DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Privacy preservation in Data Markets for IoT devices: A Systematic Review

Ilias Soto-Alaoui

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Information Systems

Privacy preservation in Data Markets for IoT devices: A Systematic Review

Schutz der Privatsphäre auf den Datenmärkten für IoT-Geräte: Ein systematischer Review

Author:	Ilias Soto-Alaoui
Supervisor:	Gonzalo Munilla-Garrido
Advisor:	Prof. Dr. Florian Matthes
Submission Date:	15.10.2020

I confirm that this bachelor's thesis in information systems is my own work and I have documented all sources and material used.

Munich, 15.10.2020 Ilias Soto-Alaoui

Acknowledgments

I would like to thank Professor Dr. Florian Matthes for the opportunity to write my thesis at his chair for Software Engineering for Business Information Systems (sebis).

I would also like to thank Ömer Uludag for providing feedback on the core structure of this review.

Finally, I would like to thank my supervisor Gonzalo Munilla-Garrido for his support during this research project. The discussions we had and his attention to detail have definitely made me a better researcher.

Abstract

Big data is considered to be the key to unlocking the next wave of growth in productivity, especially as the concept of the Internet of Things (IoT) increasingly becomes a part of the modern world. The exponential growth of data has created applications in various landscapes, such as e-healthcare industry, many mobile crowdsensing systems, machine learning, or smart mobility, just to name a few. One of the issues that arises from this situation is that data owners do not know how to effectively utilize their data after collecting it, while other entities could thrive on its findings and would be willing to pay for it. To solve this issue, several so-called data marketplaces have been implemented. However, one of the requirements that most of these platforms do not fully support is data privacy for data owners, a problem that has sparked the interest of researchers around the globe who have tried to come up with proposals on how to design privacy-preserving data markets. To this day, no proposal has established itself as the best solution. In this thesis, we provide an overview of the technologies used in the most relevant proposals and studies concerning privacy-preserving data markets for IoT devices. For this, we conduct a systematic literature review (SLR) as described by the leading researchers in the field of "Evidence-Based Software Engineering" (EBSE). Within this SLR, we identify 50 studies that are relevant according to our selection criteria. Our methodology is designed to minimize bias effects. As for the findings of this thesis, firstly, we identify the main research gaps and problems for privacy-preservation in data markets for IoT devices. Secondly, we describe the current research topics for the implementation of privacy-preserving IoT data marketplaces and structure them into four pillar perspectives and one use-case specific perspective. Thirdly, we discuss and classify the privacy-preservation techniques used and/or mentioned in the selected papers. The implementation of each technique is given with a series of advantages and disadvantages that should be considered for each individual use case and technology combination. Fourthly, we give five broad directions as a research outlook. Moreover, a mapping of the studies' meta-data is presented in order to introduce the selected studies in an aggregated and graphical manner. This allows the reader to know where, when, how and by whom the selected studies were conducted.

Contents

1. Introduction

1.1. Introduction

Big data is considered to be the key to unlocking the next wave of growth in productivity [1]. Nowadays, the total amount of data in the world is exploding, with almost 90% of the data having been created in the last two years [2]. The data sources are diverse, especially as the concept of the Internet of Things (IoT) becomes more and more a part of the modern world [1]. The IoT can be understood as a networking paradigm that integrates billions of digital sensors, smart nodes, people, services and other physical objects that are capable of realizing seamless information connection, interaction and exchange [3]. The exponential growth of data has created applications in various landscapes, such as e-healthcare industry, many mobile crowdsensing systems, machine learning, or smart mobility, just to name a few. One of the issues that arises from this situation is that data owners do not know how to effectively utilize their data after collecting it: if it is not it actively used, the data remains static and it forms individual information islands. Meanwhile, other entities could thrive on it but don't have the means necessary to obtain it. The logical solution to this problem is to trade the collected data in order to maximize both personal utility for the owners and consumers. To this end, several data trading platforms such as Factual [4] or Snowflake [5] have been implemented to serve as stages where the trading takes place, but they fail to fulfil several requirements for a fair and secure data market [6]. One of the requirements that most of the existing platforms do not support is data privacy for the users (owners and consumers), a concept we discuss and define in section 2.3.

The lack of privacy in most existing data markets has sparked the interest of various researchers around the globe who have tried to come up with proposals on how to design privacy-preserving data markets that comply with modern privacy requirements coming from users and regulations, while also accomplishing other conditions (such as market fairness) and staying economically viable. Since this is a relatively new research field, the optimal combination of technologies and theoretical ground used in these proposals is still in a developing phase, which is why no proposal has established itself as the best solution.

1.2. Research objectives

In this thesis, we aim at providing an overview of the technologies used in the most relevant proposals and studies concerning privacy-preserving data markets for IoT devices. Our

goal is to summarize the state of the art of techniques that enable the design of such data marketplaces, while also providing insight on the challenges that cause the need for these techniques. Furthermore, we also address future research directions that are being adopted by the leading scientists in the field. In order to formalize our goal, we specifically aim at answering the following research questions (RQ):

1. RQ1: What are the current research gaps and problems in the implementation of privacy-preserving data markets for IoT devices?
2. RQ2: What are the research topics that have been addressed in the intersection of privacy preservation and data markets for IoT devices?
3. RQ3: Which privacy preservation techniques have been used to enable data markets for IoT devices and what are their current deployment impediments?
4. RQ4: What are the future research directions of the application of privacy-preserving techniques on data markets for IoT devices?

1.3. Research approach

In order to answer these questions, we conducted a systematic literature review (SLR) as described by the leading researchers in the field of "Evidence-Based Software Engineering" [7]. Systematic literature reviews aim at summarizing the existing evidence concerning a technology, identifying gaps in current investigation and providing background on scientific topics in order to correctly position new research activities. Our SLR was designed in order to minimize bias effects and was defined by three main phases:

Phase 1: Planning the review. Within this phase, we developed a *review protocol* that specifies the procedures, objectives and deadlines, among others, that are to be undertaken during the SLR. The research questions listed above were defined as one of the first steps during this research phase. For the procedures, we developed a study search strategy that incorporated the selection criteria for the included studies, as well as a data extraction strategy and a synthesis strategy.

Phase 2: Conducting the review. In this phase we carried out the automated search and the data synthesis following the protocols defined in the first phase. Concretely, we identified 50 studies that are relevant according to our selection criteria.

Phase 3: Reporting the review (Documentation and data visualization). The result of this phase is the finished thesis.

Moreover, before answering the RQs, we carried out a mapping study of metadata contained in the selected studies (e.g. author, publication year, publication source) in order to provide more context on the types of studies we selected as well as an overall overview of where and when the relevant research has taken place.

The thesis is structured as follows: Section 1 provides the introduction. In section 2, we present some background definitions on the concepts of privacy, data markets and IoT-devices. Section 3 provides an overview of the found related work. The details of the methodology we used for finding and synthesizing the selected studies are thoroughly described in Section 4. Regarding the results of the thesis, we conducted a mapping of the meta-data of the studies, which is presented in section 5, while the RQs are answered in section 6. The key findings and limitations of the review are discussed in section 7, and section 8 concludes the thesis.

2. Foundations

This section is dedicated to providing the theoretical foundations for the remaining of this thesis. Therefor, it mainly aims at defining domain-specific terms as well as key concepts.

2.1. Internet-of-Things

Big data is considered to be the key to unlocking the next wave of growth in productivity [1]. Nowadays, the total amount of data in the world is exploding: about 2.5 quintillion bytes of data are produced every day [8], with almost 90% of it having been created in the last two years [2]. The big data application industry is growing at around 10%, which is almost twice as fast as the traditional software field [9]. The data sources are diverse, especially as the concept of the Internet of Things (IoT) becomes more and more a part of the modern world [1]. The IoT is a term that was firstly named by the British technology pioneer Kevin Ashton, who used it as the title of a personal presentation in 1999 [10]. Since then, the definition of the IoT has evolved throughout the years, mainly because there have been many new technologies contributing to the core functionality, creating new applications for the IoT such as machine learning and real-time analytics, among others [11].

[12] defines the IoT as the ubiquitous connection of everyday objects whose dramatically increasing deployment has enabled tremendous interactions among physical objects, which brings improved efficiency, accuracy, and economic benefits while reducing human interventions. Even if this characterization generalizes all singularities of the IoT, it is one example of a definition that is still too broad for our research purposes since it doesn't describe which types of devices are "everyday objects" or which workflows are carried out.

On a more concrete characterization, [11] describes the IoT as "a system of interrelated computing devices, mechanical and digital machines, objects, animals or people that are provided with unique identifiers (UIDs) and the ability to transfer data over a network without requiring human-to-human or human-to-computer interaction". [11] states that the created ecosystem by the IoT mainly consists of web-enabled smart devices that use embedded systems, such as processors, sensors and communication hardware to collect, send and act on data they acquire from their environments. This definition is much more useful, since it identifies three essential steps in IoT workflows (data collection, data transfer and data processing).

[13] describes the IoT with a higher level of characterization, outlining the relevant technologies (and their evolving capabilities) that have enabled the modern IoT environment.

The categories of technologies that are identified in this study are RFID technology-enabled devices, wireless sensor networks, smartphones and technologies enabling the cloud computing paradigm. Furthermore, [13] also develops a reference model describing the entities and information flows of IoT applications, which are based on [14]. In this IoT model, the authors consider four main types of entities and five types of information flows. The four entities are:

1. Smart things, which are everyday things augmented with information and communication technology (ICT).
2. Subjects or recipients: Humans that have two different roles in the model: Firstly, they can be subject to data collection by the smart things surrounding them. Secondly, they can be recipients of data or services. A person in this model can also be both subject and recipient at the same time.
3. Infrastructure: smart things are connected to services via an infrastructure with different characteristics ranging from low-power lossy networks to powerful Internet backbones possibly traversing different intermediate gateways and servers.
4. Services: Backends host services that gather, combine and analyze data from many smart things to offer a value-added service to the end-user.

The five types of information flows in the model, which are nothing else but phases in the process, are:

1. The interaction phase, where the data subject actively or passively interacts with the smart things in his environment, thereby triggering a service.
2. A collection phase, where smart things then engage in the collection of information and relay it to the corresponding back-end via the available interconnection networks possibly with the help of intermediate gateways.
3. The processing phase, where backends analyze the information in order to provide the triggered service.
4. The presentation phase, where the service is provided to the data subject by the surrounding smart things according to the instructions by the backend.

In this thesis, we adopt this broader definition of IoT and will refer to these phases when describing challenges, research topics and privacy-preserving techniques within an IoT environment.

IoT devices have several technical unique characteristics that make them well-suited for some applications and inadequate for others. The most distinctive feature of IoT devices is that they are lightweight. This makes them very flexible in terms of reproducibility and costs, but limits them in aspects such as computing power. We discuss how some features of IoT devices are challenging for the implementation of privacy-preserving data markets in section

6.

For device owners, one of the issues that arises after the collection phase is that these owners do not know how to effectively utilize their data after collecting it: if it is not it actively used, the data remains static and it forms individual information islands. Meanwhile, other individuals, companies and so on, could thrive on this data but don't have the means necessary to obtain it. The logical solution to this problem is to trade the collected data in order to maximize personal utility for both owners and consumers. To this end, several data trading platforms and proposals have arisen in recent years. These platforms are the subject of discussion of the following section.

2.2. Data marketplaces

The data markets mentioned above are still in the initial stages due to the shortage of feasible protocols that ensure fulfilling the requirements set by both users and legislations [1]. Data trading creates a win-win situation for both data owners and consumers: as data is more useful if it doesn't remain static, the principle of data trading pushes the data as a dynamic flow, realizing its commercial value and providing benefits for all participating parties.

The basic principle of data markets is that a data owner (also referred to as collector or seller) makes profit by selling data, and the data consumer (also referred to as buyer) pays to obtain this data. Most models of modern data markets introduce third-party-actors such as a data broker or a platform manager, who can be either part of the data market implementation or are supplied or administered by external companies. The responsibilities of these additional parties range from administering the platform to collecting the data, identifying potential buyers and distributing (resell) it to them, mostly by requiring a transaction fee. Other common additional parties or entities are service providers, who in many cases are responsible for processing the data before transferring it out to the buyers. It becomes apparent that the names, responsibilities, roles and number of additional parties that are part of the infrastructure named in the previous section can be different for individual implementations. It also becomes apparent that the structure of these additional entities is also fundamental for the core functionality of a data market and its underlying protocols. We discuss how the arrangement of these additional components plays a crucial role in preserving privacy within some of our selected studies in section 6. As a general rule, we can summarize that IoT data markets have a data owner, a data collector and a trading platform in the middle that constitutes of a series of additional (hardware-based or software-based) components.

There are several ways to categorize existing data markets:

[15] identifies two types: In the first type, the data owner is the seller, and the data is sold on the platform managed by the owner him- or herself (this is mostly the case where

data owners are large institutions). In the second type, a third party provides the trading services, which heightens the risk of malicious attacks, a challenge discussed in section 6. [16] also divides data markets into two types from the perspective of data providers: in this study, the first type of data markets deals with independent providers, while the second type deals with multiple data providers to publish their data either for free of for a fee. [6] distinguishes between the types of data, where so-called general platforms enable data exchange of all kinds of data, while specialized platforms trade only certain kinds of data. In [17], the distinction of the data markets is based on the target audience: the first type includes data markets that target individual customers (such as smart-home owners), who might expect comfort and convenience through some kind of automation, while the second group focuses on supporting business activities by collecting and analyzing sensed data in industrial domains, the customers mostly being companies.

In this thesis, we do not limit our research by the classification type of a data market as described above, but rather by the requirement that some of the traded data is private and must therefor be correspondingly protected. The majority of our selected studies ([18] as an example) explicitly state that existing data markets (such as Factual, Dataplaza, Infochimps or Qlik) do not fulfil privacy requirements in satisfactory manner , which constitutes the central research issue in this thesis. Privacy is only one of the many requirements that data markets should ideally pursue. Other requirements may be market fairness, economic viability or legal practicability. Even though we do not focus on these requirements in this thesis, we consider that they set up the framework in which IoT data trading must take place. Figure 2.1 illustrates the general concept of data marketplaces for IoT devices.

In order to better understand the privacy requirement, the following section aims at finding an adequate definition.

2.3. Privacy

Even though the privacy requirement in data markets is legally demanded by several laws, one of them being the famous European General Data Protection Regulation (GDPR) [19] written in 2016, its foundation comes from a social state of unsettlement and uncertainty in the world's population, in particular in data owners. [20] affirms that data owners are worried that they have lost control over their information and report to be highly concerned about their informational privacy. The word "control" is key: [21] states that data owners should have full control over the terms and conditions which under the data is shared, meaning which parts, when and with whom. Likewise, [20] also expresses that people (in this case, that owners) must find the right balance between withdrawal and disclose in the given context. In a similar manner, [22] defines privacy as "the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others", mentioning that in the case of data markets, selling private

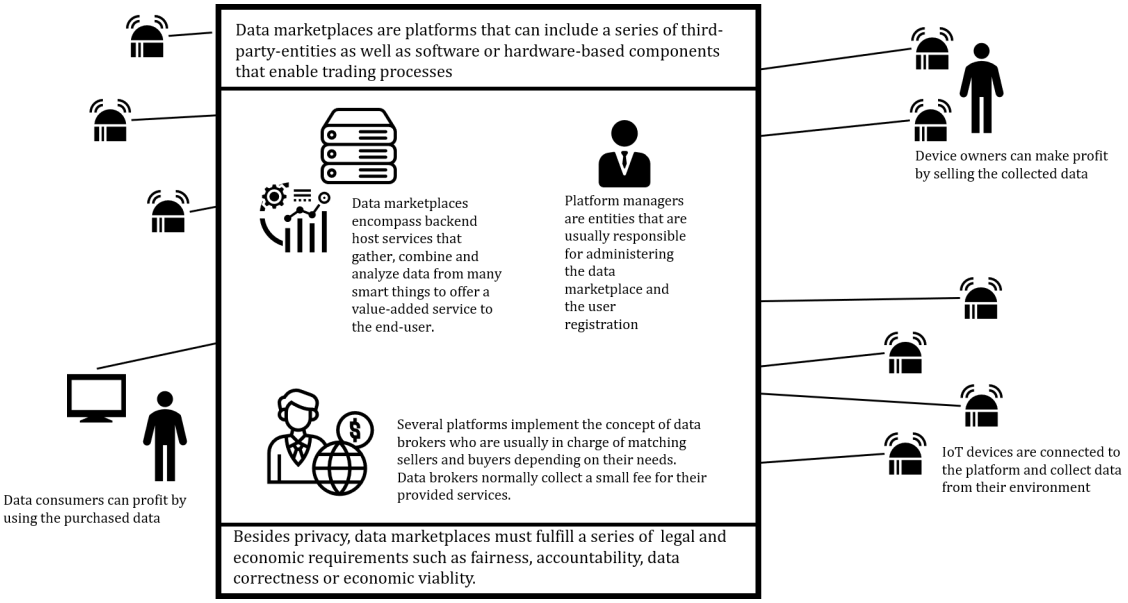

Figure 2.1.: IoT datamarkets

data does not necessarily mean losing privacy, as long as the users are informed about what is being sold. Definitions with a similar approach focusing on self-determination and control are found in [23], [24] and [13].

There are various definitions mentioned in field-relevant studies. This is because“privacy is far too vague a concept to guide adjudication and lawmaking, as abstract incantations of the importance of privacy do not fare well when pitted against more concretely stated countervailing interests” [25]. For example, [26] defines privacy as the interest of sustaining a personal space free from interference by other entities. Contradicting each other, [17] defines privacy as a human right, while [27] defines it as a commodity in the context of data markets. Within all of the found definitions, the one this thesis contemplates is illustrated by [28], a study that sets three goals in order to achieve privacy:

1. The first goal is transparency. Transparency means that all stakeholders must be informed about the data lifecycle and what happens to each data item over time. This goal must be valid for data owners, data buyers and additional entities within the trading process in order to avoid information asymmetries. In practical terms, this means that the data market platforms must inform both sellers and buyers on the data lifecycle for each transaction.

2. Intervenability, the second goal, means that data owners should be able to intervene at any time during the data lifecycle so that they can withdraw or change their consent at any time. This goal is also described by the self-determination and control focused definitions at the beginning of this section.

3. The third goal is unlinkability, which states that data from multiple sources should not be combined in such a way that together they would violate the other two goals.

One important aspect of unlinkability is to prevent re-identifiability, that is, that singular datasets that originally could not be linked with an individual or a personal identifier, must not lose this property when combined with other data sources. Since this is a central issue tackled by many of the studies identified in our SLR (section 6), we shortly describe how personal data in data markets (be it from a physical person or a device with a unique ID) can be classified concerning this particular issue. [29] categorizes personal information using three main classes:

- Explicit identifiers are data attributes that directly reveal an entity's identity, for example, the full name and the social security number of a person.
- Non-explicit identifiers are the ones that can be combined with background knowledge in order to reveal the identity, for example a zip code and a birth date.
- Sensitive attributes can be utilized to extract private information about an entity, for example, data sensed using real-time activity tracking.

The differentiation of these data types plays a crucial role in the context of IoT: [17] states that it is important to understand that different IoT solutions capture different types of data in different contexts in order to efficiently design data market protocols that fulfil the requirements of both privacy and economic-viability.

[18] states that the central problem of modern data markets regarding our definition is that they are not transparent and users are not fully aware of what happens to their data and cannot control data use. This directly confronts the first and second goals in our privacy definition. As a result of this opacity, a more or less legitimate trade of data in so-called shadow markets has evolved [20].

As a final note in this section, we mention that many studies use the self-determination and control-focused definition of privacy (see above) in order to simplify the solution to privacy as a pricing problem: since privacy means that data owners are able to determine the conditions under which their private data is sold, selecting a proper price for which users would carry-out trading accomplishes privacy in an efficient way. While this reasoning may be logical, the scope of this thesis is to provide privacy as defined above, and not to focus on an optimal pricing model that can solve the privacy requirement.

3. Related work

This section identifies and summarizes our found research relevant to privacy preservation in IoT data marketplaces. The presented related work further extends on the foundations presented in the last section. We note that several of the following papers are also part of the selected studies used in the meta-data mapping and the SLR.

[17] is very well suited as an introduction on privacy preservation in IoT data marketplaces. It is a study with a very similar goal to the one in this thesis providing a high-level analysis of privacy-preserving techniques. Another study that is fit for introductory purposes is [30]: it is a three-page long article that describes some of the challenges and research opportunities in IoT data markets in general, privacy being one of them.

[1] provides a summary on the state-of-the-art on data market research and its different components by carrying out a survey, privacy preservation being one of the components. The summary on privacy preservation is brief. One of the main chapters of this study focuses on pricing models, which is a perspective we do not discuss in detail in this thesis. Other shortly examined topics of [1] are the design of platforms, as well as some mechanisms that specifically protect digital copyright.

Another survey study is [3]. It aims at describing types of privacy breach cases and some preservation techniques that are used to counter these breaches. Similarly to [1], this study does not go into a deep level of detail when describing privacy-preserving techniques. [3] also includes a rigorous outline on the balance between privacy preservation and data analysis from a service provider's perspective. Furthermore, it reviews several relevant aspects in pricing procedures as well as game theoretical approaches and auction schemes used in data markets.

Contrary to this thesis, [13] focuses more on (classifying) the challenges that come along with privacy-preserving IoT data markets rather than the solutions to overcome them. The structure for classifying these challenges is partly used in our SLR in section 6. Moreover, [13] also provides an insight on how different laws have protected privacy before the well-known GDPR.

Furthermore, [20] analyzes internet users' preferences for privacy in data sharing in order to uncover mental models of these preferences and their motives, barriers and conditions for a privacy-preserving datamarket. In this study, the focus lies on a self-determination and control-based definition of privacy, and not so much on other features such as re-identification

or pricing.

There are also several studies that enquire research questions similar to the ones in this thesis while being more focused on a particular use case or industry:

[31] surveys privacy-preserving methods for the crowdsourcing use-case, while also illustrating essential challenges that must be tackled in future research. The authors state that the main responsibility for future work is defining a management paradigm that is based on crowdsourcing application requirements, recruiting qualified participants, distribute tasks and coordinating with participants until task completion. Therefor, the focus of this paper lies on the task management function of trading platforms and the countermeasures that task management can implement regarding privacy issues. Furthermore, the study defines major dimensions (classes) of crowdsourcing and propose three techniques to mitigate privacy challenges.

[32] provides an outline of privacy approaches and methods such as homomorphic encryption or anonymization in order to protect a person's privacy in the health industry with the example of an application that collects private data. The privacy-preservation methods are divided into two scenarios: outsourced computation, which relates to everything that happens to users' data outside the perimeter of their personal devices and home networks, and information sharing, which relates to situations where parties have to contribute to gain information from one or more data sources. The pros and cons of each method in both scenarios are described. We mention these scenarios in section 6 when discussing privacy preservation techniques' applications.

[33] is a study that is very relevant to this thesis in terms of structuring. It illustrates inter-organizational dataflows occurring in the Internet of Production (IoP) and identifies several security and privacy demands and challenges that originate within these dataflows. Furthermore, the study provides a survey of technical building blocks to meet the requirements these challenges pose and propose next steps for research. In this thesis, we use the proposed building blocks as a basis for our structure in section 6 and modify these blocks according to our findings.

Finally, there are other studies that do not conduct an analysis of privacy preserving techniques or challenges, but provide a set of guidelines and general instructions that must be followed in order to preserve privacy in IoT data markets. For example, [34] states that designing IoT applications is much more difficult that designing desktop, mobile or web applications and proposes therefor a set of detailed guidelines that are aimed to be used by software engineers to develop privacy-preserving IoT applications. The guidelines can be used for end-user applications as well as middleware programs and follow the "privacy by design" approach, discussed in section 6.3.4.

In conclusion, we identify that even though there are many studies that focus on providing information on privacy preservation in IoT data markets, there is not one literature review that can summarize everything we aim at in this thesis: Firstly, we did not find a single study that approaches these questions using the rigorousness of Evidence Based Software Engineering (EBSE): Neither of the found studies conduct systematic literature reviews or systematic mapping studies in order to increase the level of academic strictness and reduce bias as defined by relevant literature [7]. Secondly, leaving the used methodology aside, none of the found studies provide a comprehensive analysis of current challenges, topics, technical solutions and future directions that arise in the implementation of modern privacy-preserving data markets. The lack of these findings being summarized into one study could slow down potential research and future implementations.

4. Methodology

This chapter describes in detail the combination of methods and guidelines we used to accomplish our research goals.

Kitchenham et al. firstly presented a formal framework to conduct a Systematic Literature Review (SLR) within a research paradigm called Evidence-Based Software Engineering in 2004 [35]. Amongst other, this framework included a thorough description of the main phases that are to be run in order to successfully conduct SLR and so enhance the rigor in future software engineering research. The proposed guidelines were then improved by the same authors in 2007 [36].

In broad terms, the methodology consists of three main phases, which include several activities. These phases are *planning the review, conducting the review and reporting the review*:

Phase 1: Planning the review. The result of this phase is a *review protocol document* that specifies the procedures, objectives and deadlines, among others, that are to be undertaken during a SLR. This phase encompasses the following activities:

1. *Identification of the need for a review*: The need for SLRs emerges from a necessity of researchers to summarize the state of the art or the existing knowledge about a particular topic or area. Prior to carrying-out a SLR, researchers should ensure that such a study is indeed required. In our case, we found no studies that provide a comprehensive analysis of current challenges, topics,technical solutions and future directions that arise in the implementation of modern privacy-preserving data markets, as mentioned in section 3.
2. *Specifying the research question(s) (RQ)*: This is the most important task in a SLR. The RQs define the rest of the methodology as well as the quality and relevance of found papers.
3. *Development of a review protocol* that includes:
    - Background or rationale for the survey
    - The research questions previously specified
    - Search strategy (including search strings and resources)

- Selection criteria and procedures: The set of criteria used in the search phase (as well as the corresponding processes) in order to define if the found publications are relevant for the thesis.
- Data extraction strategy: This point describes the procedures used to extract data from individual publications into so-called extraction forms.
- Synthesis strategy: This aspect refers to the planned practices used to summarize and combine the extracted data.
- Project timetable

4. *Evaluating the review protocol* (optional)

Phase 2: Conducting the review. The outcome of this phase encompasses the results of the automated search, the study selection and the data synthesis. Each activity of this phase has its corresponding part in the review protocol:

1. Identification of research
2. Selection of primary studies
3. Study quality assessment as well as Data extraction and monitoring
4. Data synthesis

Phase 3: Reporting the review (Documentation and data visualization). The result of this phase is the finished thesis. The included activities are:

1. Specifying dissemination mechanisms
2. Results visualization
3. Formatting the main report
4. Evaluating the report (optional)

In this thesis, we base our methodology on this framework in order to answer the RQs listed in section 1 and again in section 4.2.1. An overview of our research workflow is depicted in Figure 4.1.

As mentioned in the introduction (section 1), we carry-out a SLR in order to address our research goals. Additionally, we also realize a short meta-data mapping study in order to provide more context on the types of studies we selected as well as an overall overview of where, when and how the relevant research has taken place.

The following section outlines the origins and characteristics of SLRs compared to other study types. After that, we depict our definitions for the first phase listed above (planning the review) by constructing a review protocol. Furthermore, for each step defined in the protocol, we also describe how the second phase (conducting the review) was specifically carried out in order to answer our RQs.

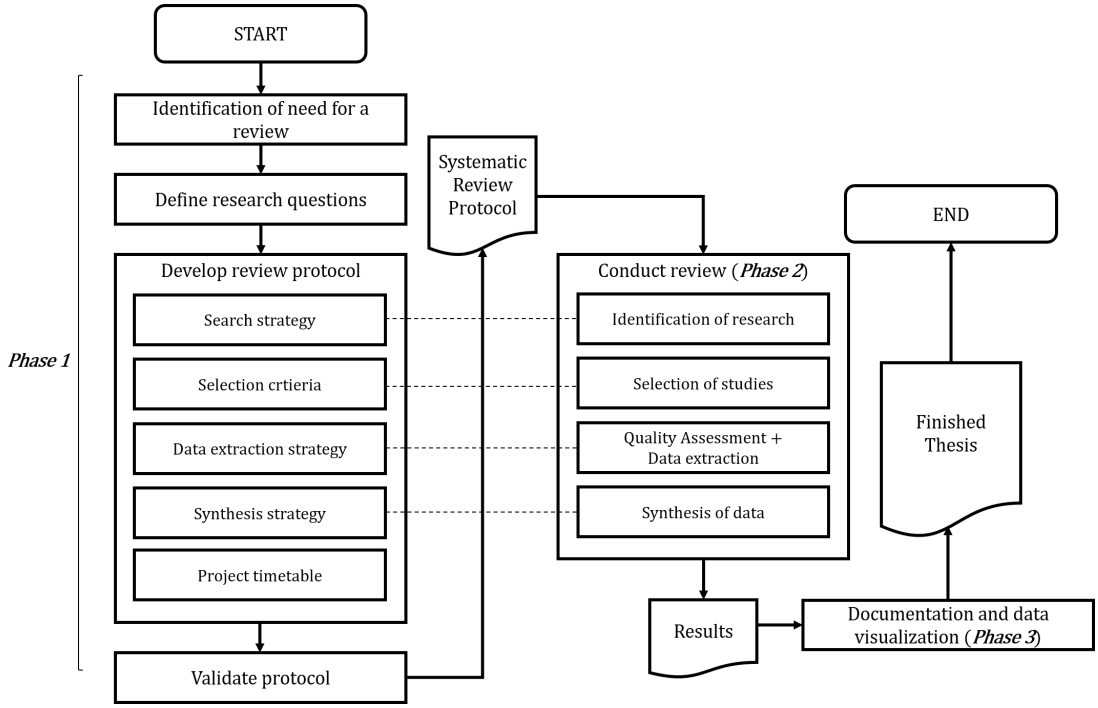

Figure 4.1.: Overview of methodology workflow

4.1. Background

4.1.1. Systematic Literature Reviews

SLRs have their origins in the medical field. Kitchenham et al. firstly suggested that the guidelines for evidence-based research used in medicine could also be transferred to software engineering [35], thus laying the foundations of what we now call Evidence-Based Software Engineering (EBSE). The need for this approach was derived from the lack of empirical research in the software engineering field compared to other sciences, which resulted in a shortage of rigor in many cases [37]. The concepts of EBSE were firstly put in practice with SLRs [36], which aimed at:

- Summarizing the existing evidence concerning a technology.
- Identifying gaps in the current investigation in order to point to certain areas for future research.
- Providing a framework or background in order to correctly position new research activities.

all of this while using a trustworthy, rigorous and auditable methodology that can be reviewed and replicated in a fair manner [35]. SLRs are referred to as "secondary studies", and the studies they analyze are called "primary studies". In contrast to conventional (non-evidence-based) reviews, SLRs have the following features [36] [38]:

- They start by defining a strict review protocol that specifies the research questions and the precise methodology that will be used.
- They define and document a search strategy that aims to detect as much of the relevant literature as possible.
- They require explicit inclusion and exclusion criteria to assess all of the potential relevant primary studies.
- Both SLRs and conventional (narrative) reviews are retrospective, observational studies and are therefore subject to systematic and random error. An example of a common error is the author's bias. SLRs focus on minimizing the error margin caused by this phenomena.

This thesis uses many of the guidelines for SLRs, including a protocol framework, in order to plan the review process and address our RQs.

4.1.2. Systematic Mapping Studies

This thesis also presents a short Systematic Mapping Study (SMS) based on the research questions. SMSs (also called scoping studies) are similar to SLRs because they take advantage of its methodology [39]. Nevertheless, there are some differences between both study types. The main difference is that a conventional systematic review makes an attempt to aggregate the primary studies in terms of the research outcomes and investigates whether those research outcomes are consistent or contradictory, whereas a mapping study usually aims at "only" classifying the relevant literature and aggregate studies with respect to previously defined categories [40]. In other words, a SMS is a form of SLR that intends to 'map out' the results from a more general perspective rather than answering a narrow research question. Even if the discrepancy between both study types is somewhat fuzzy [41], we find some concrete differences that can help us choose the particular guidelines we may prefer [40] [42]:

- **Research question**: In SLRs, the research question tends to be more specific and is related to the outcomes of empirical studies. It is of the form: 'Is technology/method A better or not than B?'. In SMSs, the research question has a more general character and is related to research trends (e.g. main researchers, activity focus, types of studies).
- **Breadth and Depth**: In a SMS, a larger number of articles can be considered as they don't have to be analyzed in such detail. Thus, a larger field can be structured. This characteristic is reflected in the search strings and inclusion/exclusion criteria that is used to filter out relevant papers.
- **Search process**: In SLRs, the search process is defined by the research question, whereas in SMSs, it's mainly defined by the topic area.
- **Process stages**: The early stages of a mapping study are generally very similar to those of SLRs. In later stages, where SLRs concentrate on aggregating the extracted

data, the analysis of SMSs is more concerned with classification of the available studies (meta-analysis instead of a thematic analysis). The level of extraction is therefor different.

In this thesis, we decided to also carry-out a mapping using the meta-data of the selected studies we use for the SLR.

One important general advantage of SMSs is that they are very well-suited for visually summarizing results in a map [39]. Furthermore, from an educational perspective, SMSs can provide an excellent starting point for students' projects and theses if the question of classification is reasonably tractable [40]. For these two reasons, we use some of the princples of SMSs and aim at providing more context about the studies that we found before presenting the results of the SLR, which constitute the main part of this thesis.

The following chapter provides a detailed description of the protocol used in order to select and synthesize the relevant studies.

4.2. Phase 1: Planning the review

This section describes in detail the activities and steps used in order to generate a *review protocol document*, which is the result of the first phase. The protocol specifies the procedures, objectives and deadlines, among others, that are to be undertaken during our SLR. The discussed activities are illustrated in figure 4.2

We abstain from describing the first and the last step in detail, since sections 2 (foundations) and 3 (related work) already disclose the need for a review, and the protocol validation was done informally.

4.2.1. Research questions

The research questions are the most important part of any SLR [36]. They are the base for the used data sources, the search terms, the inclusion and exclusion criteria and the report standards [43]. According to Budgen et al. [7], a good research question is one that

- Can be purposeful both for practitioners as well as researchers (even though SLRs are mainly aimed at further research activities).
- Will lead to changes in current Software engineering practices or will increase confidence in the value of current practices or beliefs.
- Identifies discrepancies between commonly held beliefs of a technology and the actual reality of these technologies.

Based on these criteria as well as the initial trigger for this thesis (the need for an overview of privacy preservation in data markets for IoT devices), we defined the following four research questions:

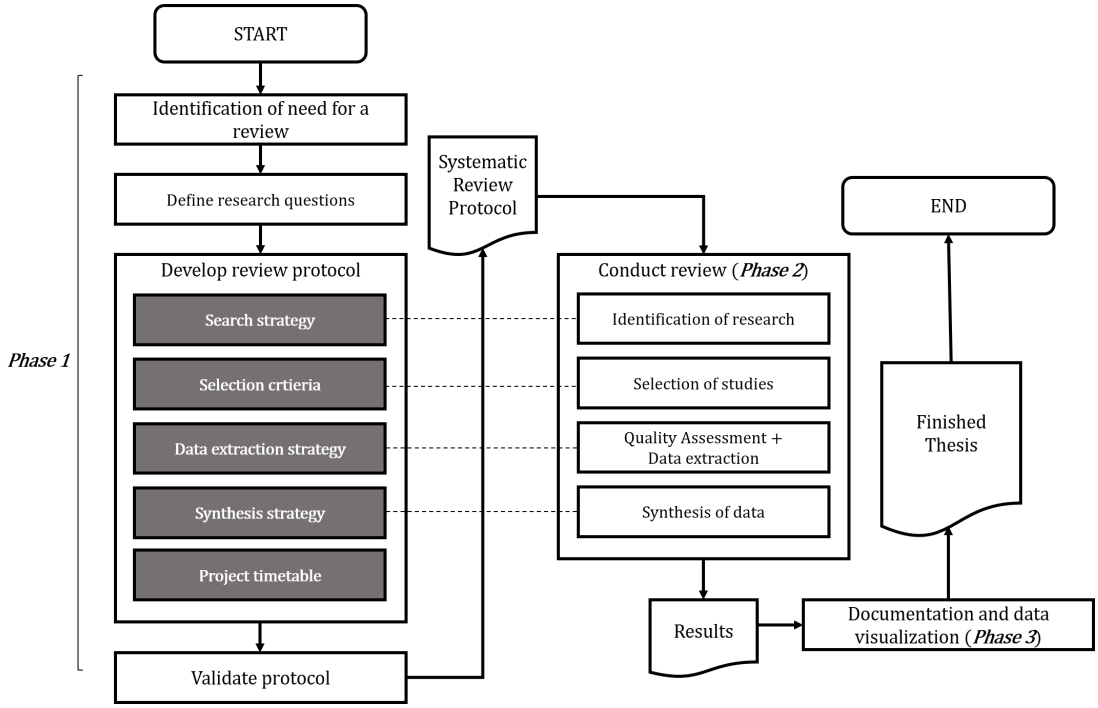

Figure 4.2.: Planning the review: relevant steps covered in this section

1. *RQ1: What are the current research gaps and problems in the implementation of privacy-preserving data markets for IoT devices?*

   The purpose of this question is to present an analysis on current problems and research gaps that still haven't been resolved in order to ensure absolute privacy preservation in data markets for IoT devices. We expect these problems to be of different different types (e.g. technical, logical, economical) and will aim at classifying them in a structured manner.

2. *RQ2: What are the research topics that have been addressed in the intersection of privacy and data markets for IoT devices?*

   The expected result of this question is to provide an overview of the topics that are being investigated in research concerning privacy enablement in data markets for IoT devices. This should list the core topics specifically addressing data markets for IoT devices as well as more general areas that may also be of use for other applications (e.g. universal architectural approaches). We expect that the topics can be clustered into a small amount of categories in order to determine few future directions.

3. *RQ3: Which privacy preservation techniques have been used to enable data markets for IoT devices and what are their current deployment impediments?*

   The goal of this question is to determine the state of the art in concrete techniques that have already been successfully implemented in data markets for IoT devices. We expect a series of techniques that are used depending on the requirements of individual data markets.

4. *RQ4: What are the future research directions of the application of privacy-preservation techniques on data markets for IoT devices?*

   Using the findings of the previous questions, this one aims at providing an outlook for future research. We intend to be able to directly link the found problems (see first RQ) to future research directions.

As mentioned above, the content of these questions would affect not only the search strings used to find relevant primary studies, but the research approach and the overall strategy.

4.2.2. Search strategy

The goal of a search strategy is to devise a way of finding as many primary studies as possible. These studies must be relevant regarding the research questions whilst considering the constraints of time and human resources [7]. The search strategy describes and justifies the way in which specific searching methods, such as automated and manual searching of papers, are combined. Most SLRs involve a combination of search methods [38]. For our purposes, the strategy concretely makes use of a series of studies that comprise the manual search results, as well as the search strings and databases for the automated research. Moreover, we explain why we selected those precise resources and approaches.

The end of this section contains a summary of the used search strategy.

Measuring the completeness of the results

In order to be able to evaluate its quality, this thesis aimed at achieving an acceptable level of *completeness* as described by Budgen et al. [7] and Zhang et al. [44]. Our search strategy is mainly based on these two sources. The term of completeness refers to the question: *How complete should the set of papers be and how will we know whether we have achieved this target?* In general terms, there are some contextual differences based on the research type:

- For quantitative reviews, completeness plays a much more crucial role than for qualitative reviews. Based on the research questions, this thesis is of a qualitative nature.

- Tertiary studies should have a high level of completeness if they are aimed at being the key-reference document for the community (not our case because ours is a secondary study).

Since our knowledge on the research area was limited in terms of knowing the relevant papers, we used the completeness metrics proposed by Zhang et al. [44]:

- **Sensitivity** (also called *recall*) is the proportion of all relevant studies that are found by an automated search.
- **Precision** is the proportion of the studies found that are relevant to the research questions being addressed.

Both metrics are calculated as follows:

$$Sensitivity = \frac{R_{found}}{R_{total}} \tag{4.1}$$

$$Precision = \frac{R_{found}}{N_{total}} \tag{4.2}$$

Where:

R_{found} is the number of relevant studies found
N_{total} is the total number of studies found
R_{total} is the total number of relevant studies (found or not found)

It goes without saying that both high sensitivity as well as high precision are desirable. Since the number R_{found} is not a constant that can be categorically be found for literature reviews (there is no way of knowing exactly how many studies in this world are relevant for a particular topic), calculating the (real) recall is impossible. For this matter, Zhang et al. [44] suggest a series of mechanisms to find the relevant studies that will later be used as an artificial R_{total}. In order to find our base set of papers that would adopt this role, we decided to combine these mechanisms by:

- Conducting a preliminary manual search in the TUM-OPAC System (the search engine from the TUM library). The string we used for this search was *"privacy data markets IoT"*. The selection criteria for this manual search was the same as for the automated search (see section 4.2.3).
- Include the papers that we had selected prior to starting this thesis.

The resulting papers of these three activities would result in the R_{total} that we would use for the final evaluation of sensitivity and precision levels. In the following, we call this set of literature *base literature*. For our case, the base literature consisted of eight studies. In order

to know which studies were part of the base literature, we provide an overview table in the Appendix. Overall, we aimed at achieving a sensitivity level of 70% as recommended by [44]. If the automated research wouldn't reach these levels, the search strings would be refined until doing so, resulting in an iterative search process, which is further summarized below.

Selection of sources for automated search

Within automated searches, there are two main decisions that have to be made: the first one is deciding on the sources that will be searched. Appropriate sources include publisher's sites (PS) and general index engines (IE), whilst a mix of the two types is best [7]. The difference between both is that IE mainly index the work published by various publishers, whilst PS refer to the online literature databases provided by the publishers to facilitate easy retrieval of the published literature [45]. We decided to use seven electronic data sources (EDS) for conducting the main literature search based on the following criteria:

- The EDS are either mainly focused on Software Engineering or they are for general purposes but have a large enough number of papers so that each topic is thoroughly represented.
- The EDS are listed as the most used by SE researchers in a study carried-out by Chen et al. in 2010 [45].
- Access to the EDS is granted for TUM students.

Moreover, we decided not to use *Google Scholar* as an EDS because, although it does identify unpublished material, it is often not possible to find a reliable source document that can be cited correctly and is guaranteed to remain publicly available [7]. We also did not include *CiteSeerX* as an EDS because, even if it contains an important set of papers, we found that the search filtering options were too limited and thus inaccurate for our purposes. Table 4.1 lists the EDS we chose for conducting our automated research.

Table 4.1.: EDS used in automated search.

ID	Name (Acronym)	EDS Type
EDS1	IEEE Xplore (IEEE)	PS
EDS2	ACM Digital Library (ACM)	PS
EDS3	ISI Web of Science (WoS)	IE
EDS4	ScienceDirect (SD)	PS
EDS5	SpringerLink (SL)	PS
EDS6	Wiley InterScience (WIS)	PS
EDS7	SCOPUS (SCOPUS)	IE

Search terms

The second important decision that has to be made within automated searches is specifying the search strings that will be used. One problem with constructing the search strings is that terminology in software engineering is neither well-defined nor stable, making it difficult to identify reliable keywords [7]. Another problem is that EDS have different limitations and conditions regarding the complexity and structure of a search string [38]. We chose our initial search strings based on the following criteria:

- We included the central keywords of the research questions.
- We included central keywords found intuitively in the titles and abstracts of the base literature.
- We used the tool *Sketch Engine* in order to analyze and include further keywords and phrases based on the full texts of our base literature. The Sketch Engine is a leading corpus tool which has been widely used in lexicography [46] and for which TUM students have full access. The tool features that we used were the following: Firstly, we used the feature "Keywords", which lists multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, matches the typical format of terminology in the language. Secondly, we used the "Thesaurus" feature which generates synonyms for selected word based on the context in which these words appear in the corpus. Thirdly, we used the feature "Wordlist", which lists the frequency of all words (brute force). Finally, we used the feature "Word Sketch", which summarizes the words behavior in the corpus. Further details on WordSketch are sound in the Appendix.
- We used common Buzzwords in IoT found in an article by Riverbed Technology [47] as well as the ones we found in [48].
- We included technical synonyms for our keywords.
- We used three categories of search terms following the research focus. The first category includes terms concerning privacy enhancement, the second category includes terms involved with data markets and the third category includes terms regarding IoT and mobility topics.

Using these criteria, the initial strings in the search categories were:

C1: *privacy OR private OR encryption OR encrypted OR encrypt OR data protection*

C2: *data market OR data marketplace OR data trading OR data broker OR data trader OR data auction*

C3: *Internet of Things OR Internet of Everything OR IoT OR Sensor OR Connected Devices OR Networked Devices OR Smart Devices OR Controller OR Edge Computing OR Cloud Infrastructure OR Machine to Machine OR M2M OR Web-of-Things OR WoT OR Mobility OR Automotive OR Vehicle OR Car OR Automobile OR Industry 4.0 OR Smart Grids OR V2V OR IIoT OR machine learning OR mobile OR cyber-physical OR microservice OR microcontroller OR micro-service OR micro-controller OR blockchain OR neural network OR smart learning OR automated driving OR autonomous driving OR smart city OR smart factory*

The three search categories were combined by using the Boolean "AND" operator, which entails that an article had to include (at least) one term of each of the three categories to be retrieved. In other words, we searched:

C1 *AND* **C2** *AND* **C3**

We omitted plurals, different verb tenses as well as British and American spelling variations, since the used search engines automatically looked for these. As mentioned in the previous section, these search strings should be refined and/or complemented in an iterative manner depending on the search results and the achieved levels of sensitivity and precision. Also, due to query limitations in some search engines, we had to adapt the search strings accordingly by using the keywords we found most relevant.

Overall search strategy overview

The search strategy used for this thesis can be summarized with the following steps using the framework of Zhang et al. [44]. These steps can also be seen in the flowchart depicted in figure 4.3 [49].

1. Identifying electronic data sources EDS and other venues for preliminary and automated search.
2. Establishing base literature with the preliminary search in order to perform completeness checks. This base literature depends directly on the research questions (Zhang et al. use the term *quasi-gold-standard* instead of base literature).
3. Define or elicit search strings.
4. Conduct the automated research using the identified EDS and the previously defined selection criteria (see section 4.2.3).
5. Evaluate the search performance calculating the sensitivity of our results. If the sensitivity level is lower than the desired one, go back to step three.
6. After reaching the desired sensitivity level, add additional papers to the selected studies using useful references found in the main search. We decided to include this step -

which was not described in the methodology literature - in order to further increase the quality of our SLR.

7. End search and proceed with data extraction. The data extraction strategy is not part of the search strategy and can be found on section 4.2.4.

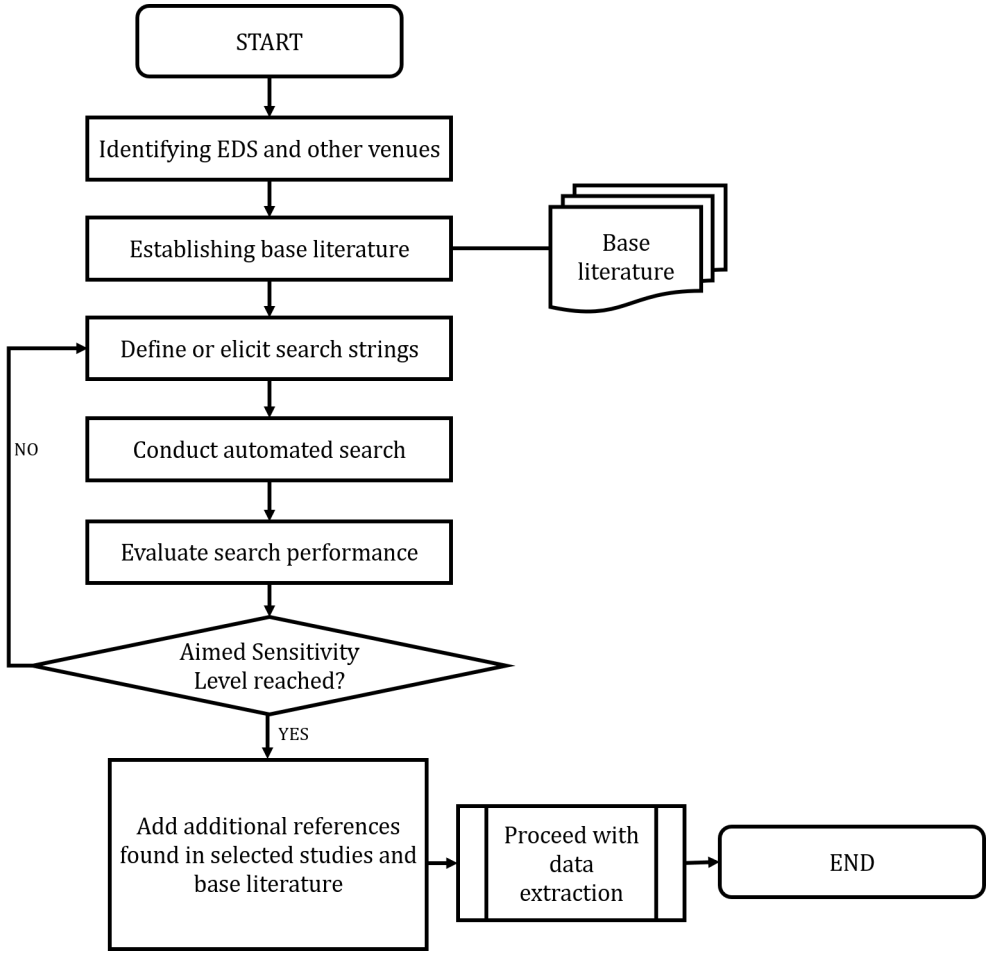

Figure 4.3.: Overview of the search strategy

4.2.3. Selection criteria and selection process

The search process in various EDS results in a set of papers that are candidates for being relevant to the research questions. These candidate papers must pass a series of criteria, which are the focus of this section. There are different levels of effort required in the numerous criteria, which is why the ones with less effort required are checked earlier: the first criteria (such as language detection) are designed to exclude papers that are clearly irrelevant, whilst the more advanced ones require reading of abstracts or even the full text in order to assert the

relevance of the inquired paper. We expressed the criteria in two sets: one for inclusion and one for exclusion criteria. Based on a SMS by Philipp [39], we defined seven facets with their corresponding inclusion and exclusion criteria. In some EDS, the inclusion criteria could be given as a search-parameter.

Facet 1 (F1): Coarse focus

A publication is only considered relevant if it primarily lays focus on privacy enhancement research for data markets. The topic must also be within the field of computer science and technology. Any other papers whose main field is either of legal, philosophical, moral or economical nature, would not be considered. We note that several of our studies included perspectives that were not solely technology focused. These papers would also be included if the spotlight is aimed at computer science technologies, more concretely on privacy-enhancing technologies for data markets.

Facet 2 (F2): Narrow focus

The paper must explicitly focus on privacy within data marketplaces or its underlying foundations, such as their technology, platform, data flow or other scientific knowledge that can be applied in the following contexts:

- Use in the framework of the Internet of Things.
- Employment for mobility technologies.
- General industry applications.

As mentioned in section 2.3, we do not focus on methodologies that define privacy protection as providing an optimal pricing procedure in order for data owners to be willing to disclose their privacy. Thus, studies that were mainly pricing-focused were also excluded from the selection.

Facet 3 (F3): Publication channel type

The found source must be either a conference publication, a journal publication or a workshop publication. Given the nature of our RQs, we did not expect many studies from workshops, but did not explicitly exclude them.

Facet 4 (F4): Language

Only publications in the English language were included. This is because it is the only language fluently spoken by everyone involved in the research project. This selection criteria also limits the effort whilst constructing search strings and extracting the relevant data.

Facet 5 (F5): Duplicates

Duplicates identified in the automated search were excluded. They can ocurr and are very likely to happen since the EDS partially deliver the same publications, as shown by Chen et al. [45]. By excluding the duplicates the time required for the filtering process can be considerably reduced. Duplicates between the automated search and the base literature are kept in order to perform the completeness check at the end of the selection phase.

Facet 6 (F6): Peer-reviewed publications

We only included publications that were subject to a formal peer-review, leaving out grey publications and thus enhancing the results' quality. This is the reason why we excluded EDS such as Google Scholar.

Facet 7 (F7): Full-text access (open access)

We only included publications that could be accessed with the student's rights provided by the TUM.

An overview of the selection criteria is displayed in table 4.2.

Furthermore, we decided to allow any publication date as well as any number of pages in the found sources. The facets F3, F4, F6 and F7 comprise the early stages of the relevance check since their criteria can be verified very quickly (they are inherent in the search). Therefor, these facets were filtered first. Facet F5 would be filtered next, since it includes the removal of duplicates. Lastly, we filtered facets F1 and F2, since they are associated with the content of the papers and need dedicated reading in order to evaluate their relevance. Because of their nature, these two facets (F1 and F2) were reviewed in three steps in order to assess if the corresponding papers were relevant:

1. At first we filtered the papers' content based on the **title**. This step could very quickly exclude numerous papers that evidently weren't suited for our research. This point involves the facet F1 since it's a coarse-grained screening step.

2. The next stage was to filter the papers based on their **abstract** or executive summary. Some abstracts give a very clear indication on the research purposes of a paper, whilst others can still be relatively vague for us to determine if a paper is relevant or not, which is why a third screening step is needed.

3. The last filter would be reading the **whole paper** in order to definitively know if it is relevant for the research questions.

These three steps would be performed by both scientists carrying-out the research in order to minimize the bias effect. In case of conflicts regarding the relevance decision, both scientists would resolve these conflicts in a personal meeting until an agreement is reached. The overall selection process is depicted in figure 4.4.

Table 4.2.: Selection criteria used to find relevant papers

ID	Facet	Inclusion criteria	Exclusion criteria
F1	Coarse focus	The privacy and data market topic must be within the field of computer science and technology.	Any other privacy and data market sub-field.
F2	Narrow focus	The paper must explicitly focus on privacy within data marketplaces within the defined applications.	The paper does not explicitly address this research direction.
F3	Publication channel type	Conference publication OR journal publication (full text) OR workshop publication	The paper is any other type of publication.
F4	Language	English.	Non-english.
F5	Duplicates	Publications are unknown to the filtering process.	Publication has been already processed.
F6	Peer-review	The publication has been peer-reviewed.	The publication is a grey publication.
F7	Full-text access	TUM-Access granted	TUM-Access not granted

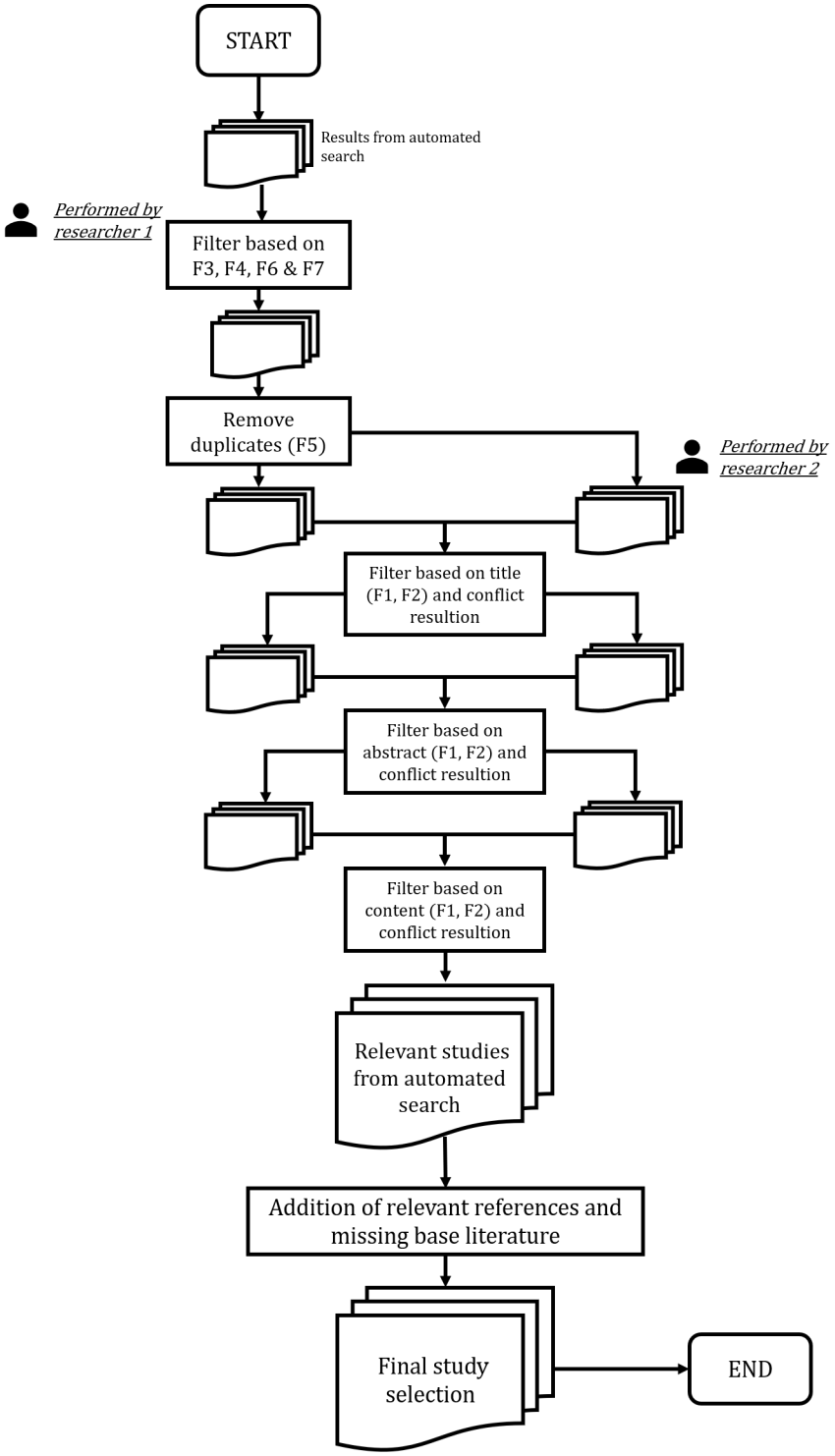

Figure 4.4.: Overall selection process

4.2.4. Data extraction strategy and synthesis strategy

SLRs can be of a qualitative or a quantitative nature. Given our RQs, this thesis presents a qualitative review. [7] notes that qualitative systematic reviews are more susceptible to bias and the data may be less amenable to statistical analysis, which is why data extraction and synthesis are very closely linked and should be combined within a single process. For this reason, we describe both strategies in this section.

The extraction of data was done in Microsoft Word documents using data extraction cards. These cards were designed to extract all the information needed for the SLR as well as the data-mapping study. For this, the most important data types and information classes were previously defined. In general terms, the extraction cards had two parts: the first part included meta-data that was later used in our mapping (section 5), while the second part extracted information needed in order to answer the RQs. Concretely, the first part - the meta-data - included the points listed in Table 4.3. The second part - the relevant data for our SLR - included the points listed in Table 4.4.

Our synthesis method is very similar to the *Narrative synthesis* method as described by [7]. We adapted this description in order to define three major steps:

1. Developing a preliminary synthesis of the findings of the primary studies.
2. Exploring relationships in the data.
3. Improving the preliminary synthesis and return to point 2 until the RQs are answered in a satisfactory manner.

4.2.5. Project timetable

The thesis was aimed at being finished in a time spectrum of 26 weeks, from which most of the effort was invested in the study selection and data collection steps. The timetable was highly oriented at the methodology workflow presented at the beginning of this chapter.

4.3. Phase 2: Conducting the review

Following our overall workflow (see Figure 4.1, we carried out the individual steps of Phase Two (Conducting the review) in accordance with the procedures defined in the review protocol (Phase One). This section summarizes our experience in the second phase. The discussed activities are illustrated in figure 4.6.

4.3.1. Identification of research

The first step in the second phase of our research was to identify a large set of papers on the research topic in order to achieve high levels of sensitivity and precision. For this, we used the search string and the seven EDS defined in Phase One (see 4.2.2).

Table 4.3.: Data extraction cards: extracted meta-data

Field name	Description
Name of the study	This point included the name of the study
Author(s)	This point included the author or authors of the study
Cite count	The cite count for each study was rechecked on the 10.09.2020
Publication year of the study	This point included the publication year of the examined study
Country or countries	This point included the countries in which the institutions where the authors of the study conducted their research are located
Found in EDS	Several studies were found in more than one EDS
Publication channel type	as defined in the previous section, the publication channel types we considered for the selected studies were journals, conferences and workshops
Publication channel name	This point reveals in which journal, conference or workshop this study was published or presented
Publication source	This point describes the institutions in which the studies were written and/or conducted (e.g. universities or private organizations)
Research type	This point was based on [39]. The three categories Evaluation research, Philosophical paper and Solution proposal are described in Table 4.5
Research approach / research methodology	This point was based on [39]. The classification scheme for these categories is presented in Table 4.6.
Research contribution	This point was based on [39]. The classification scheme for these categories is presented in Table 4.7.
Specific branch	This point described in the study was only focused on one branch (e.g. medical, automobile, etc.) or the ideas were branch-agnostic.

Table 4.4.: Data extraction cards: information for SLR

Field name	Description
Main topic	We used this point in order to know in one or two sentences what the paper is about.
Privacy definition	We included definitions of privacy as described in the paper in order to potentially use them in section 2.
Definition of data markets	We included definitions of data markets as described in the paper in order to potentially use them in section 2.
Study goal	This point was used to describe the current situation as presented by the study, the issues that are to be tackled in the study as well as how they are meant to be approached. This information was essential in order to answer RQ1.
Research questions	If available, this point refers to RQs explicitly defined in the study.
Study findings	In this point we summarized the findings and solutions to the problems and goals defined above. The information was essential in order to answer RQ2.
Privacy preserving logic	This point was meant to serve as a concretization of how the privacy was preserved in the proposed solution, if available.
Concrete techniques and technologies	This point listed the techniques and technologies used in order to preserve privacy in the study. The information was essential in order to answer RQ3.
Challenges	This point summarized the challenges that arose from the privacy-preserving techniques. The information was important in order to answer RQ3.
Future work	Here we described the future work that the authors explicitly mentioned. This point was important in order to answer RQ4.
Reason why this paper should be included	In this point, we briefly wrote why the paper is relevant for our findings.

Table 4.5.: Classification scheme of research types as described by [39]

Research type	Description
Evaluation research	Techniques are used in companies, and an evaluation of the technique is carried out. Therefore, it is shown how the technique is implemented in practice (solution implementation) and which effects the implementation has with regard to benefits and drawbacks (implementation evaluation). This also involves the identification of problems in the industry.
Philosophical papers	These papers present a new perspective on existent things by organizing the domain into a taxonomy or into a conceptual framework.
Solution proposal	There is a proposed solution to a problem, the solution can be either novel or a significantly enhanced version of an existent technique. The benefit and applicability of the solution is demonstrated by a small example or argumentation.

Table 4.6.: Classification scheme of research approaches as described by [39]

Research approach	Description
Survey	If the study collects quantitative and/or qualitative data with the help of a questionnaire or interviews.
Design and creation	The development of a new IT product or artifact, or new model/method.
Case study	If one of the following conditions is fulfilled: 1) The study states one or more research questions and some of them or all are answered with a case study. 2) The study evaluates a theoretical concept empirically by implementing it in a case study.
Theoretical	If the study is of a theoretical nature but not explicitly describes that a grounded theory approach was used.
Not applicable	Either if study does not define the applied research method or and it cannot be inferred or interpreted from study.

Table 4.7.: Classification scheme of research contribution types as described by [39]

Contribution type	Description
Model	Presentation of an observed reality with the help of concepts or related concepts resulting from a conceptualization process.
Theory	Establishing relationships between cause and effect from determined results.
Framework or methods	All Models that relate to the design of software or to the management of development processes.
Lessons learned	List of outcomes, which is analyzed directly from the research results obtained.
Guidelines	A list of advice represents a summary of the research results obtained.

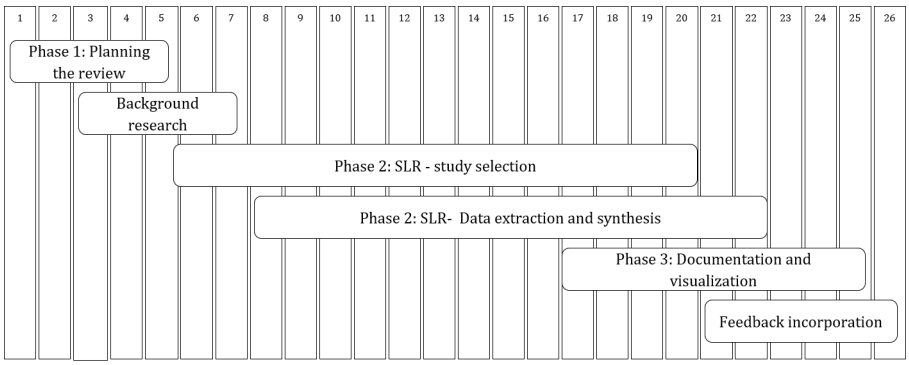

Figure 4.5.: Project timetable

Overall, we found 1155 studies that would later be filtered in the second step 4.3.2). We note that SCOPUS had ScienceDirect (SD) was the only EDS where we were not able to use the complete Search String defined in the protocol definition because it only allowed 8 boolean operators per search. Our defined string had a total of 49 operators (including the AND operators), which is why the search on ScienceDirect was conducted with the reduced search string shown below. Since the third set of search terms was initially the biggest one (see 4.2.2, we decided to reduce the first two search sets to 7 terms, leaving space for one more search term of the third set:

1. privacy OR encryption OR encrypted OR data protection
2. data market OR data marketplace OR data trading
3. Internet of Things OR Internet of Everything OR IoT OR Sensor OR Connected Devices OR Networked Devices OR Smart Devices OR Controller OR Edge Computing OR

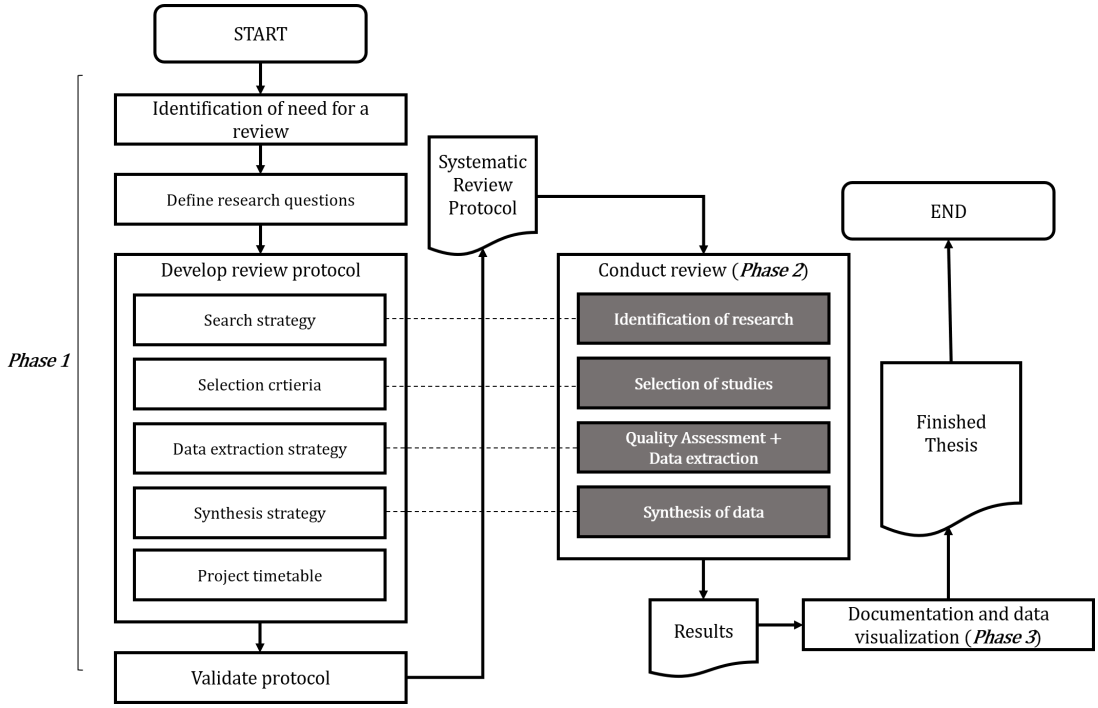

Figure 4.6.: Conducting the review: relevant steps covered in this section

Cloud Infrastructure OR Machine to Machine OR Mobility OR Automotive OR Vehicle OR Car OR Automobile OR Industry 4.0 OR Smart Grids OR machine learning OR cyber-physical OR microservice OR microcontroller OR micro-service OR micro-controller OR blockchain OR neural network OR smart learning OR automated driving OR autonomous driving OR smart city OR smart factory

As in our intial search string (see 4.2.2), the three search categories were combined by using the Boolean "AND" operator. The restriction of boolean operators meant that our search had 26 iterations, one for each term in the third set. For example, one string for a search iteration was:

(privacy OR encryption OR encrypted OR data protection)
AND
(data market OR data marketplace OR data trading)
AND
Sensor.

After carrying out the search using our search string, we exported the metadata of all found studies either in *.bib* or in *.csv* formats (depending on the possible exporting options) in order to extract the resulting titles of the search. Our search experience was especially positive with ACM and IEEE given their intuitive search mechanisms, while SCOPUS had the advantage of allowing several exporting formats. Table 4.8 summarizes the search step:

Table 4.8.: Identification of studies per EDS

EDS	Complete search string used	Fields analyzed	Query date	# of results
IEEE	Yes	Metadata	29.04.2020	49
ACM	Yes	Metadata	30.04.2020	311
WoS	Yes	Metadata	01.05.2020	40
SD	No	Metadata	01.05.2020	143
SL	Yes	Metadata	01.05.2020	69
WIS	Yes	Metadata	02.05.2020	225
SCOPUS	Yes	Metadata	02.05.2020	318
TOTAL				1155

4.3.2. Selection of studies

Once the automated search was finished, we selected our relevant studies using the selection criteria defined in subsection 4.2.3. For this, the first task was to remove duplicate studies. In order avoid reading the titles of all studies manually from the exported files, we coded a rudimentary program in Java that would only read the title lines of *.bib* files. The program can be found in the Appendix.

After removing the duplicates, we were left with 1005 studies, meaning that in this first filter we only had to remove approximately 13% of the found results.

The next step was to filter the remaining studies based on their title. This was done by each researcher separately. Researcher 1 selected 156 studies for the next phase, while researcher 2 selected 127. Out of these studies, 99 were selected by both researchers, leaving 85 studies open for a conflict resolution. After this resolution, 20 additional studies (119 in total) were selected for the following phase. The number of selected studies, which was approximately 10% of the total found studies in the automated search, is a clear sign that the title was a very efficient indicator for the filtering of relevant studies.

The next step was filtering the studies based on the abstracts of the studies. Following the same logic as the previous step, the final result left 79 studies to be fully read in the following step.

The filter based on the content of each study again followed the logic of the previous steps. This step is also the process that took most time to complete in the writing of this thesis. Within this step, we came across several papers that we had to dismiss from the final selection. The most frequent reasons were:

- The studies were too focused on the pricing prespective of privacy preservation.
- The quality of the papers was low.
- Even though many aspects of the studies pointed in the right direction, they did not preserve privacy in a secure enough manner.

This filtering step left us with 37 main studies. Within these 37 studies, we found six out of the eight studies included in the base literature, providing a sensitivity level of 75% and thus resulting in a successful first iteration of our process.

Finally, as described in section 4.2.2, we added additional references found in our selected studies as well as base literature that was not included in the automated search, ensuring 50 final studies for our results. The overall process is depicted in Figure 4.7. The list of studies in the Appendix mentions which papers where base studies, which ones where found in the automated search, and which ones were additional references.

4.3.3. Data extraction and synthesis

The data extraction and synthesis of the studies followed the methodology described in section 4.2.4. We used Mendeley as a reference managing tool. The meta-data used for the mapping was input into Microsoft Excel, while Microsoft Word was used both for data extraction using the extraction cards and for data synthesis before documenting the results in LaTeX.

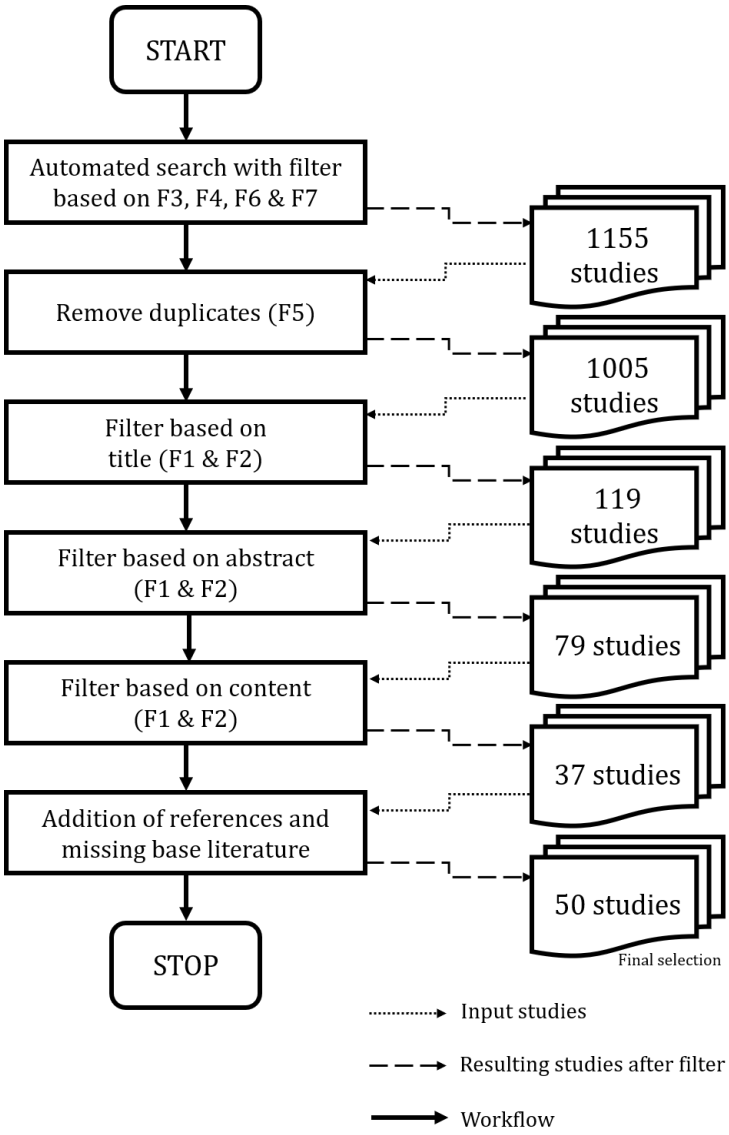

Figure 4.7.: Selection of relevant studies

5. Meta-data Mapping

This section provides a short mapping of meta-data extracted from the 50 final studies used in the SLR (section 6). The goal of this section is to introduce the selected studies in order for the reader to have more context of where, when, how and by whom these studies were conducted. The meta-data used for this section was collected using the extraction cards described in section 4.2.4. We did not map all of the data collected in the extraction cards, since some of the points did not provide useful information. Various graphical representations are based on Philipp's [39] work.

5.1. Distribution of publications per year

In order to know when the papers were written, we determined the publication year of each study. This information is also a good measure in order to know how contemporary our research questions are. 49 of our studies (98%) were published between the years 2012 and 2020, while 33 studies (66%) were published either in 2018, 2019 or 2020. This is a definitive sign that the topics researched in this thesis are modern. We note that the year 2020 presents only 7 studies. Since the automated search was carried out end of April of 2020, we could roughly argue that the year 2020 publishes 1,75 per month (and therefor 7 studies from January to April). This would account to a theoretical forecast number of studies of 21 studies in 2020. This number is consistent with the trend seen in Figure 5.1. One of our selected studies was written in 2002 [50]. This paper proposed a framework for data trading in the IoV. Even if some of the solutions presented in the study are outdated, we decided to include it as a pioneer paper since it addresses some of the challenges that are still relevant nowadays.

5.2. Geographical distribution of studies

In order to determine the geographical distribution of studies, we assigned each study to one or more countries based on the location where the research institutions are situated. Since studies are often written as cooperation research projects between universities of other types of institutions, several studies where mapped to more than one country. The results of this approach are displayed in Figure 5.2 and Figure 5.3.

Among the analyzed studies, the countries with the largest number of studies were the United States of America (USA) and China, accounting for 20 and 19 studies respectively. After these two, Australia was mentioned as a publication country in 5 studies, followed by

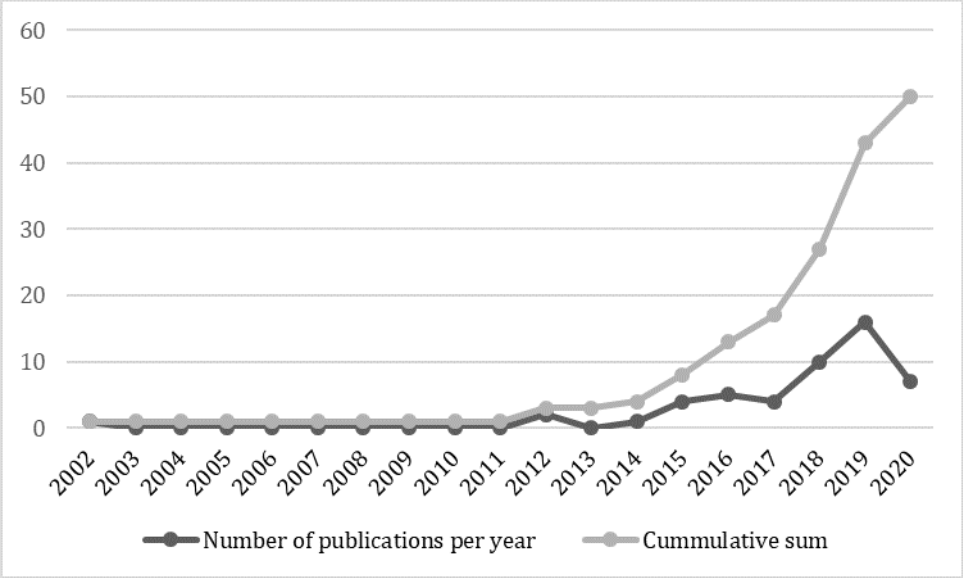

Figure 5.1.: Publications per year

Germany and the United Kingdom (UK). In total, 10 countries have published only one study.

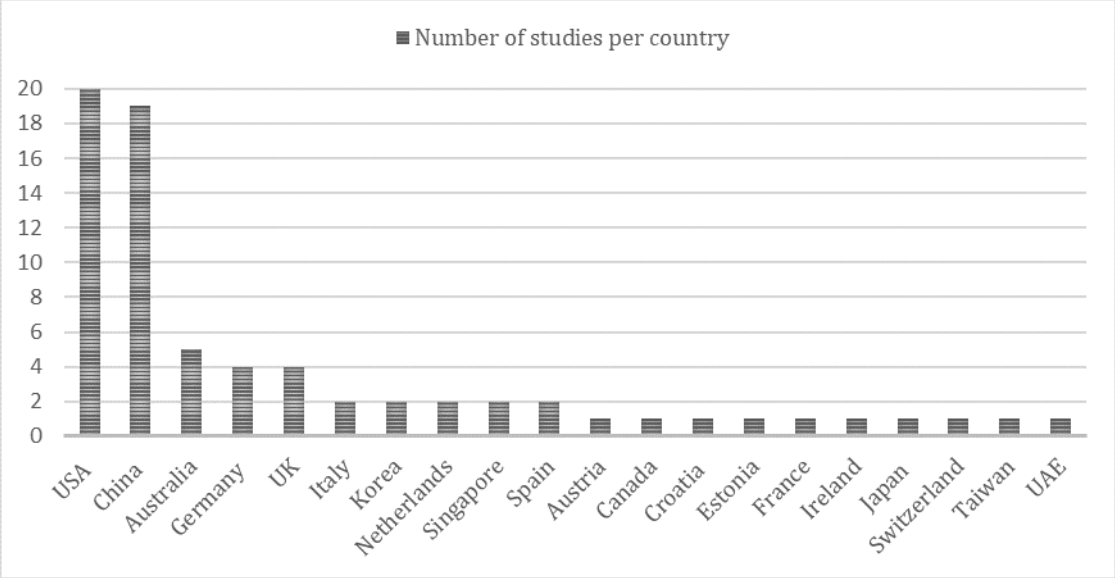

Figure 5.2.: Geographical distribution of studies

13 As far as global distribution goes, Eastern-Asia (including China, South Korea, Singapore, Japan and Taiwan) has the largest number of studies with 25, followed by North America (the USA and Canada) with 21 studies, and Europe with 19 studies. In our selection we did not find any studies from Central or South America, nor from the African continent.

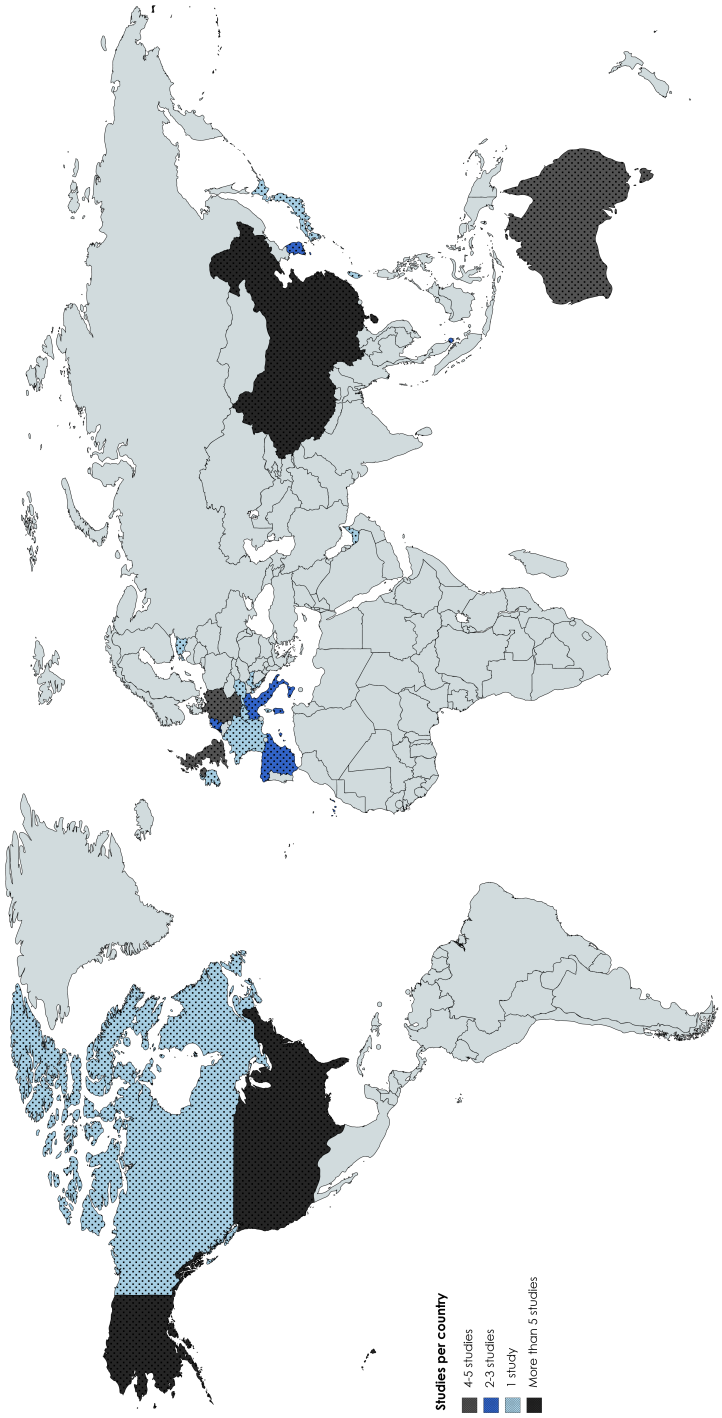

Figure 5.3.: Geographical distribution of studies by regions and continents

5.3. Most salient studies and scientists

This section discusses which were the most salient studies and the most salient scientists. For the most salient studies we defined two metrics: total citations and citations per year. For the most salient scientists we used these same two metrics, as well as a third one that contemplated the total number of studies in our final selection. The cite number for each study was extracted on 10.09.2020, as mentioned in section 4.2.4.

For the most salient studies, we found that [51], [13], [52], [50] and [53], in that same order, were the studies with the most citations, all of them having been cited over 100 times. [51] exhibited an abnormal number of citations with 794, which is a much larger number than the one showed by the next study ([13]), 415. We note that [13] and [50] are among the oldest studies, both being published in 2014 and 2002, respectively. The average of citations was of 55,72 citations per study, while the median was only 12 citations per study. A total of 10 studies have not been cited. Figure 5.4 displays a P-P plot in order to show that there was a big variance in the number of citations per study. The only trend recognizable in this plot is that older studies tend to have more citations. Regarding the second metric – citations per year – we find that [51], [52], [53], [13] and [12] are the most salient studies, in that order. The first four studies were also mentioned in using the first metric. The average of citations per year was 13,82 citations, although only 9 studies were above this average. The median for citations per year was of 4 citations per study.

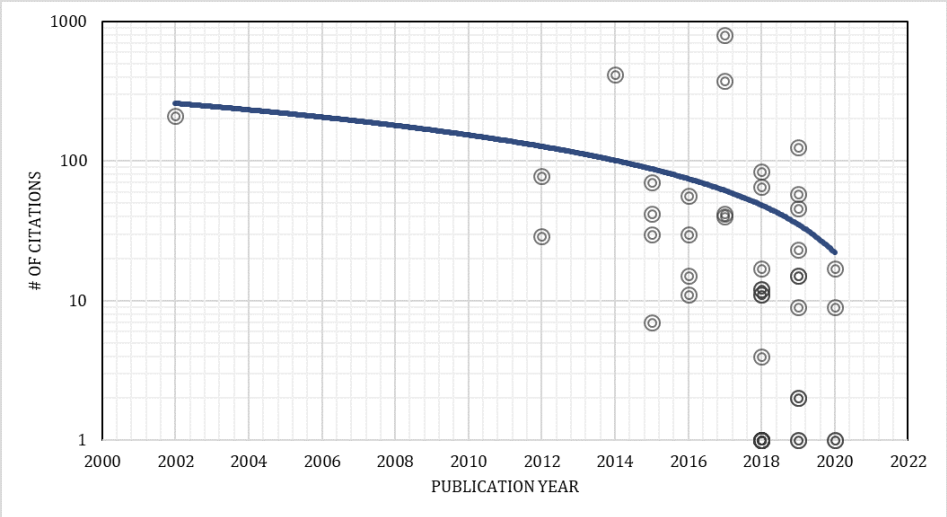

Figure 5.4.: Number of citations for each study (logarithmic scale)

For the most salient scientists, we find that out of the 187 researchers that contributed to at least one study, only one, namely Charith Perera, contributed to more than two studies (he contributed to three). His studies are [17], [34] and [54]. These studies mostly describe challenges in privacy for IoT data markets and summarize some modern solutions for these

challenges. Furthermore, 15 other researchers contributed to two studies. The rest of the scientists only contributed to one study. The researchers with the most citations (in total and per year) are the Dorri, Kanhere, Jurdak and Gauravaram, all four authors of [51], which was the study with the most citations by a large margin (mentioned above).

5.4. Publication sources (research institutions)

This statistic is used in order to know in which universities and other institutions the selected studies were researched. For this, we mapped each study to one or more sources by extracting the institutions where authors and contributors worked at. This means that, similarly to previous sections, the number of institutions was potentially greater than the number of studies.

We identified a total of 79 different institutions in our selected studies. Among these institutions, 65 (82%) were universities, 8 of them (10%) were private companies and 6 of them (8%) did not fall into any of the first two categories (e.g. NPOs or government-owned research centers). These findings are depicted in Figure 5.5.

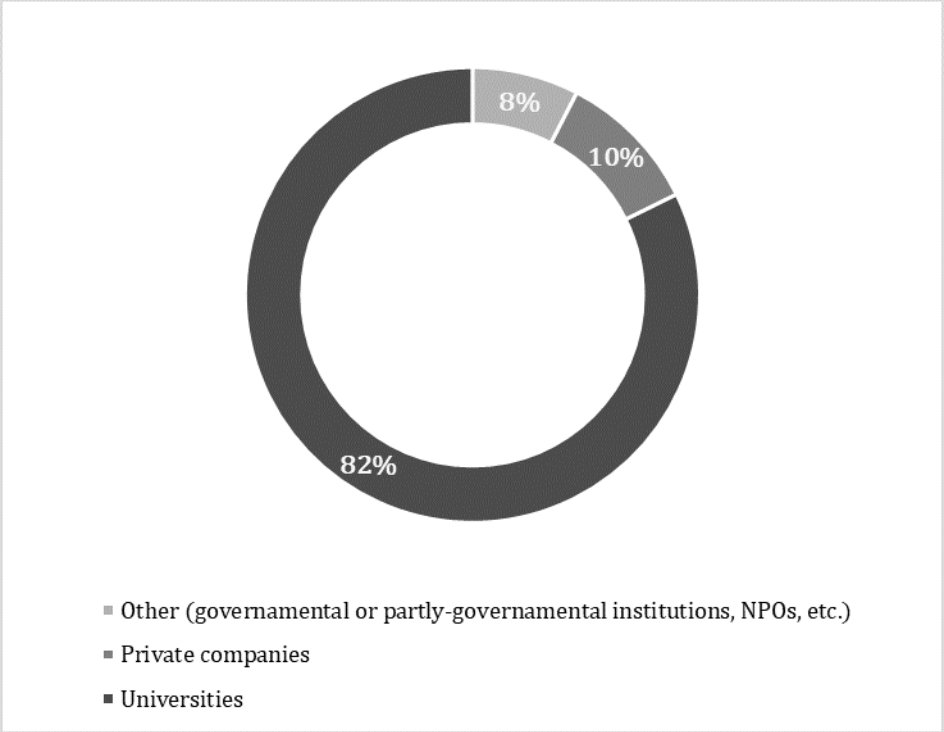

Figure 5.5.: Publication sources

Furthermore, 16 institutions were mentioned more than once, although only eight of them

were included in studies written by different authors. Moreover, three institutions were mentioned three times: the Jiao Tong University in Shanghai, the RWTH Aachen and the Open University UK, the latter including the same main author (Perera) in all three studies.

5.5. Publication channel types and publication channels

As discussed in section 4.2.3, we defined three types of publication channel types, namely journal publications, conference publications and workshop publications. Figure 5.6 illustrates how many of the studies were found in each publication channel type. As expected, given the nature of our RQs, the most common types were journal and conference publications, each of them with 24 and 21 studies, respectively. Only 5 studies were published as part of workshop proceedings.

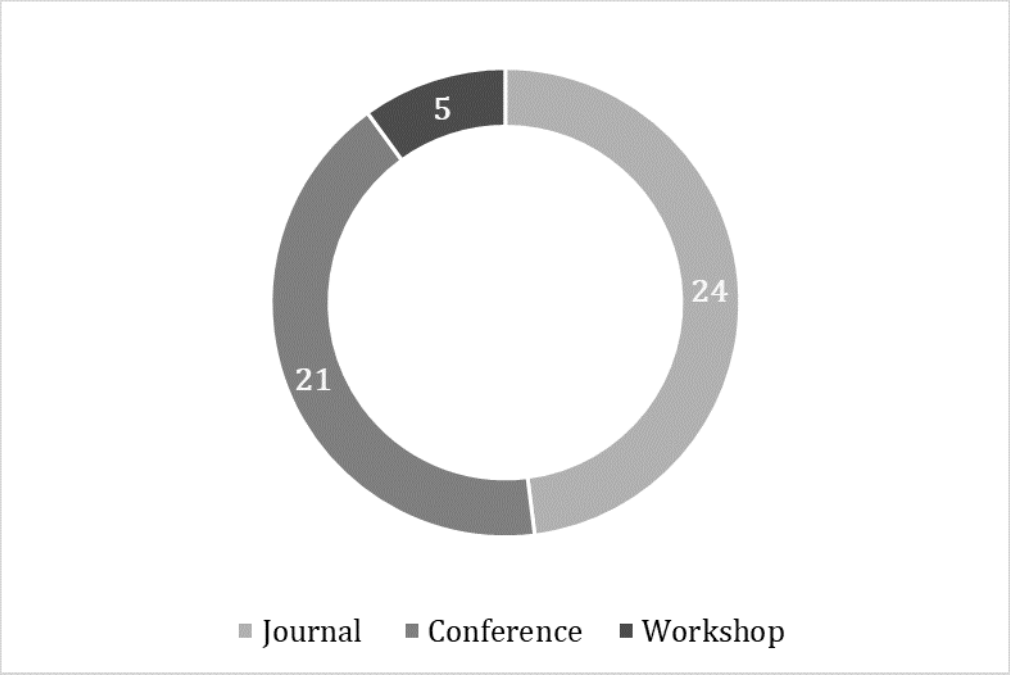

Figure 5.6.: Publication channel types

We found that only three of all publication channels appeared more than once (namely each of them twice) in our analysis. These publication channels were:

- Proceedings of VLDB Endowment (Journal)
- ACM International Conference Proceeding Series
- IEEE International Conference on Internet of Things (iThings)

5.6. Research types, research approaches and research contributions

In order to know which types of studies we encountered in our SLR, we assigned each study with a research type, a research approach and a research contribution as described in section 4.2.4. Figure 5.7 shows the distribution of research types in our found studies. 36 out of 50 studies (72%) were assigned to the research type "Solution proposal". This meant that we would encounter several studies that would propose a solution to a specific problem. Most of the time, the problems that were tackled in these studies were the ones mentioned in section 6.1, where the challenges of privacy-preserving data markets are discussed. Furthermore, we assigned 10 studies (20%) to the category "Philosophical paper" and 4 studies to the category "Evaluation research".

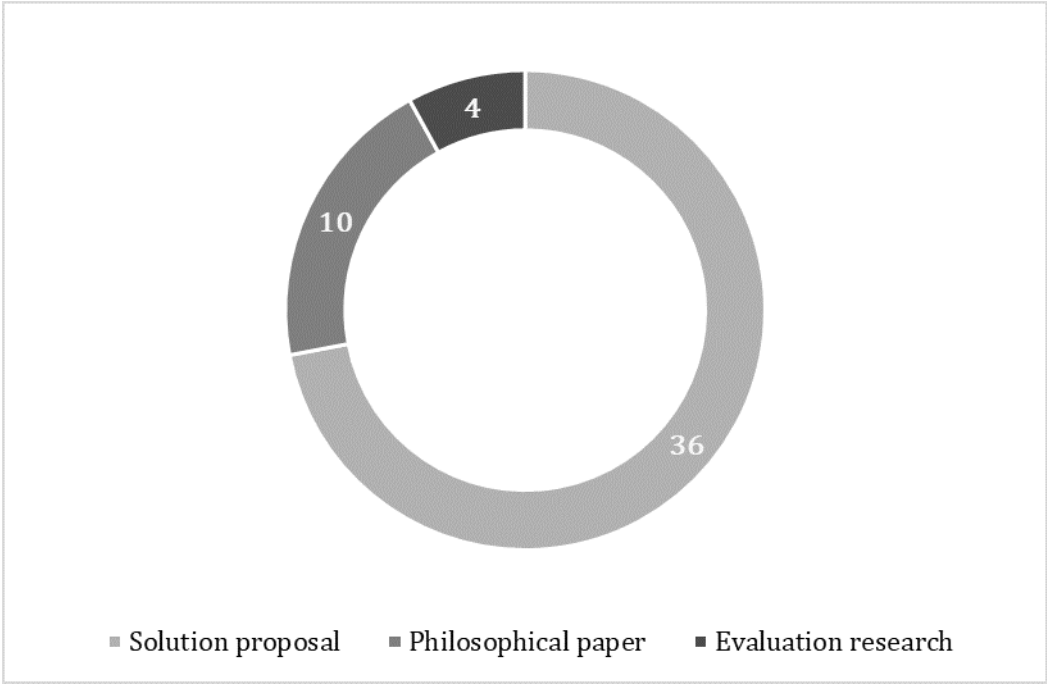

Figure 5.7.: Distribution of research types

Likewise, concerning research approaches and research contributions for the selected studies, we assigned each study with a category as defined in section 4.2.4. Figure 5.8 shows the relationship between the studies having a certain research approach type and a specific research contribution type. We clearly observe that most of the studies have the research approach "Design and creation", accounting for 38 of the total 50 studies contemplated in the SLR (76%). As for the research contribution, the most common type is "Framework or methods" followed by the contribution type "Model", both accounting for 34 (or 68% of the total) and 11 (or 22%) studies, respectively. Unsurprisingly, the combination that was most present by a large margin was the one found in studies where the research approach was "Design and creation" and that had a contribution of the type "Framework or methods",

accounting for 62% of the overall studies.

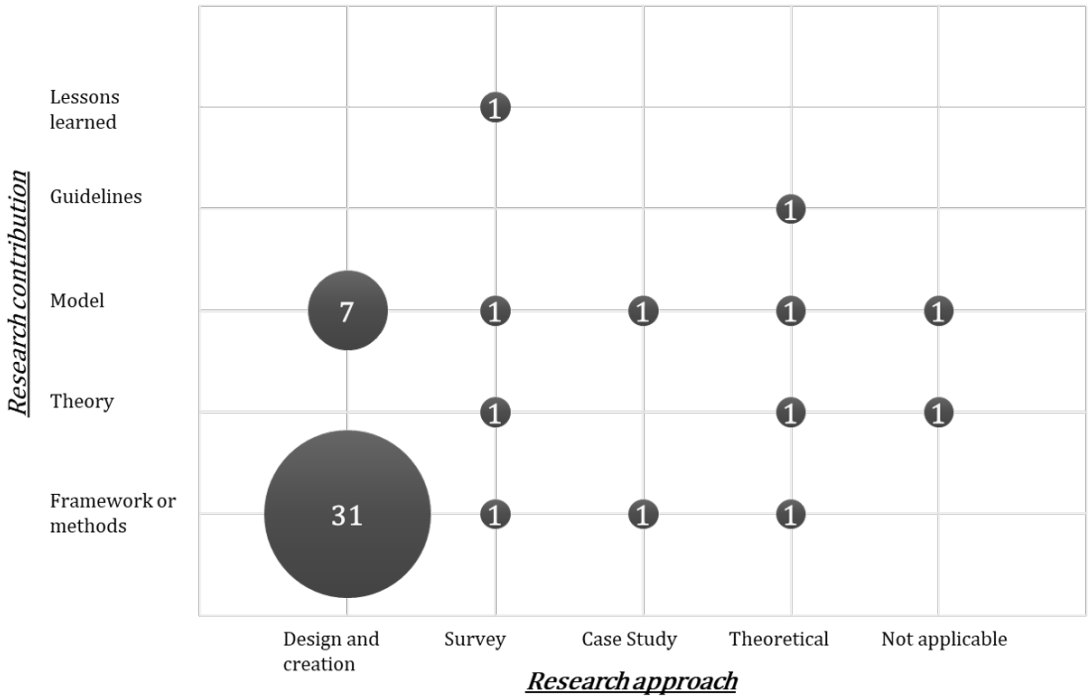

Figure 5.8.: Distribution of research approaches and research contribution types

5.7. Specific industries

This point analyzes how many of the proposed studies focus on a specific industry branch, and how many concentrate on IoT data markets that can be used for any industry sector. Figure 5.9 displays the distribution of the selected studies regarding this matter. 47 of the 50 studies are industry agnostic and their findings can be used in a series of industrial scenarios, whereas 3 studies concentrate in individual industries. Specifically, [55] focuses on data markets for mobility services, [50] proposes a solution for automotive telematics and [32] concentrates on the healthcare branch.

5.8. Effectiveness of EDS in the various steps

The last provided statistic defines how effective each EDS was for each filtering stage. In other words, we analyzed how many of the total remaining studies after applying one filter were found in a specific EDS. For this particular statistic, we did not consider the additional papers that were added to the automated search (the 13 additional references as well as base literature that was not found in the automated search). It is also important to note that several

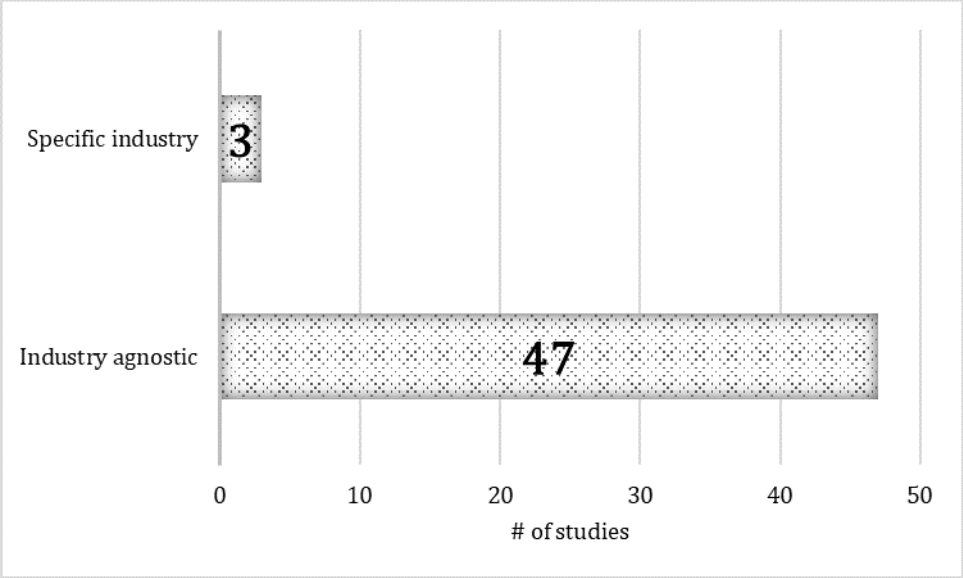

Figure 5.9.: How many studies are industry specific?

studies were found in more than one EDS.

As stated in section 4.3, the various EDS provided a total of 1155 studies before any filtering took place. The analysis summarized in Figure 5.10 shows that SCOPUS was the EDS with the most studies throughout the study selection phase. WoS and IEEE were the most effective EDS: both of them presented a low number of studies in the first step, but showed a high level of accuracy based on the search strings and the requirements for this thesis. SpringerLink was the only EDS that did not include any study in the final 37 studies selected in the main search.

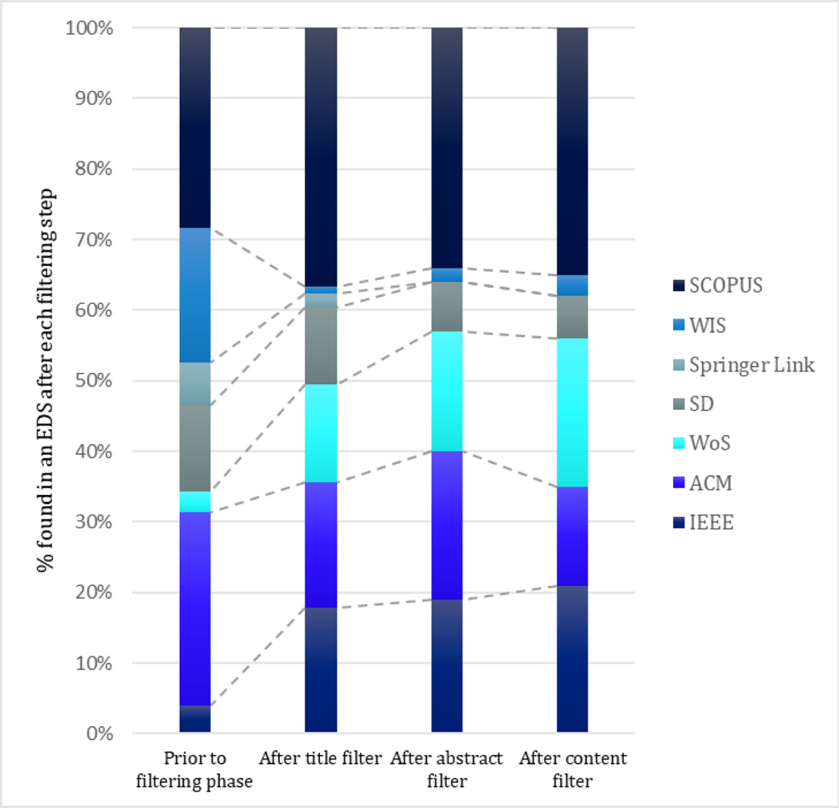

Figure 5.10.: EDS effectiveness

6. Literature Review

This section composes the main part of the thesis. While the previous section introduces the results from the found papers according to the meta-data and therefor gives us a better understanding of what types of studies we have, this section aims to take a deep dive into the content of the studies in order to adequately answer the research questions. The structure of each subsection aspires to give a brief summary of the found results in context with the research question, as well as to provide the reader with some concrete examples found in the studies in the hope that the somewhat abstract concepts presented in each subsection can also be illustrated with a tangible application.

6.1. Research gaps and problems

RQ1: What are the current research gaps and problems in the implementation of privacy-preserving data markets for IoT devices?

In this chapter, we describe the research gaps described in the literature concerning privacy in data markets in general, as well as some issues that are specific to the IoT framework. While some of the presented problems are efficiently tackled by the solutions in the found studies, others persevere or have not found an adequate solution and are therefor part of the topics to be addressed in future research (see section 6.4.

6.1.1. Third party trust

The first challenge we identify is that it is difficult to achieve fairness in data marketplaces without a trusted third-party [6]. Besides fairness, many other requirements expected in existing data markets (such as security and reliability) cannot be guaranteed without a mediating third-party [15]. [21] defines trust as a "directional relationship between two entities – a trustor and a trustee – where a trustor trusts a trustee to perform a certain action within a given scope [...]. Optionally, an action may have a qualitative modifier which implies the quality of the action".

In the context of privacy in data markets, the defined action depends on the trading scheme. For instance, in a simple one-to-one trading scheme, the data owners must be able to trust the platform and feel safe in sharing their data. On the other hand, data consumers need to be convinced that the collected data is unbiased. In some auction schemes, not only the data owners will be requiring privacy, but also the data consumers, who don't want their bids

published. If the data is collected via a third-party mediator, then it is this third-party that needs to be trusted. This is because data markets generally exhibit a transitivity property [56]: “if an entity A trusts another entity B and B trusts another entity C, a trust relation can be derived between A and C. To derive such a relation, the trust scope and the action must be the same”.

A survey conducted by [20] found out that third-party trust is the main reason not to participate in a data market. Within this reason, some of the mentioned points were:

1. "Data that is online can never be deleted"
2. "My data will be abused"
3. "I don’t trust that my data is secure"
4. "I don’t trust that the data is only used for the stated purposes"

Besides requiring trust, a mediating third-party has other issues, such as costs: schemes with a third-party suffer from complex transaction processes and a higher transaction cost [6].

6.1.2. Truthfulness vs. privacy preservation

The second problem we address is the contradiction of the goals “maximizing data truthfulness” and “preserving privacy”. It is difficult to guarantee the truthfulness of both data collection and processing if the data is constrained with privacy requirements [30]. The data consumer has to handle with the question if the provided data really is the same as the claimed data without breaching the privacy of the data owner.
If the first challenge presented in this section dealt with trusting a third party with one’s own data, this challenge deals with trusting the data owner or the data processor with their own data. There are several reasons why the data could be tampered with: A data provider may not want his/her privacy intruded with, or he/she may want to save on collection costs and may so fake the data. A data processor may want to reduce operating costs during the execution of data processing and may return an incomplete data service without processing the whole raw data set. The data processor may also reduce costs by reducing the real number of data contributors, as acquiring them may also be expensive. The problem here arises when the data processing can no longer be semantically consistent with the raw data, which makes the data consumer skeptic about the results of the processing. In addition, some digital signatures on raw data become invalid for the data processing result, which discourages the data consumer from doing verification [57].
Ensuring truthfulness and protecting the privacies of data contributors are both important in the long-term healthy development of data marketplaces [57]. In efficient data markets, mechanisms with incentive capability should be designed to encourage data owners and processors to provide their data honestly [3].

6.1.3. Accountability vs privacy preservation

Privacy preservation contradicts accountability (data traceability and provenance) as much as it contradicts truthfulness. In this case, it is not the content of the data that is questioned, but its origin. Should one party provide dishonest data, efficient mechanisms must be implemented in data marketplaces in order to hold that party accountable and be able to take countermeasures against that party (e.g. banning that party from further participating in the data trading platform). This requirement violates the privacy need for impeding reidentification.
This philosophical contradiction can be translated into concrete technical challenges. When using digital signatures, most existing signature algorithms treat signer's identity as a public parameter, while in data markets the data provider might want to protect her identity (and hence the signature) [30]. Also, data exchanges could take place across several data centers before reaching the destination. The data could take different paths for the same "copy" in order to ensure resilience. This design adds a degree of difficulty in order to accurately identify the origin of false data as well as its impacts [52]. A platform must provide a detailed history of the origins of all changes to a data object and not only simple log records of the high-level processes.

6.1.4. Legal challenges

Another set of challenges comes from the legal perspective. [13] summarized a list of grey areas that are created by the IoT regarding data markets and how these grey areas have ample space to circumvent legislative boundaries: Firstly, most pieces of legislation center around fuzzy notations of "personal identifiable information" (PII). These notations are quickly deprecated as new IoT technologies unlock and combine new datasets that can enable identification and make the distinction of PII and non-PII almost impossible. Secondly, the timeliness of legislation is a big issue since laws in most countries are passed at a much slower pace than technology advancements are researched. Thirdly, many of the privacy breaches go unnoticed: an impressive "shadow market" for private data. This market's shadow existence "undermines its long-term viability, making all its players operate at the edge of what is ethically sustainable" [18]. Lastly, the economic aspect of privacy is still favorable of those that disregard privacy legislation: on the one side, the development of Privacy Enforcement Techniques is very expensive and limits business models. On the other side, violations of privacy legislations either go unpunished or result only in comparably small fines (see [58]), while public awareness is still too low to induce great damage of public reputation.

The European General Data Protection Regulation (GDPR) has empowered its citizens by giving them a series of rights regarding their data protection [59]. The legislation aims at individuals (and not only businesses) also having economic exploitation of their own data [60].

Nonetheless, regulation on its own is not a viable solution as it is not realistic to police

the entire online behavior [61], and "hackers will always find their way" [55]. Many of the existing laws only focus on protecting the legal rights of the owner and cannot protect the privacy of the data directly; they are not preventative but rather reactive [1]. On the other hand, [62] argues that there are problems that can only be solved through legal means (and not technical ones), such as the reselling of data to third parties by buyers or malicious marketplaces.

Another problem that arises in current legal challenges is the discussion of how strict these policies should be: introducing heavy regulations could stifle the market and innovation [61].

6.1.5. IoT-specific challenges

Enabling privacy in data marketplaces for IoT devices must take into account the functional and technical challenges that IoT devices carry with them. [63] states that designing IoT applications is much more difficult than designing desktop, mobile or web applications.

The first challenge is one of heterogeneity: IoT applications require both software and hardware (e.g. sensors) to work together on multiple heterogeneous nodes with different capabilities under different conditions [64]. Each of these devices may have different characteristics, strengths, weaknesses, types of access to energy sources or built-in capabilities. An IoT data market should ideally be able to integrate the various types of devices an minimize the requirements asked for these devices. This calls or developers to have a high understanding of the diverse technologies and middleware solutions available to integrate the devices [34].

The second challenge is one of performance and storage capabilities. Many networkable devices are low in energy and lightweight. These devices must use most of their available energy and computation to executing core functionality, making the task of affordably supporting privacy more challenging [51]. This becomes a problem when using many of the privacy-preserving techniques that require expensive computation and must be carried out within the devices themselves. Also, because they are normally small in size and lightweight, IoT devices don't tend to have much storage capacity available, which is contradictory to applicable Big Data problems that arise in the IoT context [22]. Since the storage capacity in IoT devices is limited, much of the data in scenarios such as data markets is traded "all at once" [15], which requires that the trading process (including all relevant verifications and privacy enhancement techniques) happens rather quickly.

A third challenge is one of lower data quality due to the unreliability of IoT sensors and the fragility of data transmission links. The data markets should be function in a robust manner as to counter these errors and inconsistencies within the sensing and transmitting phase [30].

Another challenge comes from the heterogeneity of the data owners. In IoT scenarios there are a lot of small data owners, which is different from traditional data markets where large

parties (such as governments, agencies or companies) sell their datasets. Since there are many data owners, the data cannot be sold as one block and the price is not fixed for the entire dataset [22]. In some cases, there is also the case of ambiguous data ownership, where it is not clear who generated which data [30]. This challenge is interesting in the market design phase when developing data market frameworks.

Moreover, in IoT scenarios, the timing of measurements is more important that in traditional datasets, where the data must not always be updated in real-time [22]. In IoT, there are many use cases where the data is only valuable if the delay tolerance is in the range of a couple of minutes or even a couple of seconds [22].

Furthermore, the type of data that is collected by IoT devices (e.g. in smart home environments) can have higher privacy requirements than data in traditional data markets. Additionally, many IoT devices collect activities and behaviors 24/7, which not only increases the privacy risks, but also requires the data receiver to be online at all times [54].

Also, in many scenarios, measurements of data are often purchased in advance, meaning that the data is sold before it is collected by the IoT devices [22]. [62] describes a marketplace where the sold items are not data bits per se, but the access to data streams. This characteristic implies that negotiation frameworks for IoT data markets must be specially tailored, as well as the mechanisms for enforcing that consumers pay for data they consume and devices provide the data their owners agreed to provide [22].

6.1.6. Security challenges

When discussing security, we define it as a measure to provide confidentiality. Confidentiality refers to protecting information from being accessed by unauthorized parties or in other words, only authorized can gain access to sensitive data [65]. A failure to maintain confidentiality means that a party who shouldn't have access has managed to get confidential information, through intentional behavior or by accident. Such a failure of confidentiality, commonly known as a breach, typically cannot be remedied, meaning that "Once the secret has been revealed, there's no way to un-reveal it" [65].

In this section we do not go into detail of every single challenge existent in a security breach within data markets because that is a whole subject for itself. Instead, we aim at providing a high-level overview of possible threats that are mentioned and tackled in our found studies. It is pertinent to note that a security breach is a direct privacy violation when it comes to personal data, which is why security is one of the aspects covered when discussing relevant privacy-preserving techniques in section 6.3.1.

Firstly, we categorize the types of entities that may breach security barriers:[31] defines the adversary models, regardless of the roles they assume in various proposed solutions as either

semi-honest or malicious. semi-honest entities are assumed to follow defined protocols and do not actively alter the data to breach the privacy of participants. However, they may attempt to exploit acquired information from participants to learn their private data, while malicious entities actively try to breach the privacy of participants (e.g. with deanonymization attacks) [31].

Moreover, the relevant attacks can then be further classified based on how they take place. [61] describes the following categories:

1. "Data Forwarding" describes an attack where the seller's raw data is being forwarded to an authorized party. This attack is particularly challenging within data markets. A scenario where a buyer re-sells the acquired data is very feasible since information can be duplicated with costs almost to zero and quickly lose its value. [62] states that this particular issue within data markets cannot be solved through technical means, but rather though legal ones.
2. A "DoS-attack" happens when an attacker invokes a marketplace functionality (e.g. a smart contract) repeatedly to exhaust the network communication resources in the platform.
3. A "Repudiation and Fraud" denotes a fraudulent activity by the data buyer.
4. A "Collision" happens when a malicious actor may be the same person or there is a collision between one or more actors in different roles (e.g. a data broker and a data seller colliding to get a higher selling price).
5. An "Information Leakage during Running" occurs only during the execution of data analysis within the data market.
6. A "side channel attack" is when an adversary is trying to take advantage of the physical specificities (the implementation) of actual cryptographic /computational systems [66].

For IoT scenarios, security breaches enable various threats that can be summarized using the summary presented in [13]:

First, there is the threat of (re)-identification, which is the the threat of associating a persistent identifier with an individual and data about him / her. The threat lies in associating an identity to a specific privacy violating context while also aggravating other threats mentioned below. This threat was the most mentioned in all of our found studies and is also the focus threat in this thesis. IoT enhances this threat with modern technologies such as Facial Databases, Speech recognition or identification of devices through fingerprinting.

Second, there is the "Localization and tracking" threat, which refers to determining and recording a person's location through time and space. This threat is enabled though better GPS technologies and the increasing usage of Location-Based-Services.

Third, there is a threat called profiling, where information dossiers about individuals in order to infer their interests are compiled by correlation with other profiles. Besides privacy breaches, this threat can lead to other issues such as price discrimination or erroneous automatic decisions.

Last, there is the threat of linkage, where previously different separated systems or various datasets are linked such that the combination of data (and its sources) reveal information that the subject did not disclose in previously isolated sets. This increases the risk for re-identification of anonymized data. This risk is higher within the IoT context because of the horizontal integration between different systems (e.g. different companies and manufacturers) to form a system of systems in order to deliver new services that no single system could provide on its own [13].

6.1.7. Costs

A further challenge in the research of privacy-preserving data markets for IoT devices is the matter of the platform costs. The achievement of all needed transactions between two parties requires a few "runs" to get in touch with and it may take several negotiations (especially in auction mechanisms) to reach a consensus. For example, a buyer may need to search and call a huge amount of resources in order to find reliable data owners. Furthermore, if data owners suffer from low sales they may need to invest in additional resources in order to promote their sales [15]. Moreover, the costs of maintaining the trading platform must also be economically viable [60]. Another cost factor is the awareness of individuals: individuals are generally not aware of the threats of using IoT devices and place a low value on privacy. Informing the owners about their privacy options in order to fulfill transparency requirements can also signify a huge cost [31].

6.1.8. Pricing mechanisms

Pricing the sold data is one of the main challenges concerning data marketplaces. One could argue that if both parties agree on a price for specific data, the entire privacy problem could be solved. While this is the perspective taken by many of the papers that we found prior to the filtering phase, we decided not to concentrate on this angle. Nevertheless, even though we do not discuss the pricing solutions, we mention some aspects that are relevant when approaching this matter in section 6.2. [60] mentions two key challenges that must be resolved when researching pricing mechanisms: First, the negotiation mechanisms must be taken into account when contemplating the technical implementation. And second, there must be supporting incentives for all parties.

[31] comments on the similarities between trust challenges and pricing challenges between data marketplaces by describing the concept of "Reward-based Tasking": *"The challenge for rewarding participants in the presence of privacy mechanisms is very similar to the trust challenges*

since both require participant evaluations. However, trust models need to trace and review participants progress while incentives can be handled per task completion without linking it to other tasks.".

6.2. Current research topics

RQ2: What are the research topics that have been addressed in the intersection of privacy and data markets for IoT devices?

In this section, we discuss the current research topics found in our selected studies. We identify five perspectives from which the design of a privacy-preserving data markets for IoT devices can be approached according to the selected studies. As it is the case in the found research gaps and problems (see previous section), the found research topics and areas are highly coupled. Even though each study focuses on one perspective (e.g. the architecture of the platform), they all tend to at least mention the remaining perspectives in order to tackle the basic questions that arise from them. Figure 6.1 illustrates our classification: the first four perspectives can be seen as the building pillars of privacy-preserving data markets. These four pillars are wrapped by a fifth perspective which is use case specific. The use case lays the framework, the focus and the constraints that must be contemplated during the research of the four pillars.

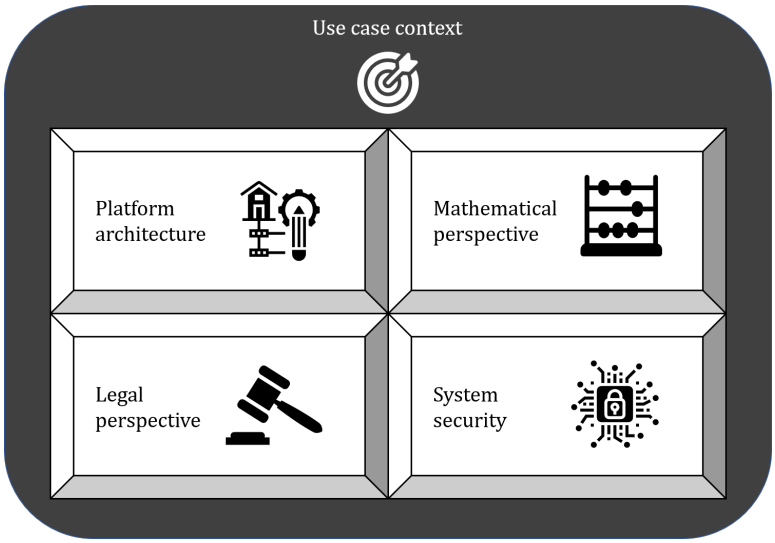

Figure 6.1.: Classification of current research topics

6.2.1. Platform architecture perspective

The first perspective focuses on the architecture of the data market. In a broad manner, when discussing architecture, we refer to the way some components (be it software, hardware or abstract entities) and their interactions are structured in order to compose a system. In a narrow manner, we refer to the definition of architecture layers and how these layers encapsulate specific services in order to reduce complexity and modulate functionality. We first discuss the most important questions in the broad architecture definition. We then describe two examples of concrete layer solutions, one centralized and one decentralized, as proposed by [55] and [17].

The first question refers to the distribution of data controlling nodes, in other words: Should the data market be centralized or distributed? As described in the previous section, a centralized control platform requires a trusted third-party, which supposes a big issue in complex networking systems. On the other hand, the lack of a third-party can lead to unfairness within data transactions. Distributed Ledger Technology (DLT) has emerged in the last years as a promising technology to tackle this contradicting issue, since it provides a distributed and trusting node-network while ensuring data verification. We discuss DLTs in detail in section 6.3.3. Out of 36 concrete solutions, 15 of them, which accounts for approximately 42%, propose a decentralized data market. All of these solutions used DLTs. Note that even in the most decentralized solutions, complete independence from a third party was not proposed by any found study, since there are many tasks that remain open and must be solved with a third-party (e.g. a company maintaining the blockchain functionality).

Another topic that must be addressed is the data storage, be it in distributed or in centralized solutions. In the context of a blockchain-based data market, [12] proposes separating the blockchain from the actual storage using Distributed Hash Tables (DHTs) [67]. The idea is that the pointer to the DHT storage address can be stored in the blockchain. When an entity requests data from the DHT, the blockchain decides whether the access can be granted or not. Besides limiting the trust issue, this solution counters the single-point-of-failure problem. The idea of providing an entry key to a separate storage entity and not transmitting the entire data on the platform is proposed by other studies such as [15] or [17]. There are two more advantages of this principle: first, it immensely reduces transmission costs. Second, it may be a necessary path for future blockchain solutions. [62] estimates that unlimited storage in blockchains will no longer be available (starting with Ethereum), and that the blockchain providers will include storage rent (e.g. for smart contracts) as part of the transaction fees. Regarding storage, another topic mentioned in the studies is that, since privacy preservation techniques first need to identify which part of the raw data is private and which not, the storing structure (the structures of the databases) and its detail level will play a key role in processing stages. This means that attributes within databases will have to be given in a very structured manner. Should the data not be well structured, [68] refers to theoretical solutions in order to determine privacy of a dataset based on their informativeness (see [69] and [70].

In terms of physically structuring the components in a data market, the concept of edge

computing is mentioned in several papers such as [12]. Edge Computing can help mitigating the IoT specific issue of low computational power since its methods process the data at the network edge rather than in the remote cloud [71] [72]. By doing so, devices at the edge can bring real-time computations that require more power (such as cryptographic computations). Solutions that don't use edge computing tend to rely on third-parties to compute expensive functions. The same principle goes for data analysis and mining algorithms [60].

Another question that is relevant to the architecture is how the market has been designed from a mathematical perspective. If the mathematical model needs a certain number of independent systems in order to function, the software components must be adequately developed. We shortly review some of the modeling approaches in market design in a subsection below.

In a more narrow and specific manner, [55] and [57] describe the architecture of their proposed solutions by designing functional layers. They describe a decentralized and a centralized solution, respectively:

1. The Identification layer is composed of mobility and other information that the nodes in the blockchain own.
2. The Privacy layer is privacy model (based on differential privacy) for accessing Location Based Services (LBS).
3. In the Contract layer are the set of smart contracts and the brokers who facilitate data transactions between nodes. Brokers are nodes in the network that arrange transactions between nodes for selling or buying transport information. To reduce possible scams, brokers need identity keys from trusted nodes.
4. The Communication layer contains so-called Decentralized Identifiers (DIDs) of the nodes who serve as endpoints to establish peer-to-peer connections without revealing any personal information. The key in this layer is that a single node will have multiple DIDs. This makes it harder for malicious parties to intercept the information while it's being transferred. Given that a node will have one unique DID per transaction, it is difficult for an attacker to correlate DIDs in the ledger to track single nodes.
5. The Consensus layer contains the consensus algorithms in which the active nodes agree to write transactions in the ledger
6. Finally, in the Incentive layer are the rewards the participating nodes.

[57] presents a centralized solution called TPDM that is made of two layers. The essence of the solution is to first synchronize data processing and signature verification into the same ciphertext space (Encrypt and then sign), and then to tightly integrate data processing with outcome verification via some homomorphic properties: In the first layer, a centralized service

provider procures the raw data and rewards the owners for it. Each of the owners has a tamper-proof device that generates a pseudo identity. A third party called the registration center maintains an online database and assigns each owner with a password for the tamper-proof device as well as the parameters for the signature scheme. The registration center is the only party that can retrieve a collector's real identity. In the second layer, the data is sent to the data consumers who can verify the data for truthfulness and correctness in case of prior processing stages.

6.2.2. Mathematical perspective

The mathematical foundations of game theory are based on the idea of finding the best strategy in a rule-constrained scenario in order to achieve a certain state. Market design, on the other hand, focuses on setting up the rules in the given scenario for the entities in it to behave in a certain way [73]. In order to enhance privacy within data markets, there are various theories and methodologies used in market design that can define the data market's functionality and structure in an abstract manner. The mathematical foundations lay the ground for concrete implementations. Data markets can make use of various subfields such as game theory, auction theory or matching theory to support their market design. In this section, we do not intend to give an overview of all the fields and theorems that are applicable for a privacy-preserving data market, but rather to summarize some important thoughts that we found in the selected studies. Our effort mainly concentrates on auction theory, since it is the data market scheme (within the ones relevant for this section) that was most mentioned in our selected studies.

36 out of our 50 selected studies propose a concrete implementation or solution. Within these studies, we find a total of nine papers that propose auctions as an implementation scheme, while many other papers mention auctions as a possible way of trading data. Auctions are among the best-known market-based allocation mechanisms [74]. This type of schemes are frequently proposed in scenarios that support mobile crowdsensing (e.g. [29] or [75]). They require a further level of privacy since not only the traded data must be protected, but also the bidding data that must be processed through a winner-selection algorithm. If this were not accomplished, the disclosure of bid information could be exploited by several entities [74]: an auctioneer could adapt its pricing strategy based on the bidders' bids to obtain more profit. Bidders may get to know others' willingness to pay and choose to bid untruthfully to get extra profit, thus tampering with the truthfulness of the auction. An external attacker could submit a bid that cannot win the auction but could increase the price paid by the winners. Also, bidding information could be related to the bidder's preferences, his/her economic situation, the geographic location or other private attributes. Besides protecting the information of the bidders, efficient auction schemes must implement mechanisms to induce bidders to truthfully submitting their bid in order to ensure fairness and avoid tampering with economic properties of the scheme [1]. Auction mechanisms can also be used in combination with a relevant privacy-preserving technique, differential privacy, which

will be discussed in another section (see 6.3.2). The interested reader can refer to [3] for a concrete implementation in an auction scenario. Auction schemes can be used in various use cases. [75] develops an auction framework for privacy-preserving data aggregation in mobile crowdsensing. In this framework, the platform selects the data collectors (called workers) based on their sensing capabilities. This solution aims to address the drawbacks of other game-theoretic models that do not ensure the accuracy level of aggregated results. Another objective of this design is to minimize the computational effort of finding an optimal subset of workers, a task that is described as NP-hard, and therefor minimizing the purchasing costs. A key element in the data aggregation scheme is that for different set of winners, different noise distributions are assigned to the workers. In other words, the privacy of each winner depends on the selection of the winner set. [76] proposes a combinatorial auction (CA) that combines homomorphic encryption, blind signatures and onion routing in order for a third party to not gain information on the biddings and the IDs of the bidders. In this single-auctioneer CA, the auctioneer sells multiple heterogeneous good simultaneously, and bidders bid on any combination of the goods instead of just one.

All of the following papers that are part of our final selection propose an auction scheme: [74], [1], [75], [27], [76], [74], [77], [29] and [23].

Apart from auctions, there were other studies that offered a mathematical solution in the design of the data markets: [78] proposes a contract design approach to find the optimal contracts when using a biased algorithm to provide privacy and shows that under this combination, buyers can achieve the same level of accuracy with a lower payment as compared to using unbiased algorithms. The key is that the sellers measure their privacy loss (and so their price) with the concept of (biased) differential privacy.

Some papers, such as [8], focus almost exclusively on a game theoretical analysis of privacy-preserving data markets. In game theory, the players are considered rational and they will always select an optimal strategy that maximizes their utility. According to [3], this rationality is more suitable for IoT frameworks than for example Multi-Party-Computation, where other entities are considered adversarial. On the other hand, [75] states that one of the problems of game-theory is that because of its intent to focus on the equilibrium status, a platform may end up with an inefficient equilibrium, i.e., the platform may not achieve a desirable accuracy level of aggregated results.

The following selected studies discuss game-theoretical approaches in privacy preserving data marketplaces: [55], [3], [8], [54] and [79].

Furthermore, in many of the data market schemes (including auctions), the efficient matching of two parties represents another problem that calls for purely mathematical theory. [80] proposes a stable matching scheme (referring to [81]) in order to match data providers and data consumers.

Lastly, another element that calls for market design theory is the determination pricing of the sold data. As stated in the introductory sections, we do not focus on pricing schemes since this would extend our scope out of proportion. While not mainly focusing on pricing procedures, some of our selected papers give a good overview of the topic: [1] studies a variety of data pricing models, categorizes them into different groups and conducts a comprehensive comparison of the pros and cons of these models. [78] is proposes a data marketplace that prices the data depending on the level of privacy it has (measured with differential privacy). [3] provides a thorough summary on auction mechanisms and used pricing approaches in IoT data markets.

The interested reader can refer to the following studies for more information on pricing: [82], [83] and [84].

6.2.3. Security perspective

When discussing security in IoT privacy-preserving data markets, this perspective encompasses all the topics that aim at tackling the mentioned challenges in section 6.1. In that section, we defined security as a measure to provide confidentiality (protecting information from unauthorized access). Since the unauthorized access to personal information is a severe privacy breach, we consider that security measures are a subset of privacy preservation techniques, which is why we discuss them in section 6.3.1.

6.2.4. Legal perspective

In section 6.1, we listed some of the challenges that regulatory parties face when they must legislate privacy in IoT data marketplaces and we discussed why some of the proposed laws have failed to effectively address this issue. Nevertheless, our selected studies mentioned some legal schemes that point into a more adequate direction.

The first relevant attempt to regulate private data in an Internet context found in our studies is dated back to 1995, where the European Parliament and Council of the EU defined religion, political and sexual orientation or race as sensitive data [68]. Several US federal laws on medical data provided lists of diseases that could be potentially discriminatory and should not be disclosed back in the early 2000's. In 2014, to mitigate the privacy issues of the uncontrolled rise of data collected in the Internet, the US Federal Trade Commission suggested the implementation of mechanisms, that is, legislations and technical solutions, that would enable consumers to access their data and give them the ability to control the use of their data according to their preferences [68]. Users would also have the option of opting out of having their data used for secondary purposes. This legislation also purposes guidelines to practice privacy-by-design (discussed in section 6.3.4). Another effort by the legislative side is the well-known European General Data Protection Regulation (GDPR),

which has empowered its citizens by giving them a series of rights regarding their data protection [59]. The legislation aims at individuals (and not only businesses) also having economic exploitation of their own data [60]. On the other side of the globe, the Act of the Protection of Personal Information in Japan defines the management of anonymized datasets [85]. Under this act, it is possible to distribute anonymized datasets in B2B scenarios without additional user consents.

In our study selection we came across one paper that was almost solely focused on the legal perspective for data marketplaces [18]. This paper provides a thorough set of technical, legal and structural requirements in order to achieve a data market that enables privacy in a realistic manner. Concretely, the study proposes a market that contains four "spaces" and assigns standardized data handling and data exchange rules to each of these spaces. Every entity is assigned to one space, although one company may take on different roles when providing services. The rules in each market space need to be enforced by technical infrastructure and regulation. We will shortly summarize these four market spaces:

The first market space includes customers and companies (hereafter called "Customer Relationship Holders" or CR-H) directly involved in a service exchange. The customers are willing to provide their personal information in a service context if it is necessary for service delivery or if they receive appropriate returns for their data. No shadow markets are allowed. For this market space, the proposed legal requirements are:

- Mandatory '1 visible partner' rule where the users know who they are dealing with.
- Recognition of customers' PI ownership, where data is not owned, but only processed by the companies and only usage rights to PI can be traded for customers to better participate in personal data markets.
- Obligation to offer one service option with minimum information use at reasonable quality and price.

The second market space includes the distributed computing and service infrastructures that enable today's business relationships. This space includes all companies providing services to the CR-H that directly enable and enrich customer relationships. The legal requirements for this space are:

- Liability of CR-H for handling personal information in accordance to policies
- Obligation to audit an accountability management platform.

The third market space includes services that grant customers ownership of their personal information and manage it in a privacy-friendly way (enabled by trusted third parties or personal data vaults). The legal requirements for this market space are that there is an

obligation of trusted third parties to act on behalf of the customer and keep the personal information secure.

The fourth and last market space grants equal access to anonymized data to all market entities that need it. Participants in this "safe harbor for big data" can provide and process as much data as they want, but the data they handle must be anonymized. Each time personal information is transferred to this space, it must pass an anonymity frontier that is operated by the best available techniques (BATs), which must be constantly reviewed. The legal requirements for this space are:

- An open and reciprocal access to safe big data
- Heavy sanctions for violation of anonymity requirements
- Penalization for re-identification, even when it doesn't cause any harm

6.2.5. Use case context

Up until now, the four previously defined perspectives must be taken into account in the research regarding privacy-preserving data markets for IoT devices. This last section refers to contemplating the specific use case when considering all the other perspectives. A use case encompasses all the possible sequences of interactions between all the possible systems and users in a particular environment and related to a particular goal [86]. In other words, a use case describes the specific scenarios in which the data market is implemented. The goal of this section is to stress the fact that particular use cases urge for individual solutions throughout the data market research process. We name some of the use cases found in the final studies and describe which specific technologies can be used for them. Therefor, this sections already covers some specific privacy preservation techniques for these individual use cases, which represents a logical transition to the next section.

On of the applications of IoT data markets is machine learning. In machine learning, systems benefit from large quantities of diverse training data [87]. Machine learning techniques such as deep learning, are viable approaches to exploiting the value of big data [88]. Unlike traditional analysis techniques, machine learning is capable of thriving on growing datasets [1]. [87] proposes a decentralized marketplace that runs on a blockchain and uses differential privacy as well as Trusted Execution Environments (TEEs) to automize machine learning processes on private data.

Another use case for IoT data markets is Smart Mobility. [55] introduces a Blockchain framework for a smart mobility data market where the transportation data is shared across multplie entities. The paper describes a six-layer model of the blockchain framework for smart mobility data transactions. In this study, there are specific algorithms designed to anonymize locations in Location Based Services (LBS) such as Cloaks or Geomasking. Another used

technique is a model of differential privacy called Geo-indistinguishability (GeoInd).

[89] and [75], among others, describe the case of crowdsourcing as a use case for IoT data markets. Crowdsourcing has gained has gained increasing popularity over the last years as it can be adopted to solve many challenging question-answering tasks [89]. In crowdsourcing, a crowd of users contributes their efforts, largely reducing the costs of asking applications. However, there are various challenges in this scenario, e.g. the homogeneous quality of answers given by different entities. Because of this, crowdsourcing and crowdsensing scenarios can need game-theoretical-focused market designs more than other scenarios.

Scenarios with Cyber-physical-Systems (CPS) are another field of interest in the research for data marketplaces, especially because they are able to foster device-to-device communication without humans in the loop. [33] analyzes the privacy risks and available techniques in the Internet-of-Production (IoP), which takes the idea of CPS one step further to afford collaboration between manufacturing processes to establish a "production-to-production" communication.

As [32] states, IoT data markets in the health industry need a higher privacy protection than other scenarios due to the level of data sensitivity found in personal records.

6.3. Privacy preservation techniques

RQ3: Which privacy preservation techniques have been used to enable data markets for IoT devices and what are their current deployment impediments?

In this section, we discuss the privacy-preservation techniques, methods and approaches found in our main study selection. As it is the case in previous passages, this section aims at describing the techniques in a summarized and elementary manner since the level of detail for each of them is a research area for itself. Where applicable, we illustrate these techniques with practical examples found in our studies. Most of these examples encompass a combination of the proposed techniques in order to tackle numerous challenges within one self-contained solution.

Since protection techniques provide room for a limited set of operations on the secured data, the selection of these techniques affects the design of privacy-preserving algorithms [32]. For example, a framework using homomorphic encryption (discussed in section 6.3.2 must express its analysis algorithms in terms of homomorphic additions or multiplications.

When classifying privacy-preserving techniques, [32] identifies two basic relevant scenarios: The first is called "outsourced computation", and it refers to everything that happens to

user's data outside the perimeter of their personal devices and home networks. The service providers (SP) in this scenario can be seen as trusted SP or as honest but curious SP (see section 6.1). The second scenario is called "Information sharing", which assumes that data must be shared between data consumers and producers in order to produce useful results. The entities in this scenario cannot always be trusted. While this classification is useful to understand the scenarios in which privacy-preserving techniques may be needed, it is not satisfactory due to its level of inexactness. Pennekamp et al. [33] define five building blocks that categorize security and privacy enabling techniques in the IoP. Based on this work, we identify four blocks for classifying the privacy preservation techniques in data markets for IoT devices. These blocks embody the structure of the following subsections. Figure 6.2 illustrates the classification of the identified blocks as well as the discussed privacy preservation techniques and their supporting technologies. We recognize that some of the mentioned techniques could fall into several of the building blocks, which is why in those cases, we decided to stick with the classification proposed in [33].

6.3.1. Data security

The first block is concerned with data security. As mentioned in previous sections, we define data security as providing confidentiality or protecting information from unauthorized access.

Encryption methods

The most basic form of achieving confidentiality is encrypting sensitive data [33]. This way, only authorized entities, with the help of a decryption key, can decipher a ciphertext back to plaintext and access the original information.

Even though not all of our selected studies explicitly mentioned encryption methods in their solution, we can safely assume that since encryption is already a living part of every IT system, every solution would implicitly contain at least basic encryption technologies in order to preserve data confidentiality. Within the studies that mention encryption, the level of detail varies mightily.
Within the IoT context, data encryption plays an important role since applying supersecure schemes such as AES-256 encryption in order to protect data from potential adversaries seriously jeopardizes computations and services provided by IoT devices [32] [1] [90] [91][89]. The computation costs do not only occur in the encryption or decryption phase, but they also increase the communication requirements and overhead [89]. Another problem with encryption is that many of IoT-relationships can be short-termed, which signifies that the number and IDs of entities that are allowed to access the encrypted information must be updated dynamically. This can also call for much higher costs in key management and data re-encryption efforts [33].

Since symmetric encryption schemes are much faster that asymmetric schemes, one solution that is widely adopted in practical terms is that the data is encrypted with a symmetric

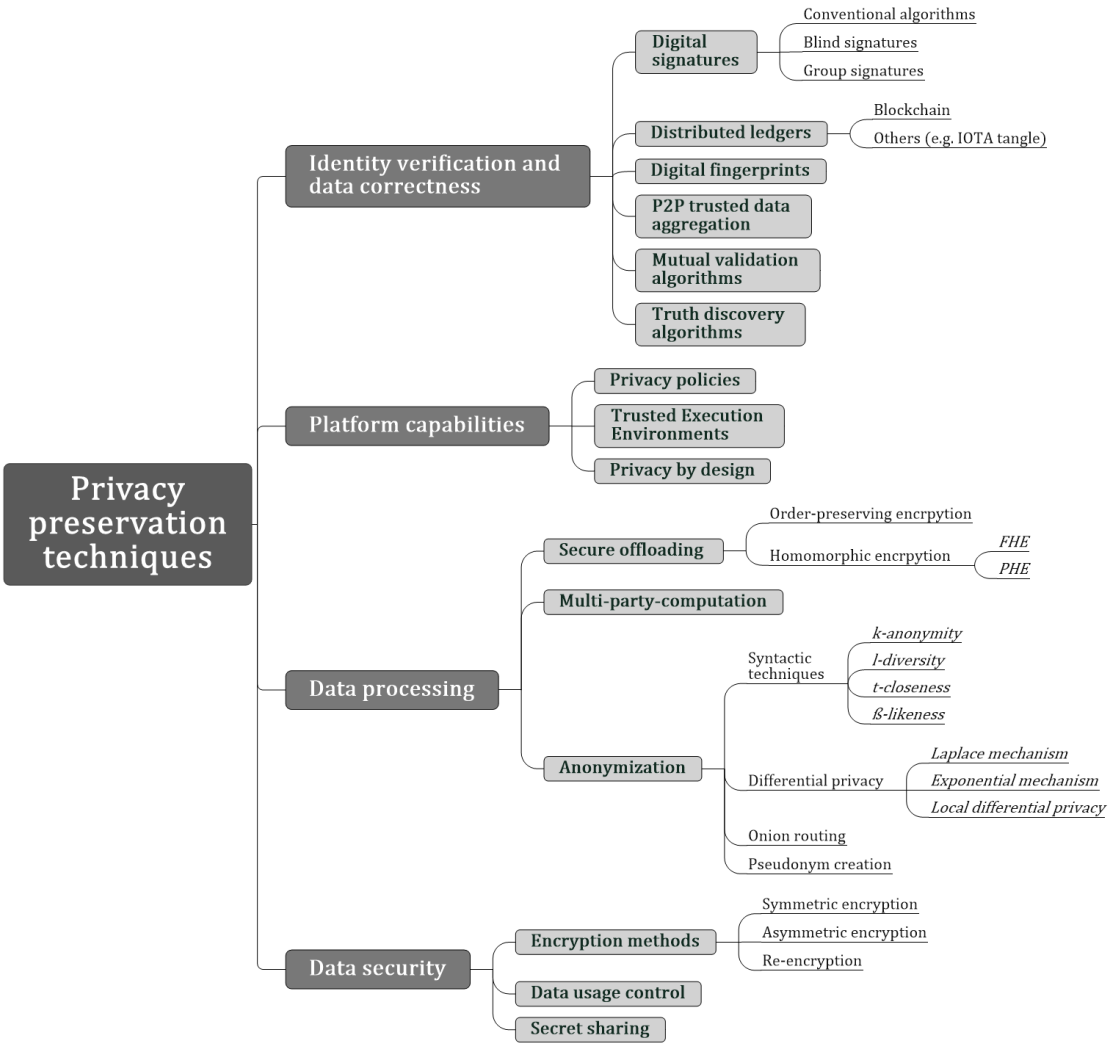

Figure 6.2.: Classification of privacy-preservation techniques and technologies

key, and the symmetric key is then encrypted asymmetrically. This pattern is explicitly implemented by studies such as [77] or [92].

[12] refers to a technique called re-encryption [93], which enables data encrypted under one public key to be transformed to data under another public key without decrypting the message. This technique can be useful for IoT data markets, especially in DLT scenarios (refer to [12] for more information).

Data encryption in the broader sense is not only used to solely provide confidentiality. It can also be used for verification (see digital signatures in 6.3.3) or data processing (e.g. homo-

morphic encryption in 6.3.2), which is why we find encryption schemes in other building blocks.

As for encryption schemes used entirely for confidentiality purposes, the most named encryption algorithm was AES (asymmetric encryption), explicitly mentioned in [15], [85], [94], [77] and [54]. There is no explicit mention of a symmetric algorithm.

Data usage control

Data usage control refers to allowing distributing decision regarding data access to multiple parties [33]. As mentioned in the section above, this problem is significant in the IoT since many of IoT-relationships can be short-termed, which signifies that the number and IDs of entities that are allowed to access the encrypted information must be updated dynamically [33].

[53] proposes a proactive key-rotation system [95] in combination with a key manager. In order to confine the purview of a leaked key, a quorum of compute nodes form a key management committee and run a distributed protocol to manage keys used by contract TEEs. A contract TEE reaches out to the key management committee to create or retrieve keys.

Data usage control, more than a technique, is still a theoretical concept more than an established functional system.

Secret sharing

Secret sharing allows information sharing with several entities while ensuring confidentiality. To reveal the information, a subset of entities, which individually have only parts of the secret, must collaborate to reconstruct the original information allowing a certain degree of data control [33] [96] [97] [32]. Three of our found papers use secret sharing in their implementation ([98], [99] and [74]).

[98] proposes a platform called Sharemind that allows for multicomputing of data in the analysis stage in order to achieve privacy. It uses a combination of Multi-Party-Computation (MPC) and secret sharing. In this framework, each data "donor" processes input transactions using secret sharing and sends one share to a set of data miners, who run an analysis algorithm on the secret shared data. When the miners complete this stage, each miner publishes a share of the results to previously agreed entities who can reconstruct the result. This way, a data provider does not have to trust any of the miners, given that no collision takes place.

[74] proposes a scheme where a Cloud-Service-Provider collects encrypted bids from bidders in an auction scheme and cooperates with another party called the Crypto-Service-

Provider to share the bids using secret sharing.

[99] is discussed in section 6.3.2.

6.3.2. Data processing

This building block is relevant for procedures that require a computational functionality that takes place outside of the physical scope of data owners. The category of data processing covers techniques that aim at hiding information during computations from unintended recipients, i.e., they extend the concept of simply limiting access to data to techniques that can also operate on or with data in a secure manner [33].

Secure offloading

Secure offloading covers cryptographic methods that allow entities to operate directly on ciphertext [33]. After the relevant operations are carried out, the entities with the corresponding key are then able to decrypt the ciphertext in order to obtain a result [100]. Differently from secure computation, in secure offloading, an entity can operate on the ciphertext of the complete plaintext. We identify two techniques: Order preserving encryption and homomorphic encryption.

Order-preserving encryption schemes (OPE) preserve the ordering relationship among encrypted values, enabling the indexing of these values [32]. The upside of these schemes is that they are very efficient in terms of computation. The downside is twofold: on one hand, they assume that potential adversaries do not know the original data distribution, giving them a weak privacy notion [32]. On the other hand, it greatly limits the types of algorithms that can be used in the processing stage.

Homomorphic encryption techniques generally offer a set of operations that support a much larger number of processing algorithms, which is why this technique is used in six of our found studies. The encryption schemes can be classified as fully homomorphic (FHE) or partially homomorphic (PHE) [101]: FHE schemes provide the strongest notion of homomorphism. They support multipliable operations (currently addition and multiplication), allowing more computation to be performed over encrypted data. In PHE schemes, only a single operation can be performed on the ciphertext, for example, addition or multiplication. Some encryption schemes that are partially homomorphic are the ElGamal encryption scheme, the Paillier cryptosystem and the Boneh-Goh-Nissim encryption scheme. Four out of the six studies used PHE, one study uses FHE and the remaining one did not specify.

[77] proposes an auction scheme with four entities. Besides sellers and bidders, there are two more parties: the auctioneer and the intermediate platform (both of them trustless). The core idea of the solution is the following: The sellers and bidders encrypt their data

and the bids using the public key of the auctioneer, so the auctioneer is the only one who could decrypt this data with his secret key. They then send the data to an intermediate platform. There, all bids and data are padded with a randomly generated pad. After this, the intermediate platform sends this bids and data to the auctioneer. At this point, neither the auctioneer nor the intermediary platform can decrypt the data or the bids in order to get the original information. The auctioneer decrypts the encrypted bids (and data), compares the bids and selects a winner without really knowing how high each bid was. This is possible by the selected homomorphic cryptosystem Paillier that was initiated with the bit-padding. The data is then encrypted with the public key of the winner and sent to the intermediary platform. The platform applies homomorphic de-padding to the data and sends it to the winner. The winner can then decrypt the key with his secret key and thus decrypt the sold data. Since everyone (including intermediary platform) has the public key of the auctioneer, the data could be manipulated before reaching the auctioneer, who then analyzed the data in order to get a winner. Therefor, the digital signature schemes are used in the solution. The signature can only be created by using a secret key and can be verified using the public key. The used Paillier cryptosystem [102] supports addition operations (used in the winner determination algorithm) and digital signatures, which is why it is implemented in the study.

Besides [77], there were other studies that described a solution using homomorphic encryption. [57] proposes a system called TPDM (also mentioned in section 6.2). The essence of the solution is to first synchronize data processing and signature verification into the same ciphertext space (Encrypt and then sign), and then to tightly integrate data processing with outcome verification via some homomorphic properties. In order to achieve a tradeoff between functionality and performance, this solution uses a PHE scheme called the Boneh-Goh-Nissim (BGN) cryptosystem. This cryptosystem facilitates one extra multiplication followed by multiple additions. [99] proposes a blockchain-based outsourcing computation (DOC) scheme that is specialized in IoT called BeeKeeper, where servers can perform FHE computations according to the requests of data owners. The used cryptographic primitive is called Fully Homomorphic Non-Interactive Verifiable Secret Sharing (FHNVSS). The proposed scheme allows transactions, including the responses and the supplementary data from the servers, to be publicly verified in a blockchain. [6] proposes two trading schemes that use blockchain: The first scheme achieves direct raw data exchange, while the second scheme achieves data statistics trading using homomorphic encryption and data organization in a data structure called Merkle Accumulative Tree (based on normal Merkle Trees), where the leaf nodes are encrypted data and the non-leaf nodes contain a hash value and a cumulative sum of ciphertexts. [76] proposes an auction scheme that uses homomorphic encryption, blind signatures and onion routing for a third party to not gain information on the biddings and IDs of the bidders.

The main problem of homomorphic encryption is that it requires very high computation complexity and overhead that is not possible for all data consumers and poses a great challenge in the IoT context, as discussed in section 6.1 [103], [77], [32], [57]. FHE is several

times more costly than PHE [32]. [32] also states that additive homomorphic encryption schemes (such as Paillier) must be complemented with other methods in order to guarantee full protection.

The interested reader can refer to [104] and [101] for further details on theory of homomorphic encryption schemes.

Secure multiparty computation

In secure multiparty computation (MPC), multiple entities can jointly compute a function without revealing individual inputs to each other [105] [32], i.e., we refer to protocols between multiple distrusting stakeholders to either jointly compute a result or to exchange information obliviously [33]. The main purpose of these protocols is to defend the mechanisms that make sure that parties don't learn more information than their own when processing data.

One special case of MPC is the two-party computation. Two central concepts of this case are oblivious transfer and the principle of garbled circuits: Oblivious Transfer (OT) is a cryptographic primitive which in its simplest flavor (1-out-of-2 OT) a sender has two input messages M_0 and M_1 and a receiver has a choice bit c. At the end of the protocol, the receiver is supposed to learn the message M_c and nothing else, while the sender is supposed to learn nothing [106]. In other words, a sender doesn't know which piece of information it has sent to the receiver. Garbled circuits are protocols that enable two-party secure computation where the function must be described as a Boolean circuit providing basic logic gates such as AND, XOR and OR [107][32]. Garbled circuits make use of oblivious transfer protocols to function.

[74] presents an auction solution that cryptographically hides the bids from all participants until a winner is determined. It uses XOR-homomorphic encryption, secret sharing as well as garbled circuits. The data oblivious auction algorithm and its basic operations are implemented as an auction circuit and then used in a cloud-based framework.

[98] proposes a platform called Sharemind that uses MPC as its central concept. In this solution, the data donors (owners) use secret sharing (discussed in section 6.3.1) to send their data to data miners in order for them to analyze the data.

The advantage of MPC using garbled circuits is that the circuits are extremely expressive [32]. The drawbacks of MPC are that they require expensive computation and intensive communication, which can significantly increase the network latency [32] [89] [98]. Another challenge in MPC is that the protocol must provide mechanisms so that the entities do not collide [98].

Furthermore, MPC protocols reduce the accountability because the individually provided inputs are only locally available and no external verification is possible without cooperation

[33]. One enhancer for MPC is the principle of zero-knowledge-proof in order to enforce honest behavior while maintaining privacy. The zero-knowledge proof could be used for an entity to prove that its behavior is correct according to the processing protocol. Because of the principle of zero knowledge proofs, the entity would not have to compromise the privacy of its private data in order to provide the proof [108]. Zero-knowledge-proofs have recently been used as a verification mechanism in some blockchains such as ZCash or ZCoin [109], which is relevant for us since we will be discussing blockchain in section 6.3.3.

Anonymization

Non-explicit identifiers and sensitive attributes, as described in previous sections, can be protected with anonymization methods [29] [110] in order to avoid re-identification attacks, which is why anonymization is a key element in modern online scenarios [20]. However, without proper theoretical foundation, anonymization can suffer from background-knowledge-based attacks [32]. For this reason, the methods must be carefully implemented and selected based on each scenario. One key difference between anonymization and secure oflloading or MPC, is that anonymization protects information to be read by anyone outside of a given frame, and not only by parties within the data processing stage. In a data market scenario, this means that if the data is correctly anonymized, the receiver (buyer) will also not be able to reverse the anonymization function in order to gain specific information and identify the data owner. In this sub-section we describe our findings for the most used anonymization techniques found in the selected studies. Some of these studies, such as [68], define data suppression (elimination) as part of the anonymization palette. Even though this definition may be satisfactory for some scenarios, we will not concentrate on it and will therefor not analyze which data should be suppressed.

[111] classifies data anonymization techniques into two categories: syntactic techniques aim at satisfying a syntactic privacy requirement, for example, that each release of data must be undistinguishably related to no less than a certain number of individuals in the population (see k-anonymity below). Semantic techniques aim at satisfying a property that must be accomplished by the mechanism chosen for releasing the data, for example, that the result of an analysis process must be insensitive to the insertion or deletion of a tuple in the released dataset (see differential privacy below). Semantic data protection techniques have recently been proposed to protect the privacy of both data respondents and individuals who are not included in data undergoing public release [111]. While syntactic techniques traditionally guarantee data protection preserving the truthfulness of the released information, semantic techniques typically add noise to the released data. Noise addition perturbs the original content of the dataset, thus achieving privacy at the price of truthfulness [111].

Syntactic techniques (k-anonymity, l-diversity, t-closeness).

In the set of semantic techniques, we identify three main definitions: k-anonymity, l-diversity and t-closeness. For k-anonymity, a dataset release is defined as having the k-

anonymity property if the information for each individual contained in the release cannot be distinguished from at least $k-1$ individuals whose information also appears in the release [60]. Since the problem of optimal k-anonymization (i.e., minimum-information- loss) is NP-hard, researchers have proposed several heuristics [112]. The most used heuristic is a method called identity generalization [29], which consists of substituting the original values with more general ones.

[23] is a good example of how generalization can be used. The study proposes a dataflow-focused scheme to implement privacy policies (see section 6.3.4) defined by a user given a certain context (a condition). The anonymization level in the solution is set by defining a view over the set of fields (the data tuples). To construct views, the application relies on Domain-Generalization-Hierarchy (DGH), where the parameter k (given by the user) indicates the number of generalization steps that have to be applied in order to obtain the desired view of the field. For example, given the value $v = 104$ for a field f, the two functions $h1$ and $h2$ with $h1(v) = 100$ and $h2(h1(v)) = 50$ are the 1-generalization and the 2-generalization results, respectively [23].

As another example, in Location Based Services (LBS), k-anonymity can be achieved by hiding the real location in a set of similar but fake locations. Dummy locations, cloaks and geomasking are some of the methods that can be used for achieving k-anonymity in LBS [55]: in dummy locations, a set of fake locations is sent to the LBS together with the real locations. Cloaks involve sending a region containing the real location. Geomasking implicates randomly displacing the real locations outside of an inner circle but within an outer circle (the real location is displaced "inside a donut") [55].

[29] proposes a framework where a set of k-cooperative users can jointly work by forming a crowdsensing coalition, increasing the anonymity privacy protection measured by k-anonymity. The total coalition payoff is divided among the cooperative users based on their marginal contributions to the total data quality at the end of the sensing service.

One main problem with k-anonymity is that it does not work well if sensitive data attributes lack diversity [29]. For this reason, another semantic technique called l-diversity requires that each equivalence class has at least l "well-represented" values. We define an equivalence class as a set of data samples with the same anonymized data attributes [29]. Two remaining issues of this solution are that firstly, this property fails to protect the data against attacks arising from an adversary's knowledge of the frequency of each sensitive value and secondly, that it fails to guarantee privacy when the distribution of the values differs substantially among equivalence classes and from the overall distribution [110]. As a rectification to this problem, a third technique called t-closeness proposes a condition where the distribution of sensitive values within each equivalence class must be close to their distribution in the entire original dataset. Technically, t-closeness can be achieved by adding random noise to sensitive data attributes [29]. [110] states that these approaches are not effective enough for microdata

publishing, which is why it proposes a model called ß-likeness. The goal of this model is to ensure that an adversary's confidence in a tuple's sensitive attribute's value should not increase in relative terms by more than a threshold after seeing the published data. It protects infrequent values by providing them with special attention, while more frequent values are disallowed from assuming frequency values of 1, which means that they are not allowed to always be in an equivalent class. The ß-parameter defines the privacy constraint for less frequent values.

[68] states that a problem for data sharing (or data trading) scenarios is that modern methods that aim to achieve k-anonymity protect all records homogenously, which can hamper the utility of the data. The is why the study proposes a policy and context-based approach to redefine anonymity requirements on the current (mainly time depending) context. We discuss this solution in section 6.3.4 in relation with policies implementation. Another problem for these approaches mentioned by [16] is that they are not well-suited to be applied in the context of a multi-source data market.

Differential privacy.

After having discussed syntactic anonymization methods, we now give a short overview of the semantic techniques. As mentioned above, these techniques aim at providing semantic privacy guarantees defined by the data holder prior to data publication. Semantic data protection techniques have recently been proposed to protect the privacy of both data respondents and individuals who are not included in data undergoing public release [111]. The central technique that we discuss in this section is differential privacy because it is the most widely mentioned in research by a large margin. Out of 36 concrete solution proposals found in our selected studies, 13 of them, which amounts to almost 37%, use differential privacy as a privacy-preserving technique. [80] states that differential privacy is the existing de facto standard for privacy protection in statistical databases. The success of differential privacy in recent year lies in its rigorous theoretical foundation. As opposed to k-anonymity, l-diversity and t-closeness, differential privacy can be applied in the context of a multi-source data market with less effort [16]. In technical terms, we define this technique as follows: "given two datasets T and T' differing only in one tuple, an arbitrarily function K (typically the release function) satisfies $\epsilon - differentialprivacy$ if and only if $P(K(T) \in S) \leq exp(\epsilon) \times P(K(T') \in S)$, where S is a subset of the possible outputs of function K and ϵ is a public privacy parameter" [111]. This means that the released dataset satisfies $\epsilon - differentialprivacy$ if the removal (insertion) of one tuple from (into) the dataset does not significantly affect the result of the evaluation of function K. Generally speaking, differential privacy is achieved by adding randomized noise (typically using the Laplace distribution) to query output, so that the adversary cannot identify if one individual is or is not in a dataset, no matter what side information the adversary has [16]. Knowing the noise distribution allows to compensate it when analyzing the released dataset.

Differential privacy can be implemented using a series of mechanisms. As mentioned before, the most used mechanism involves using the Laplace distribution. This mechanism, along with the exponential mechanism and the method of local differential privacy, are the ones that we found in our studies. Other mechanisms such as $(\epsilon, \delta) - differentialprivacy$ [111], which is a relaxation of the definition above, were not included in our studies. Out of the 13 studies that use differential privacy, 3 mentioned using the Laplace mechanism, two use local differential privacy and one implements the exponential mechanism. The rest of the studies did not specify.

[80] suggests a data market framework with a market manager that acts as a broker between the provider and the consumer. In this data market, a negotiation technique is proposed in order to determine the noise parameter ϵ used in $\epsilon - differentialprivacy$. [92] is an addition to this study, where the proposed framework is implemented and tested in a web application. [78] proposes a contract design approach within a data market to find the optimal contracts when using a biased algorithm to provide privacy. The key component of this solution is an algorithm that not only adds a zero-mean noise to the private data, but also a bias. In this bias, the data broker only collects data from those entities that value their privacy at a low enough level to make participation in a given database worth-while. By choosing the bias term, a contract can be designed for the buyer to obtain the desired accuracy level at a lower cost. [75] aims at tackling the problem of strategic behavior by the "workers" (the data providers) who may be tempted to add more noise into their sensing data to enhance their data privacy. The idea is that the noise parameter for each worker is determined by externalities induced from other workers' participation, thereby indirectly determined by the platform. The authors note that in practice, workers may refuse to participate if they are allowed to add only a low level of noise into data. [103] proposes a crowdsensing scheme called DPDT that anonymizes both buyers and sellers with differential privacy. In this design, there are two basic steps, data pricing and data collection: In data pricing, the buyers define for which price they are ready to buy some data. Each buyer has a private (real) valuation and a public bid. Which may not be truthful. This is allowed since the authors state that absolute truthfulness in this scenario may lower the revenues in the platform (approximate truthfulness). In the data collection step, the platform concentrates on selecting a minimum number of crowd workers to collect sense data in order to reduce costs. For this, it publishes the real data-collection tasks together with some fake collection tasks in order to protect the buyer's privacy, both tasks having a similar distribution but being subjected to differential privacy (exponential method). The results of the sensing task are also perturbed with the exponential mechanism on order to protect the sellers. [55] applies differential privacy in the context of Location Based Services (LBS) with a special technique called Geo-indistinguishability (GeoInd).

[32] and [68] identify two problems with differential privacy: firstly, they state that differential privacy may not be ideal for applications requiring the highest quality models. Secondly, with today's approaches, it is difficult to satisfy differential privacy in a non-interactive setting. Moreover [113] mentions that an individual's private attributes could be inferred from

differentially private data with non-trivial accuracy, while the added noise could dominate small values in the results of aggregate queries.

The set of selected studies that use differential privacy or other noise inducing methods in their proposed solutions are [87], [55], [89], [78], [75], [16], [27], [103], [53], [60], [80], [92] and [114].

The interested reader can refer to [111] for a comprehensive summary on syntactic and semantic privacy preserving techniques and further use-cases that are more general not related to IoT data markets.

Other anonymization techniques.

A further anonymization technique we find is onion routing, where after a series of re-transmission steps through a node network, a sender remains anonymous because each intermediary node knows only the location of the immediately preceding and following nodes [17]. This technique is used in [76]. The most widely used implementation of Onion Routing is the Tor Project [115]. While appealing, there are some drawbacks for this solution: Implementations such as Tor are designed to support low-latency communication, which is incompatible with our vision of IoT data markets. Moreover, Tor is often blocked by IT departments within organizations or even subject to state-level censorship by some governments [116]. These, among others, are reasons why onion routing cannot be the (only) anonymization technique in IoT data markets.

Moreover, [57] and [31] mention pseudonym creation as a further anonymization technique, which guarantees a very low level of privacy protection since it does not protect the entities from task tracing or re-identification attacks.

6.3.3. Identity verification and data correctness

As discussed in section 6.1, two of the main challenges for privacy-preserving IoT data markets are guaranteeing data truthfulness and ensuring accountability, which means providing mechanisms that can hold a party accountable for undertaking some action. These two requirements for a functioning data marketplace are intuitively contradictory to the privacy demands that we aim to research and satisfy in this thesis. This section refers to mentioned mechanisms that enable, at least to a certain degree, the conjunction of both requisites. [33] describes these mechanisms as approaches that range from providing physical aspects of work piece to providing evidence for the origin and correctness of digital information. [33] also states that "while different in scope, these approaches have in common that their ability to attest the authenticity and integrity of information contradicts the desire of stakeholders to remain [entirely] untrackable".

Digital signatures

Digital signatures are a family of primitives that rely on asymmetric cryptography. They are usually employed to provide data integrity, data origin authentication and non-repudiation (actions of an entity cannot be denied) [15]. Customarily, authentication processes with digital signature algorithms consist of three steps [15]:

- Key generation based on a security parameter. The result of this step is a public key pk as well as a corresponding private (secret) key *sk*.
- The signature step is done based on a message *m* and on *sk*. The result of this step is the digital signature.
- The verification step is based on *m*, on *pk* and on the signature. The result is a Boolean value that is only true if the signature is a valid signature of *m*.

There are several implementations of digital signature algorithms, such as RSA, EGamal (used in [57]), DSA or ECDSA (used in [99] and [85]), each of them with a series of advantages and disadvantages. The interested reader can refer to [117] for further details. In this section, we describe some solutions proposed our selected studies and discuss how these solutions deal with challenges and opportunities that arise from using digital signatures:

One challenge occurs in scenarios that include a third party: If a digital signature scheme is applied to the plaintext space, the data consumer, be it buyer or third-party, needs to know the content of raw data for verification [57]. However, by employing a conventional public key encryption scheme to build the ciphertext space, only the buyer can verify this information in order to preserve the privacy of the seller. If a third-party service provider (e.g. a data analyst) must verify the data, it automatically must also have access to the raw data and not only the encrypted one. As a solution, [57] uses partially homomorphic encryption in order to facilitate both data processing and verification. The essence of the solution is to first synchronize data processing and signature verification (with ElGamal algorithm) into the same ciphertext space (Encrypt and then sign). This paper is also discussed in section 6.3.2.

[79] uses the concept of blind signatures between a data provider (owner) and a broker in an IoT data market in order to achieve data truthfulness while respecting privacy from a third party. The data provider sells access keys to data streams (called session keys) to data consumers so that these can access the data streams for a certain period. The promise of the provider is that only one session key will be available for each product. A broker party is in charge of creating the channels where the streams are placed as well as certifying that the session keys are in fact as promised. In order for the broker to not steal the keys for himself, the data provider "blinds" the session key with the broker's public key and send the blinded key to the broker. The broker signs (certifies) the key and returns the signature to the provider who removes the blinding factor. The provider can then send the session key as well

as the signature for him/her to verify the certificate. [76] does also use blind signatures in the proposed solution.

[94] proposes a blockchain-based fair data trading protocol that makes use of ring signatures to enhance the anonymization of the data provider's identity. Ring signatures provide full anonymity: in a ring signature [118], a user chooses a group of users called a ring to generate a signature. A verifier is convinced that the signature is generated by a member of the ring but cannot reveal which member actually generated the signature. Ring signatures are part of so-called group signatures. Within group signatures, there are some techniques that provide a high degree of anonymity but can also be traced back to the group member that issued the signature in order to ensure accountability. These types of signatures were not mentioned in our selected studies. The interested reader can refer to [119] for further details.

[77] implements digital signatures in a solution that also uses homomorphic encryption. The authors use the Paillier cryptosystem since it supports digital signatures in the encrypted data.

[57] and [30] mention a central challenge with digital signatures that had no satisfactory solution: digital signature schemes, which verify the received signatures one after another, may fail to satisfy the stringent time requirement of IoT data marketplaces while also incurring significant communication overhead because of the required maintenance of digital certificates, becoming a process bottleneck in IoT scenarios with limited computational power. None of the selected papers did not have a satisfactory solution to this problem.

Distributed ledgers

Distributed ledger technologies (DLT) have proven to be a suitable approach to improve immutability features and auditing processes in data trading systems [33]. They allow establishing a persistent decentralized record of information and past dataflows [33]. DLT are software infrastructures maintained by a peer-to-peer network in which the network nodes must reach a consensus on the states of transactions submitted to the distributed ledger in order to validate them [60]. In this section we mainly refer to the DLT type that is most mentioned in our selected studies: blockchain.

In blockchain, the transactions within a period are packed into a block in regular intervals. Miners, which are nodes in the network, are responsible for validating the transactions in the block and finding a hash that satisfies a cryptographic nonce. For each block, the node that is the first of the network to find the right hash and validating the block can get rewards in the form of a cryptocurrency. The later block is connected to the former block to form a chain of blocks. This process is supported by other techniques such as digital signatures and hash pointers (each block has a hash pointer to the previous block) [15]. This way, the ledger is recorded and stored permanently. There are four types of blockchains [55]: public closed, public open, private closed and private open. In public blockchains, anyone can do

transactions and have access to the ledger, which is why, in the context of data markets, raw and unencrypted private data should be stored off-blockchain in scenarios using public blockchains [55]. One of the key aspects of blockchain technology is the design of the consensus algorithm. Most designs are based on byzantine fault tolerance [12]. In the context of blockchain, technology enabling smart contracts is widely used. A smart contract is a set of promises in digital form that include the protocols within the involving parties perform on these promises. They not only define the protocols, but they enforce them in order to reach agreements between distrusting parties. While typically existing in blockchains, they further enforce trust through integrity assurance, meaning that even the creator (programmer) of a smart contract cannot feasibly modify its code or subvert its execution. In DLT scenarios, smart contract systems replicate data and computation on all nodes, so that individual nodes are able to verify the correct execution of the contract. Smart contracts have a unique address and they act as independent actors whose sole objective is to transact the assets given in a certain set of rules that involved parties agreed upon [15] [53]. Out of the 36 selected studies that proposed a technical solution for a data market, 18 (50%) use DLT. Out of those 18, 16 are blockchain-based, while 7 of those 16 use the Ethereum blockchain. Other named blockchains were the Oasis blockchain (used in [87]), the Hyperledger Fabric blockchain (used in [55], [15] and [85]) and the Tendermint blockchain (used in [53]). The remaining studies did not mention which blockchain they use for the implementation and testing.

Besides providing a platform that enhances accountability and correctness verification by the trustless nodes in the network, using blockchain as a base technology in data markets has other advantages: it contributes a decentralized architecture that removes the problem of a single-point of failure as well as reducing costs that could be issued by an intermediary third party. It also provides consensus mechanisms that do not require trust in single nodes, but in the majority of the nodes, thus resolving the problem of third-party trust. This fact is also enhanced by the inalterability of the records in the blockchain, which enhances accountability. Moreover, several blockchains offer the integration of smart contracts that can be programmed in order to realize complex transfers. Furthermore, blockchains also provide convenient ways of peer-to-peer payment in form of cryptocurrencies [6] [94]. Nonetheless, the use of blockchains in privacy-preserving data markets also brings some challenges:

Firstly, although smart contracts enable verifying correctness, they also require disclosure of contract inputs, which could include private data. This also poses an economic problem since users in the blockchain could copy the data and sell it on other platforms. As a solution to this problem, [87] uses so-called Trusted Execution Environments, which are discussed in section 6.3.4. Since smart contracts allow the enforcement of a data provider's constraints, one of these constraints could refer to the level of privacy on the sold data, which can be enhanced through perturbation methods such as differential privacy. [87] uses this principle and combines TEEs with differential privacy in order to achieve the desired privacy requirements by each user.

For our purposes, another issue that arises is associated with the physical properties of IoT devices: blockchain principles demand high resources for the consensus mechanisms such as Proof of Work (PoW), while causing high latency for transaction confirmation. [51] proposes eliminating the concept of PoW and the need for coins. The solution is developed in a case study optimized for smart homes but can be used in other contexts, especially in industrial scenarios. The idea is the following: Each smart home is equipped with an always online, high resource device, known as "miner" that is responsible for handling all communication within and external to the home. Each smart home has local storage and a local private blockchain that is stored and mined by one or more resource capable devices, which is always online. All transactions to or from the smart home are stored in the local private Blockchain.

[12] addresses the problem of high resources demand by also using (Intel-manufactured) TEEs and a concept called "Proof of Useful Work" (PoUW). The smart idea is that the miners compute useful work for Intel, and in return Intel provides workers with a proof of their work so that the workers can build a block. The study combines this mechanism with a technique called certificateless cryptography, which solves a further problem of blockchain implementations: the only way to check whether an IoT device has rights to access certain data or not is to verify some credentials only known to this device. This issue is crucial in scenarios that use blockchains, since the verification credentials are exposed to all nodes. In other words, the problem is that the miners should not have any knowledge of the credentials to perform authentication. This implies that the system must have some cryptographic mechanism that allows an IoT device to be identified and verified by other parties without utilizing a secret value such as password. Certificateless cryptography is a technique where a trusted third-party called key generation center creates a partial private key based on a user's identity. The user then utilizes the partial private key and its own secret value (mostly randomized) to establish a private key. Since the secret value is only known to the user, the key generation center is not able compute the private key. The user also creates the public key based on the secret value and makes it public. With this solution, a public key can be verified whether it belongs to certain user or not. The only drawback of certificateless cryptography is that the public key of a user, even though can be verified, needs to be pre-broadcasted.

[53] proposes a platform that also combines TEEs and blockchain protocols with smart contracts. The study focuses on mitigating some of the challenges of combining TEEs and blockchains such as manipulation of host scheduling, unrelated attack vectors, key management and TEE crash. This study proposes a variety of detailed techniques specific to this problem as well as a framework to increase blockchain performance.

[99] tackles storage related issues in the blockchain. It proposes a blockchain-based outsourcing computation (DOC) scheme that is specialized in IoT called BeeKeeper, where servers can perform FHE computations according to the requests of data owners.

[61] not only tackles the identity verification and data correctness problem, but it also

combines this with data processing within the platform in order to address the problem of reidentification when aggregating the data as well as the problem of copyright (user selling the data somewhere else). The buyers are not able to obtain access to the seller's raw data, as they only get access to the findings that they require.

A blockchain-based trading protocol using ring signatures and similarity learning is described in [94]. This approach implements a method for data verification and correctness, and it gives the data consumer a chance to "take a look" into the sold data in order to decide if a purchase is worth. The key idea here is that the data provider sends encrypted data in chunks to the consumer, where each chunk is encrypted with a different key. The data consumer can randomly challenge some data blocks and the data provider responds with the unencrypted data as well as the secret key to decrypt the challenged chunks. This way, the consumer can verify the encrypted data he/she received for correctness and quality by comparing the ciphertext and the unencrypted data. The data consumer can finally use similarity learning in order to decide if he/she is going to purchase the data or not.

Besides blockchain, the other DLT that was mentioned in our selected studies was the IOTA tangle. IOTA is a DLT that aims to solve problems of scalability [17]. The IOTA ledger is structured as a Direct Acyclical Graph (DAG) called the Tangle [120], where graph vertices represent transactions and edges represent approvals: to issue a new transaction it is necessary to approve two previous tip transactions (transactions that have not yet been approved) [17]. IOTA also uses PoW. On top of the solution, Masked Authenticated Messaging (MAM) [121] is a protocol that allows transmission, access, and verification of encrypted data streams [79]. The IOTA Foundation launched a data marketplace [122] prototype that not only allows putting data on Tangle without any trusted cloud services, but also enables trading on Tangle where privacy and integrity meet with MAM, although [79] and [60] state that the platform has two issues: It is centralized at the point that new devices require manual approval from the IOTA foundation, it is still unscalable (even if it's faster than for example the Ethereum blockchain) and that with increasing requests, the number of failing nodes and the amount of noise also increase highly. [79] proposes a solution that combines the IOTA tangle for channel publication of data (via Masked authenticated Messaging) and Ethereum Smart contracts. [62] uses IOTA as an example, but it mentions that the solution can be deployed in any other DTL platform.

The selected studies that included a DLT solution are [15], [87], [55], [99], [99], [12], [123], [79], [53], [94], [85], [52], [61], [6] and [21].

Other approaches

Besides digital signatures and DLT, we identify four more techniques that are used for identity verification and data correctness:

Digital fingerprints [33] are unique identifiers on a physical work piece. Even though

they work in theory, they are difficult to realize in practical scenarios because securely and verifiably attaching a unique identifier to a physical object is not always possible.

[3] mentions a mechanism of peer-prediction-based trustable data aggregation [124], which creates incentives for honest reports and therefor enhances data correctness. In this scheme, a participant is rewarded for its success in predicting the outcome of a random event involving data from other participants. In addition, the participant's utility is defined as the difference between the payment received from the data analyst and the cost resulting from the privacy losses when applying the mechanism. This technique results in almost all participants choosing to report their bids truthfully [3].

Moreover, [3] also proposes making use of mutual validation in the IoT context by benefiting of the fact that in some scenarios (e.g. if the IoT devices are physically very close to each other), there should be a high correlation among the sensed data, which could be used for data correctness verification purposes.

Lastly, we refer to a concept called truth discovery, which encompasses a series of algorithms aiming at finding the true value for a dataset in the case that different data sources provide conflicting information on it, e.g. majority voting. [89] proposes a crowdsourcing solution that uses truth discovery in order to enhance the data quality. Concretely, the study describes a two-layered mechanism that combines data perturbation using local differential privacy and a truth discovery algorithm. The core idea of the algorithm is that if a candidate answer is supported by many high-quality users, it is more likely to be a true answer. Meanwhile, if a user provides many accurate answers, he/she is assigned a high weight, which results in a highly coupled process.

6.3.4. Platform capabilities

This section describes approaches that strengthen the platform's capability to practice certain rules. In other words, it encompasses centrally deployed mechanisms that define and enforce rules to enhance data privacy [33]. One could argue that other mentioned technologies such as access control mechanisms (discussed in section 6.3.1 or smart contracts (discussed in section 6.3.3) could also fall into this category, but we decided to stick as closely as possible to the structure defined in [33].

Privacy policies

The policies approach is best suited if one regards the definition of privacy as "the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others" [22], meaning that users that sell their private data in data markets and understand exactly what data they are selling and to whom, have full privacy concerning this data. In studies proposing a policies approach, data owners

choose among a set of privacy policies (e.g. how much should a specific dataset be perturbed using k-anonymization as a metric) for each set of sold data. We found six papers ([18], [23], [50], [85], [68] and [54]) that explicitly declare using policies as part of their solution, while many others at least mention similar ideas. For example, [20] does not provide a concrete implementation, but it mentions that in data sharing scenarios, the data owners should be able to control four basic aspects: What types of data are shared, with whom (and how much is the other party trusted), for what purposes and for which benefit.

[18] provides a set of legal requirements and high-level technical solutions that facilitate the introduction of policies in international data markets. An example of the technical requirements is to have a policy language with standard vocabulary. We discuss this paper in section 6.2. In the solution proposed by [50] (which dates all the way back to 2002), the data owner is able to choose among a set of 4 privacy policies (including how data is aggregated). The amount of information transmitted is minimized by using pseudonym switching. [85] proposes a blockchain-based data market that includes two third parties: a data broker and a privacy policy manager (PPM). The PPM manages the consent information of data owners when data is requested by data brokers, i.e. it manages an individual's privacy settings and provides information for controlling flows of private data according to those privacy settings. The data broker is then responsible for anonymizing the dataset using k-anonymity as a metric.

[68] proposes a one-to-one privacy preserving mechanism where the disclosure of information is not only decided based on the data to be disclosed, but also on past data that has been published before, and thus could be used to carry-out correlation analysis. This solution tackles the problem that if collected data increases all the time, data that may once been k-anonym may no longer be. The proposed infrastructure manages both the requirements and the data record, which is sorted by the time of publication. The sorting is used to take into consideration that privacy risks associated to data tend to diminish as time passes.

[23] introduces a policy enforcement algorithm together with a model and a language for specifying policies in dataflow computing. In dataflow computing, applications are seen as directed graphs where edges represent data streams and nodes represent functional operators (e.g. reduce or filter) as well as data sources or sinks. In the proposed algorithm, if the context of a situation changes or of two policies collide, the dataflow is rewritten. This exploits the nature of the dataflow-based applications by inserting connections to new operators. The policies use domain-generalization-hierarchy (DGH) in order to anonymize the data (discussed in section 6.3.2).

The implementation of policies also has some disadvantages. Firstly, there may be multiple colliding policies, meaning that applications using this approach must also model how policies are to be weighted in specific contexts [23]. Secondly, there is no uniformly accepted global standard for electronically representing privacy policies [18]. [54] surveys

this problem by investigating how privacy has been modeled in different domains with a focus on IoT applications. Thirdly, policy enforcement also causes overhead and latency increase induced by all filters that data must pass [23]. Furthermore, privacy preferences of individual users must by acquired with minimal human intervention, since asking too much information about preferences might overload users. [54] suggests using recommender systems that are based on similar users' data in order to address this issue. Moreover, the costs of data acquisition for privacy policies should not be too costly regarding computational resources, since they can scale exponentially for a large number of transactions [54].

Trusted computing and Trusted execution environments

Trusted computing refers to technologies that enable that a computer behaves in expected ways by having hardware pieces with unique encryption keys, which makes software tampering of protected programs at least as difficult as hardware tampering, as the user would have to hack the trust chip to give false certifications in order to bypass remote attestation and sealed storage [54]. Trusted execution environments (TEEs) are secure areas of a main processor that guarantee code and data loaded inside to be protected with respect to confidentiality and integrity [87]. Four of the selected papers use TEEs in order to enhance these characteristics: [87], [12], [53] and [61]. All of these papers use Software Guard Extensions (SGX), which are security-related instruction codes built into modern Intel CPUs where TEEs can be built. SGX are able to prove the correct execution of programs (attested execution) by issuing remote attestation, digital signatures and the private hardware key over the program and an execution output [61] [53]. The regions of memory used for the execution of the code are called *enclaves*.

[87] proposes a decentralized marketplace that runs on a blockchain and uses differential privacy as well as Trusted Execution Environments (TEEs) to automize machine learning processes on private data. Concretely, the machine learning pipeline begins with the TEE remotely attesting the veracity of the consumer smart contract. After this, it runs the smart contract. The authors mention that this approach is more efficient than homomorphic encryption or MPC.

[61] introduces a secure data trading ecosystem, where both data broker and buyer are no able to obtain access to the seller's raw data, as they are only getting access to the analysis findings that they require. This tackles the problem of data reidentification as well as the problem of users re-selling raw data in other platforms. The solution is implemented in the Ethereum blockchain and divides the nodes into normal and trusted nodes: normal nodes have one Ethereum Virtual Machine, while the trusted nodes have two EVM, where one EVM is protected by SGX. The traditional contract is executed on the unprotected EVM of normal and trusted nodes, while the data analysis contract is only executed in the SGX-protected EVM on the trusted nodes. A further system counts the number of different results gotten by the trusted nodes, where the largest number of identical results is used as the final result.

Finally, the trusted nodes who have this final result are rewarded.

The blockchain-based solution proposed in [12] uses SGX in order to mitigate computation issues introducing the concept of Proof of Useful Work (PoUW). The solution in [53] uses SGX too while also describing the challenges of unifying blockchain technologies with TEEs. Both studies are discussed in section 6.3.3.

Privacy-by-design

Privacy-by-design refers to considering privacy issues at every stage of the data management. This results in consumers and third parties (such as data brokers) collecting no more data than the one they need and the platform implementing measures to refrain from collecting intrinsically sensitive information [68].

[1] mentions a minimize design strategy (included in [125]) which aims at providing the minimum amount of data at each time interval and further increases the price for larger data packages. One paper that is focused on the overall design perspective is [33]. This paper presents "a set of guidelines, as the core of a conceptual framework, that incorporate privacy-by-design principles to guide software engineers in the systematic assessment of privacy capabilities of IoT applications and platforms". The study mainly focuses on providing design guidelines to prevent unauthorized access (see section 6.3.1) and secondary usage issues (see section 6.1).

6.4. Future directions

RQ4: What are the future research directions of the application of privacy-preservation techniques on data markets for IoT devices?

In this section, we discuss the future directions that researchers in the selected studies have mentioned. In the previous sections we outline current challenges, research topics and concrete technologies that are being used or proposed in order to preserve privacy in data markets for IoT devices. We also mention the issues that these solutions pose. One could argue that each of these challenges is a subject for future research, since it hasn't been solved. While this is technically true, in this section we do not focus on every single technical issue that the solutions pose, but more on overall research trends that lie open for future research as described in the selected studies.

Some of the bigger issues mentioned have a clear research path ahead, while others do not. For example, one of the still unresolved challenges is that (raw) data purchased in one data

market can be sold in another. This is possible because of the nature of data itself and none of the selected studies was able to provide a satisfactory answer as to how this challenge is solved. In fact, [62] states that the only way to mitigate this problem is through legal measures. [77] proposes using digital watermarks (which are still not maturely researched) for this problem. This issue is an example of a scenario where the future research does not have a clear path ahead (at least as mentioned by our studies). We focus on research that does have a clear path ahead and identify five research directions:

The first concerns the physical abilities of IoT devices as well as the research of efficient algorithms. As mentioned in section 6.1, one core problem of IoT devices is that, being lightweight, they do not possess computational abilities for securing the data against privacy-related attacks on a large scale. Specifically, IoT devices are currently not well suited for making computations that require complex encryption, digital signatures, analysis of data, among others. In the proposed solutions, many papers solve this problem by shifting big computations into other parties. While these solutions sound logical, they also open up a series of issues concerning third party trust that could pose an even bigger threat than the original ones. Solving the computational and scalability challenge inherent to IoT devices would tackle many of the problems created by current solutions and would therefor signify a big step towards privacy-preservation in IoT data markets.

While the first research direction we mention is concerned with dealing with issues inherent to the IoT nature, the second direction is the opposite: it uses (other) inherent components of the IoT for solving current challenges. For example, [3] proposes making use of mutual validation in the IoT context by benefiting of the fact that in some scenarios (e.g. if the IoT devices are physically very close to each other), there should be a high correlation among the sensed data, which could be used for data correctness verification purposes. [29] proposes using an incentive mechanism design for fog computing in order to reduce analysis efforts of servers and allowing partial data processing at idle mobile devices owned by other users. In this way, owners of IoT devices would not only be paid for sensing data, but also for available computing power. Another example would be using the heterogeneity of IoT devices in order for buyers to decide which device is better for a certain type of data. Future research should further explore how IoT devices and their many characteristic features could open new use-cases and mathematical scenarios that are not available in other frameworks.

The third direction is concerned with defining standards. As it is the problem with any young technology and use case, there is a great deal of proposals as to how to standardize some of the processes carried-out in data markets. This goes from highly abstract concepts, such as the definition of a language or syntax to describe privacy related problems, to hands on implementation standards, such as universal APIs. One important standard that must be set is the syntactic and semantic definition of privacy terms in order for them to be machine readable. One of the important recent trends in this area is the increasing adoption of ontologies to privacy knowledge [54]. Ontologies define a common vocabulary for researchers who

need to share information in a domain, including machine-interpretable definitions of basic concepts in a domain and the relations among them. The standardization of ontologies would also have a big impact on concrete technologies used later; for example, it would play a big role in the manner SQL and Non-SQL databases would be implemented. Another standard that must be dealt with is the definition of APIs that can deal with the heterogeneous set of IoT devices available nowadays.

The fourth direction is research on the legal measures that are adequate for IoT data markets regarding privacy preservation. Even though it is not the focus of this thesis, the development of legal requirements will greatly determine many of the aspects in future IoT data markets. We discuss some of the challenges in sections 6.1 and 6.2.

Lastly, another future research direction that will also have a great impact on future privacy-preserving data markets is the development of effective pricing mechanisms. They will define aspects such as user acceptance and market integration, as well as possibly reinforce some proposed technologies more than others.

7. Discussion

In this section, we discuss the key findings and the limitations of this thesis.

7.1. Key findings

There is an increasing attention of scientists in the research of privacy-preservation of data markets for IoT devices
Even if our meta-data mapping in section 5 only comprised a total of 50 studies and is therefor only to be interpreted as a reference for our SLR, it showed a clear tendency of the increasing attention that scientists have had in the recent years for the implementation of privacy-preserving mechanisms in IoT data markets. 66% of the studies we found were published either in 2018, 2019, or the first four months of 2020, which is a clear sign that the topics posed in the RQs are contemporary in research facilities. The combination of the research approach "Design and creation" and the contribution type "framework or models" was found in an overwhelmingly high number of studies accounting for 62% of the total, while the research approach "Case study" and the research contribution type "Lessons learned" only accounted for 4% and 2%, respectively. This is a general sign of the maturity for the research topics.

USA and China are the countries with most contributions within the selected studies
We reiterate that our meta-data mapping in section 5 only comprised a total of 50 studies and is therefor only to be interpreted as a reference for our SLR. Even so, 79% of the found studies were written at least in collaboration with an institution located either in the USA or in China, which is a clear indication of the geographical location of research efforts on the discussed topics.

The privacy requirement for IoT data markets is a requirement that creates several conflicts within data market design
In section 6.1, we discussed the current challenges in the implementation of privacy-preserving data markets. We found that several of the requirements needed in IoT data markets are put in conflict with the privacy requirement, i.e. more privacy reduces the capabilities of other requirements. Concretely, these other requirements are fairness, truthfulness, accountability and economic viability. We note that these conflicts arise with current technological solutions and are not unsolvable for future work. In fact, in section 6.3 we mention some directions that could be taken but that are not maturely researched in order to minimize the impact of these conflicts. Moreover, the privacy requirement specifically for the IoT environment

creates a conflict with the efficiency requirement of data markets.

There are five perspectives that must be taken into account when designing privacy-preserving data markets
In section 6.2, we identify four design perspectives that are considered the pillars of the topics currently being researched. The first perspective is platform architecture, which encompasses the way various components and their interactions are structured in order to compose a system, as well as the system layers that encapsulate the services these components provide. The second perspective is the mathematical perspective, which carves the game theoretical and the market design models that are used in the data markets. Furthermore, the mathematical perspective also describes the algorithms that are used to solve the pricing subject in data markets. The third perspective is the security perspective, which encompasses all measures that provide confidentiality and protect the system from potential attacks. The fourth perspective is the legal perspective, which sets the parameters under which the data market must function according to the law and therefor adds further requirements. These four pillars are wrapped by a fifth perspective, which is the use case context perspective. This last perspective encompasses the specific IoT related use case and is the instance that indicates which technologies are most relevant for the usage context.

Encryption techniques to provide confidentiality must be adjusted to the available capabilities of the devices and to the required privacy level
In section 6.3.1 we classify several data security techniques such as data encryption, data re-encryption, data usage control and secret sharing. Each of these techniques has its upsides and downsides. One important downside with complex encryption methods such as AES-256 is that the computation and services provided by IoT devices are jeopardized given their innate capacities. Another problem with super-secure schemes is that many IoT relationships are short-termed, which means that keys and data access information must be updated dynamically, increasing the costs for managing entities and computation efforts. For this reason, it is crucial that the required security levels are correctly calculated in order to only provide the required level and not sacrificing computation capabilities and unnecessary costs. One significant trend that is also used in many other applications nowadays is that raw data is encrypted symmetrically, and the symmetric key is then encrypted asymmetrically in order to increase computation efficiency.

Within data processing stages, homomorphic encryption, k-anonymity and differential privacy are the most frequently used privacy-preserving techniques
In section 6.3.2, we discuss various techniques for processes that require computational functionality that takes place outside of the physical scope of data owners. The first widely proposed technique is homomorphic encryption, a technique we classify under “secure offloading”. The other technique in this category, order-preserving encryption, is not used often since it harshly limits the types of algorithms that can be applied in the processing stage. The main problem of homomorphic encryption is that it requires very high computation

complexity and overhead, posing an important challenge in the IoT context for devices and data consumers. That is why partially homomorphic encryption schemes are mentioned more often in our studies than fully homomorphic schemes. The second widely used technique is k-anonymity, a syntactic technique that can be achieved through mechanisms such as generalization. K-anonymity has the issue that it does not work well if sensitive data attributes lack diversity, which is why other similar techniques, such as l-diversity or t-closeness, are mentioned in other studies. The third technique that is mentioned several times is differential privacy, which is a semantic technique that has a strong mathematical foundation. The main issues with differential privacy are that they are not ideal for applications with the highest quality models and they are not well-suited for interactive settings. These three main techniques were used much more often than other techniques such as MPC, onion routing or pseudonym creation.

Blockchain and digital signatures are the two most frequently mentioned technologies for supporting identity verification and data correctness
Digital signatures and blockchain technologies are discussed in section 6.3.3. Firstly, we find that digital signatures are, to date, the best mechanism for identity verification. The main issues with digital signatures are that they surrender a certain amount of privacy in order to address accountability and correctness requirements, and that they incur significant communication overhead while being prone to satisfy the stringent time requirements of IoT data markets. For the first issue, approaches such as group signatures and blind signatures are proposed by some of the selected studies, while there is no proposed solution for the second issue. Still, digital signatures are used more often than other mechanisms such as digital fingerprints or truth discovery algorithms. Secondly, we discover that blockchain, as part of the mentioned DLTs, provides a promising solution to requirements such as data correctness and accountability given its decentralized nature, which also removes the central issue of the need for a trusted third party and can reduce maintenance costs. Blockchain is, by a large margin, the most mentioned DLT. Potential challenges of this technology are the high computational costs as well as privacy disclosure of smart-contract specifications in many cases. To mitigate the latter, one often mentioned viable solution is using TEEs in trusted nodes. We find that all the solutions that mention using TEEs implement them in a blockchain environment.

An efficient way to enable privacy preferences is through the definition of policies
In section 6.3.4 we discuss how several studies propose schemes for setting up privacy policies when trading private data. This is an efficient approach not only because it facilitates the definition of preferences by the user, but also because it aims at complying with current and future legal requirements: as lawmaking is limited in the technical specifications it can introduce in order to diminish innovation, it can define semantical decisions that must be made by the data owner in order for an application to be privacy-preserving. The challenges of this approach lie in the general complexity of re-modeling data workflows depending on the situational context. Other challenges are the definition of standards for machine-readable

privacy policies as well as a common language that can be understood by the average user.

There are five fundamental research categories scientists mention for their future work
At the end of section 6.3, we identify five main future directions that researchers in our selected studies mention. The first direction is concerned with physical abilities of IoT devices as well as the research of efficient algorithms in order to tackle many of the challenges created by current solutions. The second direction is to make use of IoT components in order to solve other challenges. The third direction is concerned with defining standards, from highly abstract concepts such as languages and privacy definitions, to implementation standards such as universal APIs. The fourth direction is concerned with researching legal measures that find a balance between privacy-preservation and other requirements in order to set up an optimal framework for innovation. The fifth and last direction is concerned with developing effective pricing mechanisms.

7.2. Limitations

This section describes the threats to validity and follows the workflow of our methodology, aspiring to identify limitations that are possible in each step.
Concerning the identification of need for a review, we stated in section 3 that we did not find any study that answers our RQs in a thorough and rigorous manner as we did. This claim was based on a manual search conducted on February 2020 in the OPAC system of the TUM as well as other EDS such as IEEE and SpringerLink. Related studies that may be found in other data sources were not contemplated at this point. Within our search strategy, we also limited our manual search and our automated search to seven EDS and the TUM OPAC system, which is why studies stored in other sources were not identified. In order to create our search string, we used a rigorous process based on the corpus of our base literature (see section 4.2.2). Even so, our search string may have been incomplete and may therefor have missed some relevant results.
Regarding our selection of studies, the filters based on study title, study abstract and study content were made by two researchers in order to minimize bias. Even so, a margin of human error and bias was not eliminated and may have an influence on the validity of our studies. As for the data synthesis, we only had a total of 50 studies in the final selection, which is why the meta-data mapping depicted in section 5 can only be used as a reference for these 50 studies and not for the entire research field. Furthermore, the topics concerning privacy-preservation in IoT data markets are highly contemporary and new discoveries are made continuously. Since the automated search was conducted at the end of April 2020, relevant innovations that have occurred or surfaced in the meantime until the publication of this thesis were not included. Finally, in order to limit the scope of this thesis, we did not go into detail of some of the research topics mentioned in section 6.2, such as the used game theoretical models, proposed pricing approaches or possible attacks by external parties.

8. Conclusion and future work

This section summarizes the findings of the thesis and presents an outlook for future work.

8.1. Summary

In this thesis, we assessed 50 relevant studies that discussed the challenges, current topics, techniques and research directions in the implementation of privacy-preserving data markets for IoT devices. The methodology for this assessment was a systematic literature review that was designed to minimize the bias effects in the findings.
Firstly, the main research gaps and problems for privacy-preservation in data markets for IoT devices were identified and thoroughly described. Some of these challenges are efficiently tackled by the solutions in the selected studies, while others persevere or have not had an adequate solution and are therefor part of the topics to be addressed in future research. Secondly, current research topics for the implementation of privacy-preserving IoT data marketplaces were described and structured into the four pillar perspectives "Platform architecture perspective", "Mathematical perspective", "Security perspective" and "Legal perspective", all of them wrapped in a fifth perspective that encompasses the specific IoT use case. Thirdly, the individual privacy-preservation techniques used and/or mentioned in the selected papers were discussed and classified into the four building blocks "Data security", "Data processing", "Identity verification and data correctness" and "Platform capabilities", a structure that was heavily influenced by [33]. The implementation of each technique was given with a series of advantages and disadvantages that should be considered for each individual use case and technology combination. Fourthly, five broad research directions were given as a research outlook for privacy-preservation techniques.
Moreover, a mapping of the studies meta-data was presented in section 5 in order to introduce the selected studies in an aggregated and graphical manner. This allows the reader to know where, when, how and by whom the selected studies were conducted.

8.2. Future work

The following recommendations for future work are based on the defined RQs in this thesis (see section 4). For the identification of current gaps and issues within privacy-preserving data markets, it would be interesting to further research which issues are being solved and which ones are still without promising solutions. Our analysis has given some indications of the technologies and problems that are being tackled, but a larger, detailed analysis of

the technological and logical constraints we face could further define future paths. In the analysis of current topics, we provide a high-level overview of perspective pillars that play a role in market design. It would be helpful to know how these perspectives can be unified in order for them to act as a single thread and not being researched independently. Regarding the privacy-preservation techniques, we mention that several of them have constraints, such as computational complexity or low levels of privacy. It would be helpful to measure these constraints with adequate metrics and map them to the mentioned solutions. Lastly, for the mentioned future directions in the selected studies, a review of the development speed for each of the identified directions could help providing new research opportunities.

A. Appendix

A.1. Selected studies

The following table presents the selected studies. As described in section 4, we started with 8 studies as base literature and ended up with a total of 50 studies in the final selection. The 50 studies were comprised as follows: 31 studies were found only in the automated search, 6 studies were part of the base literature as well as of the automated search, 2 studies were part of the base literature but not of the automated search results and 11 studies were added as additional references after the automated search. The acronyms used in the table describe at ahich phase each paper was found:

- BL means that this study was part of the *base literature*
- AS means that the study was found within the *automated search*
- AL means that the study was found as a reference in one of the other studies and was included as *additional literature* after the automated search

Table A.1.: Selected studies for the SLR

ID	Citation	Title	Inclusion stage
P1	[15]	A decentralized and secure blockchain platform for open fair data trading	AS
P2	[87]	A Demonstration of Sterling: A Privacy-Preserving Data Marketplace	AS
P3	[55]	A multi-layered blockchain framework for smart mobility data-markets	BL + AS
P4	[1]	A Survey on Big Data Market: Pricing, Trading and Protection	BL + AS
P5	[98]	A Universal Toolkit for Cryptographically Secure Privacy-Preserving Data Mining	AS
P6	[18]	A vision for global privacy bridges: Technical and legal measures for international data markets	AS
P7	[57]	Achieving Data Truthfulness and Privacy Preservation in Data Markets	BL + AS

Continued on next page

Table A.1 – *Continued from previous page*

ID	Citation	Title	Inclusion stage
P8	[20]	All of me? Users' preferences for privacy-preserving data markets and the importance of anonymity	BL
P9	[89]	An efficient two-layer mechanism for privacy-preserving truth discovery	AL
P10	[99]	BeeKeeper 2.0: Confidential blockchain-enabled IoT system with fully homomorphic computation	AL
P11	[51]	Blockchain for IoT security and privacy: The case study of a smart home	AL
P12	[12]	Blockchain for Large-Scale Internet of Things Data Storage and Protection	AS
P13	[123]	Blockchain-Enabled Peer-to-Peer Data Trading Mechanism	AS
P14	[30]	Challenges and Opportunities in IoT Data Markets	AS
P15	[78]	Contract Design for Purchasing Private Data Using a Biased Differentially Private Algorithm	AS
P16	[75]	Crowd-Empowered Privacy-Preserving Data Aggregation for Mobile Crowdsensing	AS
P17	[22]	Data marketplace for Internet of Things	AS
P18	[16]	Data Trading with Differential Privacy in Data Market	AS
P19	[33]	Dataflow Challenges in an Internet of Production: A Security & Privacy Perspective	AS
P20	[79]	Decentralized Data Marketplace to Enable Trusted Machine Economy	AS
P21	[23]	Defining, Enforcing and Checking Privacy Policies in Data-Intensive Applications	AS
P22	[27]	Differentially Private Auctions for Private Data Crowdsourcing	AS
P23	[3]	Distributed Data Privacy Preservation in IoT Applications	BL + AS
P24	[103]	DPDT: A Differentially Private Crowd-Sensed Data Trading Mechanism	AS
P25	[53]	Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts	AL
P26	[76]	Enabling privacy-preserving auctions in big data	AL

Continued on next page

Table A.1 – *Continued from previous page*

ID	Citation	Title	Inclusion stage
P27	[17]	End-to-End Privacy for Open Big Data Markets	BL + AS
P28	[60]	Ensuring personal data anonymity in data marketplaces through sensing-as-a-service and distributed ledger technologies	AL
P29	[50]	Framework for Security and Privacy in Automotive Telematics	AS
P30	[62]	Hermes: An Open and Transparent Marketplace for IoT Sensor Data over Distributed Ledgers	AS
P31	[8]	Incorporating social interaction into three-party game towards privacy protection in IoT	AS
P32	[94]	Machine learning based privacy-preserving fair data trading in big data market	AS
P33	[85]	On blockchain-based anonymized dataset distribution platform	AS
P34	[74]	On privacy-Preserving Clound Action	AL
P35	[31]	Participant Privacy in Mobile Crowd Sensing Task Management	AL
P36	[68]	Personalized privacy in open data sharing scenarios	AS
P37	[80]	Privacy Bargaining with Fairness: Privacy-Price Negotiation System for Applying Differential Privacy in Data Market Environments	AS
P38	[13]	Privacy in the Internet of Things: threats and challenges	AS
P39	[34]	Privacy-by-design framework for assessing internet of things applications and platforms	AS
P40	[54]	Privacy-Knowledge Modeling for the Internet of Things: A Look Back	BL + AS
P41	[77]	Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption	BL + AS
P42	[92]	PRIVATA: Differentially Private Data Market Framework Using Negotiation-Based Pricing Mechanism	AS
P43	[52]	ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability	AL
P44	[110]	Publishing microdata with a robust privacy guarantee	BL

Continued on next page

Table A.1 – *Continued from previous page*

ID	Citation	Title	Inclusion stage
P45	[61]	SDTE: A Secure Blockchain-Based Data Trading Ecosystem	AS
P46	[6]	Secure Fair and Efficient Data Trading Without Third Party Using Blockchain	AS
P47	[29]	The Accuracy-Privacy Trade-off of Mobile Crowdsensing	AL
P48	[32]	Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems	AL
P49	[21]	Towards a Trusted Marketplace for Wearable Data	AS
P50	[114]	Trading private range counting over big IoT data	AS

A.2. Title reader for extracted *.bib*-files

We note that the following program is only a rudimentary implementation for extracting the title lines according to the structure that we encountered in the *.bib* files. The program does not guarantee functioning for other more complex file structures.

```
import java.io.*;
public class BibTitleReader {
public static void main(String[] args) {
try {
    BufferedReader br = new BufferedReader (new FileReader("BibFileName.bib"));
    PrintWriter pw = new PrintWriter(new FileWriter("Titles.txt"));
    pw.println("The titles of the bibfile are: ");
    while (br.ready()) {
        String line = br.readLine();
        if (line.startsWith("Title") || line.startsWith("title")) {
String line2 = line.substring(7).substring(0, line.substring(9).length()-0);
pw.println(line2);
pw.flush();
    }
    }
    br.close();
    pw.close();
} catch (IOException e) {
e.printStackTrace();
} } }
```

A.3. Sketch Engine details

As described in section 4.2.2, we used an online tool called *Sketch Engine* to create a corpus of our base literature and analyze its recurring terms in order to create our search string. Within this tool, we used the following functionalities:

- The *Keywords* functionality is used to extract single word and multi-word units which are typical of the corpus. It defines what makes the existing corpus different from other corpora. The multi-word items were displayed together with links to the sentences where they were used. This was the functionality that was most useful for our means. Figure A.1 depicts the user interface and shows how we already see terms such as *Differential privacy* that will be relevant in our SLR
- The *Wordlist* function generates frequency lists for all words in the corpus (brute force approach), which means that within the first pages we found many words like "the" or "in". We had to go further into the anaylsis in order to find useful results. Users can specifically look for word types (verbs, nouns, etc.).
- The *Thesaurus* functionality automatically generated list of synonyms or words belonging to the same category (semantic field). The list was produced based on the context in which the words appear in our corporus.
- The *Word Sketch* functionality processes the word's collocates and other words in its surroundings. It can be used as a one-page summary of the word's grammatical and collocational behaviour. The results are organized into categories, called grammatical relations, such as words that serve as an object of the verb, words that serve as a subject of the verb, words that modify the word etc.

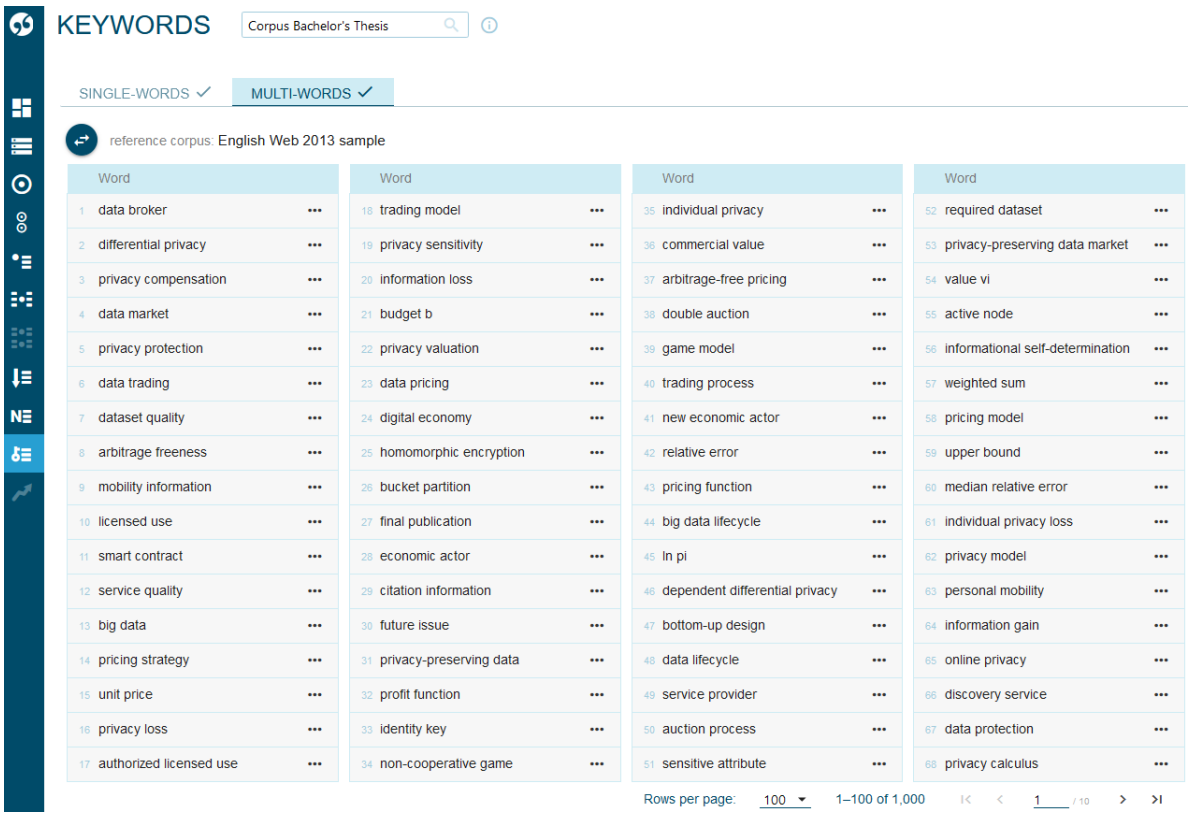

Figure A.1.: Example of the Keywords functionality using WordSketch

List of Figures

List of Tables

Bibliography

[1] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao. "A Survey on Big Data Market: Pricing, Trading and Protection". In: *IEEE Access* 6.May (2018), pp. 15132–15154. ISSN: 21693536. DOI: 10.1109/ACCESS.2018.2806881.

[2] J. Milenkovic. *30 Eye-Opening Big Data Statistics for 2020: Patterns Are Everywhere*. 2019. URL: https://kommandotech.com/statistics/big-data-statistics/ (visited on 10/06/2020).

[3] J. Du, C. Jiang, E. Gelenbe, L. Xu, J. Li, and Y. Ren. "Distributed Data Privacy Preservation in IoT Applications". In: *IEEE Wireless Communications* 25.December (2018), pp. 68–76. DOI: 10.1109/MWC.2017.1800094.

[4] *Factual Website*. URL: https://www.factual.com/.

[5] *Snowflake Website*. URL: https://www.snowflake.com/data-marketplace/.

[6] Z. Guan, X. Shao, and Z. Wan. "Secure, Fair and Efficient Data Trading without Third Party Using Blockchain". In: *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (2018), pp. 1349–1354. DOI: 10.1109/Cybermatics.

[7] B. A. Kitchenham and D. Budgen. *Evidence-based software engineering and systematic reviews*. Chapman and Hall/CRC, 2016. ISBN: 978-1-4822-2865-6.

[8] K. Li, L. Tian, W. Li, G. Luo, and Z. Cai. "Incorporating social interaction into three-party game towards privacy protection in IoT". In: *Computer Networks* 150 (2019), pp. 90–101. ISSN: 13891286. DOI: 10.1016/j.comnet.2018.11.036. URL: https://doi.org/10.1016/j.comnet.2018.11.036.

[9] T. Economist. *Data, Data Everywhere*. 2010. URL: https://www.economist.com/special-report/2010/02/27/data-data-everywhere (visited on 10/14/2020).

[10] J. Elder. *How Kevin Ashton named The Internet of Things*. 2019. URL: https://blog.avast.com/kevin-ashton-named-the-internet-of-things (visited on 10/08/2020).

[11] M. Rouse. *Internet of Things*. 2020. URL: https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT.

[12] R. Li, T. Song, B. Mei, H. Li, X. Cheng, and L. Sun. "Blockchain for Large-Scale Internet of Things Data Storage and Protection". In: *IEEE Transactions on Services Computing* 12.5 (2019), pp. 762–771. ISSN: 19391374. DOI: 10.1109/TSC.2018.2853167.

[13] J. H. Ziegeldorf, O. G. Morchon, and K. Wehrle. "Privacy in the internet of things: Threats and challenges". In: *Security and Communication Networks* 7.12 (2014), pp. 2728–2742. ISSN: 19390122. DOI: 10.1002/sec.795.

[14] ITU. *Internet of Things (IoT)*. URL: https://www.itu.int/en/ITU-T/ssc/resources/Pages/topic-001.aspx.

[15] Y. N. Li, X. Feng, J. Xie, H. Feng, Z. Guan, and Q. Wu. "A decentralized and secure blockchain platform for open fair data trading". In: *Concurrency Computation* 32.7 (2019), pp. 1–11. ISSN: 15320634. DOI: 10.1002/cpe.5578.

[16] X. Zheng. "Data trading with differential privacy in data market". In: *ACM International Conference Proceeding Series* 8 (2020), pp. 112–115. DOI: 10.1145/3379247.3379271.

[17] C. Perera, R. Ranjan, and L. Wang. "End-to-end privacy for open big data markets". In: *IEEE Cloud Computing* 2.4 (2015), pp. 44–53. ISSN: 23256095. DOI: 10.1109/MCC.2015.78.

[18] S. Spiekermann and A. Novotny. "A vision for global privacy bridges: Technical and legal measures for international data markets". In: *Computer Law and Security Review* 31.2 (2015), pp. 181–200. ISSN: 02673649. DOI: 10.1016/j.clsr.2015.01.009. URL: http://dx.doi.org/10.1016/j.clsr.2015.01.009.

[19] . European Parliament and Council of the European Union. *General Data Protection Regulation*. 2016. URL: https://gdpr-info.eu/ (visited on 10/08/2020).

[20] E. M. Schomakers, C. Lidynia, and M. Ziefle. "All of me? Users' preferences for privacy-preserving data markets and the importance of anonymity". In: *Electronic Markets* (2020). ISSN: 14228890. DOI: 10.1007/s12525-020-00404-9.

[21] A. Colman, M. J. M. Chowdhury, and M. Baruwal Chhetri. "Towards a trusted marketplace for wearable data". In: *Proceedings - 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC 2019* Cic (2019), pp. 314–321. DOI: 10.1109/CIC48465.2019.00044.

[22] K. Mišura and M. Žagar. "Data marketplace for Internet of Things". In: *Proceedings of 2016 International Conference on Smart Systems and Technologies, SST 2016* (2016), pp. 255–260. DOI: 10.1109/SST.2016.7765669.

[23] M. Guerriero, D. A. Tamburri, and E. Di Nitto. "Defining, enforcing and checking privacy policies in data-intensive applications". In: *Proceedings - International Conference on Software Engineering* (2018), pp. 172–182. ISSN: 02705257. DOI: 10.1145/3194133.3194140.

[24] A. F. Westin. *Pivacy and Freedom*. The Bodley Head Ltd, 1967.

[25] D. Solove. "A taxonomy of privacy". In: *Law Review* 477 (2006), pp. 477–560.

[26] R. Clark. *Introduction to Dataveillance and Information Privacy, and Definitions of Terms*. 1997. URL: http://www.rogerclarke.com/DV/Intro.html (visited on 10/14/2020).

[27] M. Shi, Y. Qiao, and X. Wang. "Differentially private auctions for private data crowdsourcing". In: *Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019* (2019), pp. 1–8. DOI: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00013.

[28] G. Danezis, J. Domingo-Ferrer, M. Hansen, J.-H. Hoepman, D. Le Métayer, R. Tirtea, and S. Schiffner. *Privacy and Data Protection by Design – from policy to engineering*. December. 2014. ISBN: 9789292041083. DOI: 10.2824/38623. URL: https://arxiv.org/ftp/arxiv/papers/1501/1501.03726.pdf.

[29] M. A. Alsheikh, Y. Jiao, D. Niyato, P. Wang, D. Leong, and Z. Han. "The Accuracy-Privacy Trade-off of Mobile Crowdsensing". In: *IEEE Communications Magazine* 55.6 (2017), pp. 132–139. ISSN: 01636804. DOI: 10.1109/MCOM.2017.1600737. arXiv: 1702.04565.

[30] Z. Zheng, W. Mao, F. Wu, and G. Chen. "Challenges and opportunities in IoT data markets". In: *SocialSense 2019 - Proceedings of the 2019 4th International Workshop on Social Sensing* (2019), pp. 1–2. DOI: 10.1145/3313294.3313378.

[31] L. Pournajaf, D. A. Garcia-Ulloa, L. Xiong, and V. Sunderam. "Participant Privacy in Mobile Crowd Sensing Task Management". In: *ACM SIGMOD Record* 44.4 (2016), pp. 23–34. ISSN: 0163-5808. DOI: 10.1145/2935694.2935700.

[32] S. Sharma, K. Chen, and A. Sheth. "Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems". In: *IEEE Internet Computing* 22.2 (2018), pp. 42–51. ISSN: 10897801. DOI: 10.1109/MIC.2018.112102519.

[33] J. Pennekamp, M. Henze, S. Schmidt, P. Niemietz, M. Fey, D. Trauth, T. Bergs, C. Brecher, and K. Wehrle. "Dataflow Challenges in an Internet of Production". In: *ACMWorkshop on Cyber-Physical Systems Security & Privacy (CPS-SPC'19), November 11, 2019, London, United Kingdom. ACM*. 2019, pp. 27–38. ISBN: 9781450368315. DOI: 10.1145/3338499.3357357.

[34] C. Perera, C. McCormick, A. K. Bandara, B. A. Price, and B. Nuseibeh. "Privacy-by-design framework for assessing internet of things applications and platforms". In: *ACM International Conference Proceeding Series* 07-09-Nove (2016), pp. 83–92. DOI: 10.1145/2991561.2991566.

[35] B. Kitchenham. "Procedures for Performing Systematic Reviews". In: *Joint Technical Report* (2004). ISSN: 09754466. DOI: 10.5144/0256-4947.2017.79.

[36] B. Kitchenham. "Guidelines for performing Systematic Literature Reviews in Software Engineering". In: *EBSE Technical Report* (2007). ISSN: 00010782. DOI: 10.1145/1134285.1134500. arXiv: 1304.1186.

[37] D. Budgen, M. Turner, and B. Kitchenham. "Using Mapping Studies in Software Engineering". In: *Proceedings of PPIG 2008* (2008), pp. 195–204. ISSN: 18651348. DOI: 10.1007/978-3-642-02152-7_36.

[38] T. Dybå, T. Dingsøyr, and G. Hanssen. "Applying Systematic Reviews to Diverse Study Types: An Experience Report". In: *Proceedings - 1st International Symposium on Empirical Software Engineering and Measurement, ESEM 2007* 7465 (2007), pp. 126–135. DOI: 10.1109/ESEM.2007.59.

[39] P. Philipp. *Investigating the Current State of Research in Large-Scale Agile Software Development: A Systematic Mapping Study*. 2019.

[40] B. A. Kitchenham, D. Budgen, and O. P. Brereton. "The value of mapping studies – A participant-observer case study". In: (2010). DOI: 10.14236/ewic/ease2010.4.

[41] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman. "Systematic literature reviews in software engineering-A tertiary study". In: *Information and Software Technology* 52.8 (2010), pp. 792–805. ISSN: 09505849. DOI: 10.1016/j.infsof.2010.03.006. URL: http://dx.doi.org/10.1016/j.infsof.2010.03.006.

[42] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson. "Systematic mapping studies in software engineering". In: *12th International Conference on Evaluation and Assessment in Software Engineering, EASE 2008* June (2008). DOI: 10.14236/ewic/ease2008.8.

[43] K. Henz. *Vorwissenschaftliches Arbeiten: ein Praxisbuch für die Schule*. E. Dorner, 2010, p. 53.

[44] H. Zhang, M. A. Babar, and P. Tell. "Identifying relevant studies in software engineering". In: *Information and Software Technology* 53.6 (2011), pp. 625–637. ISSN: 09505849. DOI: 10.1016/j.infsof.2010.12.010. URL: http://dx.doi.org/10.1016/j.infsof.2010.12.010.

[45] L. Chen, M. A. Babar, and H. Zhang. "Towards an Evidence-Based Understanding of Electronic Data Sources". In: January 2015 (2010). DOI: 10.14236/ewic/ease2010.17.

[46] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, and V. Suchomel. "The Sketch Engine: ten years on". In: *Lexicography* (2014).

[47] RiverbedTechnology. *IoT Buzzword Bingo*. 2018. URL: https://www.riverbed.com/blogs/iot-buzzword-bingo.html.

[48] F. Firouzi, K. Chakrabarty, S. Nassif, and F. Device. *Intelligent Internet of Things*. 2020. ISBN: 9783030303662. DOI: 10.1007/978-3-030-30367-9.

[49] I. O. f. S. (ISO). "ISO 5807:1985 Information processing — Documentation symbols and conventions for data, program and system flowcharts, program network charts and system resources charts". In: 1985 (1985). URL: https://www.iso.org/standard/11955.html.

[50] S. Duri, M. Gruteser, X. Liu, P. Moskowitz, R. Perez, M. Singh, and J. M. Tang. "Framework for security and privacy in automotive telematics". In: *Proceedings of the ACM International Workshop on Mobile Commerce* (2002), pp. 25–32. DOI: 10.1145/570709.570711.

[51] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram. "Blockchain for IoT security and privacy: The case study of a smart home". In: *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017* (2017), pp. 618–623. DOI: 10.1109/PERCOMW.2017.7917634.

[52] X. Liang, S. Shetty, D. Tosh, C. Kamhoua, K. Kwiat, and L. Njilla. "ProvChain: A Blockchain-Based Data Provenance Architecture in Cloud Environment with Enhanced Privacy and Availability". In: *Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017* (2017), pp. 468–477. DOI: 10.1109/CCGRID.2017.8.

[53] R. Cheng, F. Zhang, J. Kos, W. He, N. Hynes, N. Johnson, A. Juels, A. Miller, and D. Song. "Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts". In: *Proceedings - 4th IEEE European Symposium on Security and Privacy, EURO S and P 2019* (2019), pp. 185–200. DOI: 10.1109/EuroSP.2019.00023.

[54] C. Perera, C. Liu, R. Ranjan, L. Wang, and A. Zomaya. "Privacy-knowledge Modeling for the Internet of things: A look back". In: *Computer* 49.12 (2016), pp. 60–68. ISSN: 0018-9162. DOI: 10.1109/MC.2016.366. URL: http://ieeexplore.ieee.org/document/7756262/.

[55] D. López and B. Farooq. "A multi-layered blockchain framework for smart mobility data-markets". In: *Transportation Research Part C: Emerging Technologies* 111.June 2019 (2020), pp. 588–615. ISSN: 0968090X. DOI: 10.1016/j.trc.2020.01.002. arXiv: 1906.06435. URL: https://doi.org/10.1016/j.trc.2020.01.002.

[56] A. Jøsang and T. Bhuiyan. "Optimal trust network analysis with subjective logic". In: *Proceedings - 2nd Int. Conf. Emerging Security Inf., Systems and Technologies, SECURWARE 2008, Includes DEPEND 2008: 1st Int. Workshop on Dependability and Security in Complex and Critical Inf. Sys.* (2008), pp. 179–184. DOI: 10.1109/SECURWARE.2008.64.

[57] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen. "Achieving Data Truthfulness and Privacy Preservation in Data Markets". In: *IEEE Transactions on Knowledge and Data Engineering* 31.1 (2019), pp. 105–119. ISSN: 15582191. DOI: 10.1109/TKDE.2018.2822727. arXiv: 1812.03280.

[58] F. T. Commission. *Google Will Pay $22.5 Million to Settle FTC Charges it Misrepresented Privacy Assurances to Users of Apple's Safari Internet Browser*. 2012. URL: https://www.ftc.gov/news-events/press-releases/2012/08/google-will-pay-225-million-settle-ftc-charges-it-misrepresented.

[59] *Council of European Union. Regulation (eu) 2016/679 - directive 95/46*. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[60] M. Zichichi, M. Contu, S. Ferretti, and V. Rodríguez-Doncel. "Ensuring personal data anonymity in data marketplaces through sensing-as-a-service and distributed ledger technologies". In: *CEUR Workshop Proceedings* 2580 (2020). ISSN: 16130073.

[61] W. Dai, C. Dai, K. K. R. Choo, C. Cui, D. Zou, and H. Jin. "SDTE: A Secure Blockchain-Based Data Trading Ecosystem". In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 725–737. ISSN: 15566021. DOI: 10.1109/TIFS.2019.2928256.

[62] P. Tzianos, G. Pipelidis, and N. Tsiamitros. "Hermes: An open and transparent marketplace for iot sensor data over distributed ledgers". In: *ICBC 2019 - IEEE International Conference on Blockchain and Cryptocurrency* (2019), pp. 167–170. DOI: 10.1109/BLOC.2019.8751331.

[63] C. Perera, R. Ranjan, L. Wang, S. U. Khan, and A. Y. Zomaya. "Big data privacy in the internet of things era". In: *IT Professional* 17.3 (2015), pp. 32–39. ISSN: 15209202. DOI: 10.1109/MITP.2015.34.

[64] C. Perera, C. H. Liu, and S. Jayawardena. "The Emerging Internet of Things Marketplace from an Industrial Perspective: A Survey". In: *IEEE Transactions on Emerging Topics in Computing* 3.4 (2015), pp. 585–598. ISSN: 21686750. DOI: 10.1109/TETC.2015.2390034. arXiv: 1502.00134.

[65] Mozilla Foundation. *Confidentiality, Integrity, and Availability*. 2019. URL: https://developer.mozilla.org/en-US/docs/Archive/Security/Confidentiality,%7B%5C_%7DIntegrity,%7B%5C_%7Dand%7B%5C_%7DAvailability (visited on 10/01/2020).

[66] F.-X. Standaert. "Introduction to Side-Channel Attacks". In: *Secure Integrated Circuits and Systems* September (2010), pp. 79–104. DOI: 10.1007/978-0-387-71829-3. URL: http://link.springer.com/10.1007/978-0-387-71829-3.

[67] M. F. Kaashoek and D. R. Karger. "Koorde: A simple degree-optimal distributed hash table". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2735 (2003), pp. 98–107. ISSN: 16113349. DOI: 10.1007/978-3-540-45172-3_9.

[68] D. Sánchez and A. Viejo. "Personalized privacy in open data sharing scenarios". In: *Online Information Review* 41.3 (2017), pp. 298–310. ISSN: 14684527. DOI: 10.1108/OIR-01-2016-0011.

[69] M. Rodriguez-Garcia, M. Batet, and D. Sánchez. "Semantic noise: Privacy-protection of nominal microdata through uncorrelated noise addition". In: *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* 2016-Janua (2016), pp. 1106–1113. ISSN: 10823409. DOI: 10.1109/ICTAI.2015.157.

[70] R. Chow, P. Golle, and J. Staddon. "Detecting privacy leaks using corpus-based association rules". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), pp. 893–901. DOI: 10.1145/1401890.1401997.

[71] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. "Fog computing and its role in the internet of things". In: *MCC'12 - Proceedings of the 1st ACM Mobile Cloud Computing Workshop* (2012), pp. 13–15. DOI: 10.1145/2342509.2342513.

[72] I. Stojmenovic and S. Wen. "The Fog computing paradigm: Scenarios and security issues". In: *2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014* 2 (2014), pp. 1–8. ISSN: 2300-5963. DOI: 10.15439/2014F503.

[73] A. Roth E. *The art of designing markets*. 2007. URL: https://hbr.org/2007/10/the-art-of-designing-markets (visited on 10/01/2020).

[74] Z. Chen, L. Chen, L. Huang, and H. Zhong. "On Privacy-Preserving Cloud Auction". In: *Proceedings of the IEEE Symposium on Reliable Distributed Systems* (2016), pp. 279–288. ISSN: 10609857. DOI: 10.1109/SRDS.2016.045.

[75] L. Yang, M. Zhang, S. He, M. Li, and J. Zhang. "Crowd-empowered privacy-preserving data aggregation for mobile crowdsensing". In: *Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (2018), pp. 151–160. DOI: 10.1145/3209582.3209598.

[76] T. Jung and X. Y. Li. "Enabling privacy-preserving auctions in big data". In: *Proceedings - IEEE INFOCOM* 2015-Augus.BigSecurity (2015), pp. 173–178. ISSN: 0743166X. DOI: 10.1109/INFCOMW.2015.7179380. arXiv: 1308.6202.

[77] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu. "Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption". In: *IEEE Transactions on Network Science and Engineering* 7.2 (2020), pp. 776–791. ISSN: 23274697. DOI: 10.1109/TNSE.2018.2846736.

[78] M. M. Khalili, X. Zhang, and M. Liu. "Contract design for purchasing private data using a biased differentially private algorithm". In: *Proceedings of NetEcon 2019: 14th Workshop on the Economics of Networks, Systems and Computation - In conjunction with ACM EC 2019 and ACM SIGMETRICS 2019* (2019). DOI: 10.1145/3338506.3340273.

[79] Z. J. Wang, C. H. V. Lin, Y. H. Yuan, and C. C. J. Huang. "Decentralized Data Marketplace to Enable Trusted Machine Economy". In: *2019 IEEE Eurasia Conference on IOT, Communication and Engineering, ECICE 2019* (2019), pp. 246–250. DOI: 10.1109/ECICE47484.2019.8942729.

[80] K. Jung and S. Park. "Privacy Bargaining with Fairness: Privacy-Price Negotiation System for Applying Differential Privacy in Data Market Environments". In: *2019 IEEE International Conference on Big Data* (2019), pp. 1389–1394. DOI: 10.1109/BigData47090.2019.9006101.

[81] I. Giannakopoulos, P. Karras, D. Tsoumakos, K. Doka, and N. Koziris. "An equitable solution to the stable marriage problem". In: *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI* 2016-Janua.i (2016), pp. 989–996. ISSN: 10823409. DOI: 10.1109/ICTAI.2015.142.

[82] H. Oh, S. Park, G. M. Lee, J. K. Choi, and S. Noh. "Competitive Data Trading Model With Privacy Valuation for Multiple Stakeholders in IoT Data Markets". In: *IEEE Internet of Things Journal* 7.4 (2020), pp. 3623–3639. ISSN: 23274662. DOI: 10.1109/jiot.2020.2973662.

[83] L. Tian, J. Li, W. Li, B. Ramesh, and Z. Cai. "Optimal Contract-Based Mechanisms for Online Data Trading Markets". In: *IEEE Internet of Things Journal* 6.5 (2019), pp. 7800–7810. ISSN: 23274662. DOI: 10.1109/JIOT.2019.2902528.

[84] J. Parra-Arnau. "Optimized, direct sale of privacy in personal data marketplaces". In: *Information Sciences* 424 (2018), pp. 354–384. ISSN: 00200255. DOI: 10.1016/j.ins.2017.10.009. URL: https://doi.org/10.1016/j.ins.2017.10.009.

[85] S. Kiyomoto, M. S. Rahman, and A. Basu. "On Blockchain-Based Anonymized Dataset Distribution Platform". In: *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)* (2017), pp. 85–92.

[86] M. Rouse. "What is a Use Case?" In: (2020). URL: https://searchsoftwarequality.techtarget.com/definition/use-case.

[87] N. Hynes, D. Dao, D. Yan, R. Cheng, and D. Song. "A demonstration of sterling: A privacy-preserving data marketplace". In: *Proceedings of the VLDB Endowment* 11.12 (2018), pp. 2086–2089. ISSN: 21508097. DOI: 10.14778/3229863.3236266.

[88] SkyTree. *The future is here - Why is machine learning so big right now?* 2018. URL: https://www.skytree.net/machine-learning/why-do-machine-learning-big-data/.

[89] Y. Li, C. Miao, L. Su, J. Gao, Q. Li, B. Ding, Z. Qin, and K. Ren. "An efficient two-layer mechanism for privacy-preserving truth discovery". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018), pp. 1705–1714. DOI: 10.1145/3219819.3219998.

[90] X. Yang, T. Wang, X. Ren, and W. Yu. "Survey on Improving Data Utility in Differentially Private Sequential Data Publishing". In: *IEEE Transactions on Big Data* 7790.c (2017), pp. 1–1. ISSN: 2332-7790. DOI: 10.1109/tbdata.2017.2715334.

[91] X. Yang, X. Ren, J. Lin, and W. Yu. "On Binary Decomposition Based Privacy-Preserving Aggregation Schemes in Real-Time Monitoring Systems". In: *IEEE Transactions on Parallel and Distributed Systems* 27.10 (2016), pp. 2967–2983. ISSN: 10459219. DOI: 10.1109/TPDS.2016.2516983.

[92] S. Park, K. Park, J. Lee, and K. Jung. "PRIVATA: Differentially private Data market framework using Negotiation-based Pricing mechanism". In: *Proceedings of ACM CIKM conference (CIKM'19), November 3–7, 2019, Beijing, China.* (2019), pp. 156–157. DOI: 10.1007/978-3-663-10915-0_47.

[93] S. Yu, C. Wang, K. Ren, and W. Lou. "Achieving secure, scalable, and fine-grained data access control in cloud computing". In: *Proceedings - IEEE INFOCOM* (2010). ISSN: 0743166X. DOI: 10.1109/INFCOM.2010.5462174.

[94] Y. Zhao, Y. Yu, Y. Li, G. Han, and X. Du. "Machine learning based privacy-preserving fair data trading in big data market". In: *Information Sciences* 478 (2019), pp. 449–460. ISSN: 00200255. DOI: 10.1016/j.ins.2018.11.028. URL: https://doi.org/10.1016/j.ins.2018.11.028.

[95] M. Yung, S. Jarecki, H. Krawczyk, and A. Herzberg. "Proactive Secret Sharing Or : How to Cope With Perpetual Leakage". In: *Communication* (1995), pp. 1–22.

[96] T. P. Pedersen. "Non-interactive and information-theoretic secure verifiable secret sharing". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 576 LNCS (1992), pp. 129–140. ISSN: 16113349. DOI: 10.1007/3-540-46766-1_9.

[97] A. Shamir. "How to share a secret". In: *Publications of the ACM* (1979). DOI: 10.1007/978-3-642-15328-0_17.

[98] D. Bogdanov, R. Jagomägis, and S. Laur. "A Universal Toolkit for Cryptographically Secure Privacy-Preserving Data Mining". In: *LNCS 7299 - Intelligence and Security Informatics* 7299 (2012). URL: https://link.springer.com/content/pdf/10.1007%7B%5C%%7D2F978-3-642-30428-6.pdf.

[99] L. Zhou, L. Wang, T. Ai, and Y. Sun. "BeeKeeper 2.0: Confidential blockchain-enabled IoT system with fully homomorphic computation". In: *Sensors (Switzerland)* 18.11 (2018). ISSN: 14248220. DOI: 10.3390/s18113785.

[100] X. Chen. *Introduction to Secure Outsourcing Computation*. Morgan & Claypool publishers, 2016, p. 94.

[101] M. A. Will and R. K. Ko. *A guide to homomorphic encryption*. Elsevier Inc., 2015, p. 101. ISBN: 9780128017807. DOI: 10.1016/B978-0-12-801595-7.00005-7. URL: http://dx.doi.org/10.1016/B978-0-12-801595-7.00005-7.

[102] P. Paillier. "Public-key cryptosystems based on composite degree residuosity classes". In: *Eurocrypt* (1999). ISSN: 16113349. DOI: 10.1007/3-540-48910-X_9.

[103] G. Gao, M. Xiao, J. Wu, S. Zhang, L. Huang, and G. Xiao. "DPDT: A Differentially Private Crowd-Sensed Data Trading Mechanism". In: *IEEE Internet of Things Journal* 7.1 (2020), pp. 751–762. ISSN: 23274662. DOI: 10.1109/JIOT.2019.2944107.

[104] P. Chaudhary, R. Gupta, A. Singh, and P. Majumder. "Analysis and Comparison of Various Fully Homomorphic Encryption Techniques". In: *2019 International Conference on Computing, Power and Communication Technologies, GUCON 2019* (2019), pp. 58–62.

[105] M. Stadler. "Publicly verifiable secret sharing". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1070 (1996), pp. 190–199. ISSN: 16113349. DOI: 10.1007/3-540-68339-9_17.

[106] Z. A. Genç, V. Iovino, and A. Rial. ""The simplest protocol for oblivious transfer"". In: *Information Processing Letters* 161 (2020), pp. 1–12. ISSN: 00200190. DOI: 10.1016/j.ipl.2020.105975.

[107] S. Yakoubov. "A Gentle Introduction to Yao' s Garbled Circuits". In: (2017). URL: https://web.mit.edu/sonka89/www/papers/2017ygc.pdf.

[108] S. Goldwasser, S. Micali, and C. Rackoff. "The Knowledge Complexity Of Interactive Proof Systems". In: *Society for Industrial and Applied Mathematics* (1989).

[109] C. Bhardwaj. *What is Zero-Knowledge Proof & its Role in the Blockchain World?* 2020. URL: https://appinventiv.com/blog/zero-knowledge-proof-blockchain/ (visited on 10/02/2020).

[110] J. Cao and P. Karras. "Publishing microdata with a robust privacy guarantee". In: *Proceedings of the VLDB Endowment* 5.11 (2012), pp. 1388–1399. ISSN: 21508097. DOI: 10.14778/2350229.2350255. arXiv: 1208.0220.

[111] S. De Capitani Di Vimercati, S. Foresti, G. Livraga, and P. Samarati. "Data privacy: Definitions and techniques". In: *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems* 20.6 (2012), pp. 793–817. ISSN: 02184885. DOI: 10.1142/S0218488512400247.

[112] A. Meyerson and R. Williams. "On the complexity of optimal k-anonymity". In: *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* 23 (2004), pp. 223–228. DOI: 10.1145/1055558.1055591.

[113] G. Cormode. "Personal privacy vs population privacy: Learning to attack anonymization". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2011), pp. 1253–1261. DOI: 10.1145/2020408.2020598.

[114] Z. Cai and Z. He. "Trading private range counting over big IoT data". In: *Proceedings - International Conference on Distributed Computing Systems* 2019-July (2019), pp. 144–153. DOI: 10.1109/ICDCS.2019.00023.

[115] *Tor Website*. URL: https://www.torproject.org/ (visited on 10/03/2020).

[116] R. Henry, A. Herzberg, and A. Kate. "Blockchain access privacy: Challenges and directions". In: *IEEE Security and Privacy* 16.4 (2018), pp. 38–45. ISSN: 15584046. DOI: 10.1109/MSP.2018.3111245.

[117] Z. Lyasota. *A guide to digital signature algorithms*. 2018. URL: https://dzone.com/articles/digital-signature-1 (visited on 10/03/2020).

[118] R. Rivest, A. Shamir, and Y. Tauman. "How to Leak a secret". In: *Lecture Notes in Computer Science, vol 2248. Springer, Berlin, Heidelberg.* (2001). URL: https://media.readthedocs.org/pdf/btc-relay/latest/btc-relay.pdf%7B%5C%%7D0Ahttps://github.com/namecoin/wiki/blob/master/Merged-Mining.mediawiki%7B%5C%%7D0Ahttps://ieeexplore.ieee.org/document/6032224%7B%5C%%7D0Ahttps://cryptoslate.com/ethereum-network-congestion-doubles-gas-fees-as-game.

[119] M. Bellare, D. Micciancio, and B. Warinschi. "Foundations of Group Signatures". In: *Eurocrypt* 2656 (2003), pp. 1–27. ISSN: 03029743. DOI: 10.1007/3-540-39200-9.

[120] IOTA-Foundation. *About the Tangle*. 2020. URL: https://docs.iota.org/docs/getting-started/1.1/the-tangle/overview (visited on 10/14/2020).

[121] P. Handy. *Introducing Masked Authenticated Messaging*. 2017. URL: https://blog.iota.org/introducing-masked-authenticated-messaging-e55c1822d50e.

[122] IOTA-Foundation. *IOTA data marketplace*. URL: https://data.iota.org/%7B%5C#%7D/ (visited on 10/14/2020).

[123] J. Wei, M. Sabonuchi, and R. Roche. "Blockchain-enabled Peer-to-Peer Data Trading Mechanism". In: *2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (2018), pp. 1349–1354. DOI: 10.1109/Cybermatics.

[124] A. Ghosh, K. Ligett, A. Roth, and G. Schoenebeck. "Buying private data without verification". In: *EC 2014 - Proceedings of the 15th ACM Conference on Economics and Computation* (2014), pp. 931–948. DOI: 10.1145/2600057.2602902. arXiv: 1404.6003.

[125] S. Gürses, C. Troncoso, and C. Diaz. "Engineering: Privacy by design". In: *Science* 317.5842 (2011), pp. 1178–1179. ISSN: 00368075. DOI: 10.1126/science.1143464.