collection of pseudonymised data records are under risk of re-identification by inference attacks due to the informative value of combined fields.
Problem: A record of data about an identity may contain enough information to re-identify it, even if primary identifiers are removed from the record. For

example the combination of the attributes gender, ZIP code and birth date may uniquely identify 87% of the US-American population [36]. Furthermore it may be unwanted in a system to enable anyone with access to the dataset (e.g. also insiders like administrators) to be able to link sensitive data.
Forces/Concerns: Using server-side encryption may not help, if insiders such as administrators have access to the encryption keys. Encrypting the data using end-to-end encryption (i.e. unauthorized entities do not have access to the keys) might help, however when the dataset is large the performance penalty may be unacceptable/impractical. De-identification of the data using techniques from the area of Statistical Disclosure Control may work for some scenarios. However, such techniques may remove data needed for the use case.
Solution: Instead of storing related data with a single pseudonym, split the data into small fragments, which are hard to re-identify by themselves, and assign each fragment its own unique pseudonym. Only authorized persons or systems get the knowledge of which pseudonyms (i.e. which data fragments) belong to the same identity. It is also possible to reveal only partial information about which fragments belong together, to limit access to certain parts of data records. The pattern may furthermore be combined with de-identification to de-sensitize potentially identifiable data such as birth dates (e.g. mask day and month of birth) before transmitting the data.
Benefits: Enables unlinkability of data fragments by default, while authorized entities are able to link subsets of fragments. May significantly reduce the risk of insider attacks, as insiders are unable to link fragments or establish a relation to an identity. In case of a data breach, data fragments remain unlinkable for attackers without additional knowledge. Computationally efficient, as data fragments do not necessarily have to be encrypted
Liabilities: Increases complexity of the system, as knowledge about pseudonyms needs to be managed.
Examples: In an e-health system where health records or metadata of records from patients are stored centrally, instead of storing data referring to the same person in a linkable way, data fragments may be used to split health records into small fragments. For example each medical result is stored as a separate fragment. Only the data owner (i.e. the patient) has the knowledge which pseudonyms/data fragments belong to her. When the data owner wants to share fragments with a doctor, new pseudonyms pointing to the fragments can be generated and shared with the doctor.
[Known Uses]: PIPE (Pseudonymisation of Information for Privacy in E-Health) uses data fragments for electronic health records. By default only the patient (data owner) is able to access her health records. Access to the pseudonyms is managed through a central metadata storage which is encrypted with the users keys. The data owner may decide to give access to some records to selected medical personnel by creating additional pseudonyms referring to fragments [25]. Identifiable and non-identifiable data is also unlinkable by default, so the system may be employed for secondary use, e.g. in research. Stingl and Slamanig describe a concept for an e-health portal, which uses unlinkable and undetectable

partial identities of a patient to keep separate health records for participating parties (i.e. dentist and general practitioner access different partial identities) [35]. The Fraunhofer ISST designed a concept for the German electronic health card (eGK), which uses ticket-based authorization and challenge-based authentication to allow fine-granular access control to data fragments, which are unlinkable by default [5]. Biskup and Flegel use a secret sharing scheme to assign each event in an intrusion detection system a unique pseudonym, which keeps events unlinkable until enough evidence for re-identification is available [2]. Camenisch and Lehmann propose the use of "data snippets", which are stored with unlinkable pseudonyms. A central entity is able to link those snippets and may provide de-identified subsets of the original record to authorized parties. They suggest the use a central *Pseudonym Converter*, which is able to convert pseudonyms in a blind way while providing auditability for users [4].
[Related Patterns]: Uses *Minimal Pseudonym Scope*, as every data fragment gets its own pseudonym, therefore the scope of a pseudonym is very limited. Refines *Data owner-based pseudonymisation*, as it allows the data owner in a more specific context (i.e. shared repository of data) to perform the pseudonymisation and therefore provide a more privacy-preserving solution in comparison to a trusted third party solution. Extended by *Encrypted Link*.

Pattern	**Description**
Pseudonym Converter	A separate entity, the Converter, is able to translate a pseudonym from one scope to a pseudonym in another scope.
Encrypted Link	To authorize access to *data fragments* in a way that is not detectable by third parties, encrypt pseudonyms, pointing to *data fragments*.
Anonymisation Network	Hide the network identity of a communication partner by adding anonymisation nodes between communication partners.
Data owner-based Pseudonymisation	Generate and assign pseudonyms on the data owner side instead of using a third party, to keep the link between pseudonym and the data owner hidden from other parties.

Table 1. Further patterns for pseudonymity (summary)

6 Discussion and Final Remarks

In this paper privacy patterns and a pattern language for pseudonymity is proposed, which try to close the gap between very abstract and very complex pseudonymity patterns and to ease the development of pseudonym systems. For data minimization and unlinkability *Minimal Pseudonym Scope*, *Data fragments*, and

Data hidden from pseudonymiser may be applied. To establish selective linkability, while staying restricted to a small audience, *Recoverable Identity*, *Pseudonym Converter* and *Encrypted Link* can assist. Furthermore to shift the asymmetry of power to the data owner side, the patterns *Data owner-based Pseudonymisation* and *Anonymisation Networks* are useful concepts.

To foster the adoption of privay patterns, the applicability of such patterns in system development processes needs to be evaluated to derive guidelines for developers in the context of privacy engineering processes.

The taxonomy of relations between patterns by Caiza et al. [3] is a promising approach to provide an overview of privacy patterns, because of the visual structure and the relations between patterns. Some of the pattern relations proposed there were not fully applicable in the context of privacy patterns. E.g. *leads to* specifies that a pattern is necessary, as to not leave unsolved problems. However, in our experience the existence of problems may depend on the use case (e.g. *Recoverable Identity*). Also the relation *complements* is defined as symmetric, which was not always the case here.

The importance of a relationship itself is a topic which may be discussed further. Some relationships may be redundant or not helpful (i.e. referencing *Pseudonymous Identity* from every pattern regarding pseudonymity), while others may give helpful insights.

Regarding the patterns for pseudonymity it has to be shown, whether this catalogue is complete or if there may be more patterns, yet to be discovered. An interesting question is, whether it is actually possible to check that a pattern language is exhaustive or to at least get hints where something may be missing. Another observation is the difference in the level of abstraction/complexity between the patterns. Developing a hierarchy of patterns, or clusters of complexity/abstractions could be a useful concept.

Acknowledgement This work was supported by the Ministry for Science and Culture of Lower Saxony as part of SecuRIn (VWZN3224).

References

1. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) `http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC`
2. Biskup, J., Flegel, U.: On pseudonymization of audit data for intrusion detection. In: Designing Privacy Enhancing Technologies. pp. 161–180. Springer (2001)
3. Caiza, J.C., Martín, Y.S., Del Alamo, J.M., Guamán, D.S.: Organizing design patterns for privacy: A taxonomy of types of relationships. In: Proceedings of the 22Nd European Conference on Pattern Languages of Programs. pp. 32:1–32:11. EuroPLoP '17, ACM, New York, NY, USA (2017)
4. Camenisch, J., Lehmann, A.: Privacy-preserving user-auditable pseudonym systems. In: 2017 IEEE European Symposium on Security and Privacy (EuroS P). pp. 269–284 (April 2017)

5. Caumanns, J.: Der Patient bleibt Herr seiner Daten Realisierung des eGK-Berechtigungskonzepts über ein ticketbasiertes, virtuelles Dateisystem. Informatik-Spektrum 29(5), 323–331 (October 2006)
6. Colesky, M., Hoepman, J.H., Bösch, C., Kargl, F., Kopp, H., Mosby, P., Le Métayer, D., Drozd, O., del Álamo, J.M., Martín, Y.S., Gupta, M., Doty, N.: Privacy patterns, https://privacypatterns.org/, accessed on 1st August 2018
7. Drozd, O.: Privacy pattern catalogue: A tool for integrating privacy principles of ISO/IEC 29100 into the software development process. In: IFIP International Summer School on Privacy and Identity Management. pp. 129–140. Springer (2015)
8. Falletta, V., Teofili, S., Proto, S., Bianchi, G.: P-DIBS: Pseudonymised DIstributed billing system for improved privacy protection. In: 2007 16th IST Mobile and Wireless Communications Summit. pp. 1–5 (July 2007)
9. Finster, S., Baumgart, I.: Pseudonymous smart metering without a trusted third party. In: 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications. pp. 1723–1728 (July 2013)
10. Gong, Y., Cai, Y., Guo, Y., Fang, Y.: A privacy-preserving scheme for incentive-based demand response in the smart grid. IEEE Transactions on Smart Grid 7(3), 1304–1313 (2016)
11. Gudymenko, I.: A privacy-preserving e-ticketing system for public transportation supporting fine-granular billing and local validation. In: Proceedings of the 7th International Conference on Security of Information and Networks. pp. 101:101–101:108. SIN '14, ACM, New York, NY, USA (2014)
12. Hafiz, M.: A pattern language for developing privacy enhancing technologies. Software: Practice and Experience 43(7), 769–787 (2013)
13. Henrici, D., Gotze, J., Muller, P.: A hash-based pseudonymization infrastructure for RFID systems. In: Second International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing (SecPerU'06). pp. 6 pp.–27 (June 2006)
14. Heurix, J., Karlinger, M., Neubauer, T.: Pseudonymization with metadata encryption for privacy-preserving searchable documents. In: 2012 45th Hawaii International Conference on System Sciences. pp. 3011–3020 (January 2012)
15. Hillen, C.: The pseudonym broker privacy pattern in medical data collection. In: 2015 IEEE Trustcom/BigDataSE/ISPA. vol. 1, pp. 999–1005 (August 2015)
16. Hussain, R., Son, J., Kim, D., Nogueira, M., Oh, H., Tokuta, A.O., Seo, J.: PBF: A new privacy-aware billing framework for online electric vehicles with bidirectional auditability. Wireless Communications and Mobile Computing 2017 (2017)
17. IBM Research - Zürich: Specification of the identity mixer cryptographic library version 2.4.43, https://abc4trust.eu/index.php?option=com_content&view=article&id=187, accessed on 1st August 2018
18. Lenhard, J., Fritsch, L., Herold, S.: A literature study on privacy patterns research. In: Software Engineering and Advanced Applications (SEAA), 2017 43rd Euromicro Conference on. pp. 194–201. IEEE (2017)
19. Liu, H., Ning, H., Zhang, Y., Guizani, M.: Battery status-aware authentication scheme for v2g networks in smart grid. IEEE Transactions on Smart Grid 4(1), 99–110 (2013)
20. Lu, R., Lin, X., Luan, T.H., Liang, X., Shen, X.: Pseudonym changing at social spots: An effective strategy for location privacy in VANETs. IEEE Transactions on Vehicular Technology 61(1), 86–96 (January 2012)
21. Mano, K., Minami, K., Maruyama, H.: Privacy-preserving publishing of pseudonym-based trajectory location data set. In: 2013 International Conference on Availability, Reliability and Security. pp. 615–624 (September 2013)

22. Martinez-Pelaez, R., Rico-Novella, F., Satizabal, C.: Mobile payment protocol for micropayments: Withdrawal and payment anonymous. In: 2008 New Technologies, Mobility and Security. pp. 1–5 (November 2008)
23. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. pp. 111–125. SP '08, IEEE Computer Society, Washington, DC, USA (2008)
24. Neubauer, T., Kolb, M.: Technologies for the pseudonymization of medical data: A legal evaluation. In: 2009 Fourth International Conference on Systems. pp. 7–12 (March 2009)
25. Neubauer, T., Heurix, J.: A methodology for the pseudonymization of medical data. International Journal of Medical Informatics 80(3), 190–204 (March 2011)
26. Noumeir, R., Lemay, A., Lina, J.M.: Pseudonymization of radiology data for research purposes. Journal of Digital Imaging 20(3), 284–295 (September 2007)
27. PCI Security Standards Council: Tokenization product security guidelines. Tech. Rep. 1.0, PCI Security Standards Council (April 2015), https://www.pcisecuritystandards.org/documents/Tokenization_Product_Security_Guidelines.pdf
28. Pfitzmann, A., Hansen, M.: A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management (2010)
29. Pommerening, K., Reng, M.: Secondary use of the EHR via pseudonymisation. Studies in Health Technology and Informatics 103, 441–446 (2004)
30. Rahim, Y.A., Sahib, S., Ghani, M.K.A.: Pseudonmization techniques for clinical data: Privacy study in sultan ismail hospital johor bahru. In: 7th International Conference on Networked Computing. pp. 74–77 (September 2011)
31. Riedl, B., Grascher, V., Neubauer, T.: Applying a threshold scheme to the pseudonymization of health data. In: 13th Pacific Rim International Symposium on Dependable Computing (PRDC 2007). pp. 397–400 (December 2007)
32. Rottondi, C., Mauri, G., Verticale, G.: A data pseudonymization protocol for smart grids. In: 2012 IEEE Online Conference on Green Communications (GreenCom). pp. 68–73 (September 2012)
33. Schumacher, M.: Security patterns and security standards - with selected security patterns for anonymity and privacy. In: Privacy,"European Conference on Pattern Languages of Programs (EuroPLoP (2003)
34. Seigneur, J.M., Jensen, C.D.: Trust enhanced ubiquitous payment without too much privacy loss. In: Proceedings of the 2004 ACM Symposium on Applied Computing. pp. 1593–1599. SAC '04, ACM, New York, NY, USA (2004)
35. Stingl, C., Slamanig, D.: Berechtigungskonzept für ein ehealth-portal. na (2007)
36. Sweeney, L.: Simple demographics often identify people uniquely. Health (San Francisco) 671, 1–34 (2000)
37. Thenmozhi, T., Somasundaram, R.M.: Pseudonyms based blind signature approach for an improved secured communication at social spots in VANETs. Wireless Personal Communications 82(1), 643–658 (May 2015)
38. Zhao, X., Li, H.: Privacy preserving authenticating and billing scheme for video streaming service. In: Cyberspace Safety and Security. pp. 396–410. Lecture Notes in Computer Science, Springer, Cham (October 2017)