

Revealing the Landscape of Privacy-Enhancing Technologies in the Context of Data Markets for the IoT: A Systematic Literature Review

Gonzalo Munilla-Garrido^{a,*}, Johannes Sedlmeir^b, Ömer Uludağ^a, Ilias Soto Alaoui^a, Andre Luckow^c, Florian Matthes^a

^aTechnical University of Munich, Department of Informatics, Munich, Germany

^bProject Group Business & Information Systems Engineering of the Fraunhofer FIT, Bayreuth, Germany

^cLudwig Maximilian University of Munich, Department of Computer Science, Munich, Germany

Abstract

IoT data markets in public and private institutions have become increasingly relevant in recent years because of their potential to improve data availability and unlock new business models. However, exchanging data in markets bears considerable challenges related to the disclosure of sensitive information. Despite considerable research that has focused on different aspects of privacy-enhancing data markets for the IoT, none of the solutions proposed so far seems to find considerable practical adoption. Thus, this study aims to organize the state-of-the-art solutions, analyze and scope the technologies that have been suggested in this context, and structure the remaining challenges to determine areas where future research is required. To accomplish this goal, we conducted a systematic literature review on privacy enhancement in data markets for the IoT, covering 50 publications dated up to July 2020. Our results indicate that most research in this area has emerged only recently, and no IoT data market architecture has established itself as canonical. Existing solutions frequently lack the required combination of anonymization and secure computation technologies. Furthermore, there is no consensus on the appropriate use of blockchain technology for IoT data markets and a low degree of leveraging existing libraries or reusing generic data market architectures. We also identified significant remaining challenges such as the copy problem and the recursive enforcement problem that – while solutions have been suggested to some extent – are often not sufficiently addressed in proposed designs.

Keywords: Anonymization, Big Data, Copy Problem, Data Exchange, Marketplace, Platform, Secure Computation

1. Introduction

IoT devices have been improved, mass-produced, and deployed in the past few decades through steady progress in information and communication technologies (ICTs) and motivated by a trend of data-driven decision-making, automation, and the opportunity of new business models. IoT devices' primary collective purpose is to interact with the physical world and enable the measurement and collection of events and interactions [S23]. These characteristics apply to IoT devices deployed in, for example, a factory or a powerline network and many devices employed by people such as cell phones, laptops, or wearables. The *volume*, *velocity*, and *variety* of the information generated by the IoT is immense, which drove practitioners to coin the term *big data* and develop tools for their analysis [S4]. Public and private institutions use *big data* to promote the public good, innovations, and improve products and services. Big data has become the foundation of the emerging data economy, which in Europe was worth nearly 2 % of its GDP in 2016, close to 300 billion Euros [1]. However, the generation, collection, storage, processing, distribution, and analysis of *big data* to realize such economic potentials also come with challenges for enterprises and responsibilities towards society.

Big data needs to be accessible to institutions who can harness their potential and develop innovations, lest society fails to materialize their advantages. Unfortunately, a significant share of the world's data is siloed and exploited solely by the institutions that host them [2], consequently locking the untapped potential of the data economy and hindering progress in science, business, and society. To surmount this obstacle, a paradigm shift towards openness emerged in the form of electronic data markets for the IoT, i.e., mediums for the trade of information across the Internet based on electronic infrastructure [3]. This paradigm brings potential benefits, such as increasing the efficiency of business processes, facilitating growth by unlocking new business models [4], and profiting from trading. Decision-makers in governments and businesses have recognized the economic potential of data markets and hence recently supported major projects that provide a common digital infrastructure for data-sharing initiatives such as GAIA-X [5] or the automotive-related Catena-X, which promote the collaboration of large enterprises in data markets.

Despite data markets' promise to benefit society by fostering innovation and collaboration across enterprises, these markets hold the risk of exposing individuals' and businesses' sensitive information [6][7]. Moreover, confidence in privacy protection is an essential driver of users' willingness to share information [S8]. Similarly, businesses are unwilling to bear the risk of unintentionally leaking their customers' private or business-

*Corresponding author

Email address: gonzalo.munilla-garrido@tum.de (Gonzalo Munilla-Garrido)

critical information. Consequently, the adoption of data markets is generally hampered. Additionally, while a corporation may have taken security measures for protecting collected data from unauthorized access or unintended use, data buyers might not have the same standards. Hence, in this case, exchanging data entails an additional risk that the seller needs to mitigate *before* the data are shared. Furthermore, blockchains are expected to play an essential role in the ability of institutions to trade data in tokenized form [8], but their inherent transparency further increases the need to make data exchanged in markets less sensitive [S25]. Additional trends that aggravate the negative consequences of lacking data protection for institutions are recent privacy laws such as the GDPR or the CCPA with their increasingly expensive fines for data breaches [9].

As a first reaction to these risks, practitioners and corporations have increased their systems' security. However, if data are sold and, thus, replicated, confidentiality is not sufficient to protect privacy; only managing or modifying the data in a way that enhances privacy while preserving as much utility as possible is effective [10]. Thus, institutions have started to allocate more resources to balance data utility and privacy, employing privacy-enhancing technologies (PETs) [S27]. The term PET was coined in 1995 in a report by the Dutch Data Protection Authority and the Ontario Information Commissioner [11] that explored a novel approach to privacy protection [12]. These technologies take the form of architectures built with privacy-by-design principles and policies [S39][S6], or data modifications based on heuristics or mathematical privacy guarantees. Prominent examples of PETs are differential privacy [13][14], syntactic anonymization technologies like k-anonymity [15], homomorphic encryption [16][17][18], trusted execution environments [19], secure multiparty computation [20], zero-knowledge proofs [21][22], and a set of conventional de-identification approaches such as masking, rounding, or hashing [23].

The relevance of PETs in data markets also increases with the growing adoption of IoT devices, such as in vehicles, wearables, smartphones, and the applications that stream data daily from millions of individuals' private lives to data marketplaces [24]. Despite their current relevance and growing attention [S27], researchers and institutions still find PETs challenging to understand, integrate, and deploy in IoT data markets because most PETs are technically complex and have a wide range of variations and combinations with different tradeoffs [25]. Regarding research addressing these challenges, primary studies are predominant, i.e., studies based on original designs developed or data collected by their authors, while secondary studies are less frequent. The applications proposed by primary studies range from funneling data from markets into machine learning (ML) algorithms [S2][S25][S43], crowdsourcing data into markets [S16][S22][S24][S47], adopting data markets for smart mobility ecosystems [S3], smart manufacturing [S19], smart homes [S11], and smart wearables in the health industry [S48].

On the other hand, 9 out of the 50 studies that we identified in our systematic literature review (SLR) are secondary, and out of these, four studies [S19][S23][S35][S48] cover

some of the PETs available for data markets for the IoT, yet without giving a detailed comparison of their functionalities, benefits, and limitations. The other five secondary studies [S4][S8][S14][S27][S38] perform high-level surveys revolving around challenges, non-technical privacy strategies, and user-centric perspectives on data markets for the IoT. However, none of these secondary studies provides a rigorous, *systematic* review that collects and map PETs and challenges comprehensively. Moreover, as we discuss in Section 7, we noted a low level of re-using existing components to build a more holistic architecture for data markets in related work. This may indicate a need for systematically analyzing the current seminal components, strengths, and weaknesses of solutions proposed for privacy-enhancing IoT data markets.

Consequently, we tackle the aforementioned research gaps with a comprehensive and detailed SLR that aims to guide decision-makers, privacy officers, policymakers, and researchers in the challenge of employing PETs to build or participate in privacy-enhancing IoT data markets. We guide these stakeholders by identifying, classifying, and describing how PETs are leveraged in the current body of scientific knowledge (see Section 6) and presenting key findings from our SLR (see Section 7). Moreover, for the benefit of the reader, we distill terminology from the extant literature to differentiate and navigate the concepts of PETs in the scope of this SLR (see Section 5). We also organize related work into a reference model for the use of PETs in IoT data markets in distinct categories (see Fig. 10 and Fig. C.11) and identify narrow and broad challenges that PETs can tackle or circumvent (see Fig. 8). Through mapping PETs to the distilled terminology and the identified narrow challenges, we want to support practitioners in making informed decisions about the appropriate PET to employ in the context of IoT data markets (see Table 2).

The remainder of the paper is structured as follows. Section 2 introduces the main concepts of privacy, data markets, and the IoT. Section 3 portrays how we conducted our SLR, followed by a discussion of related work in Section 4 and a distillation of terminology in Section 5. Section 6 presents the main results from analyzing the content of the studies in our SLR. Based on these results and the studies' metadata, we extract a set of key findings and artifacts in Section 7. Finally, Section 8 concludes with a summary of the results, points out the limitations of our research and presents potential avenues for future work.

2. Background

2.1. Privacy

Given the increased attention and relevance of privacy during the past decades, practitioners have provided many acknowledged definitions. For example, A. F. Westin [26] states that "[Privacy is] *the claim of individuals [...] to determine for themselves when, how and to what extent information about them is communicated [...]*". Similar definitions have been given by other authors like G. A. Fink et al. [27] or K. Renaud and D. Galvez-Cruz [28]. Despite these efforts, D. J. Solove argues that any attempt to distill a unique, timeless definition is infeasible due to privacy's multifaceted concept [29]. However, in the

field of computer science, a narrower definition may be possible through entailing an *attacker model* that accommodates for the fact that the concept of privacy would likely not have emerged in this field if transgressors would not exist: Attackers of one’s secrets tacitly give meaning to privacy. Therefore, a useful definition in the context of computer science may be F. T. Wu’s [30]: “[Privacy] is defined not by what it is, but by what it is not – it is the absence of a privacy breach that defines a state of privacy.” F. T. Wu hence defines privacy as a product of a threat model, the one from M. Deng et al. [31], in which a practitioner needs to determine what information to hide, from whom, and which harms should be prevented before defining legal and technical tools.

Once IT architectures and tools enhance privacy appropriately, other advantages emerge. For example, from an economic perspective, privacy enables data utilization across organizations and applications to create new fair products and services, and can prevent price discrimination [32]. Furthermore, employing PETs may increase the number of sources and data harvested by institutions because PETs help to overcome regulatory barriers [33], in addition to mitigating the risk of fines, and differentiating and appreciating a brand [4]. Moreover, political freedom and stability may only be achieved by unobtrusive forms of governments [34], privacy-enhancing journalism, and less pervasive forms of digital products such as social media that can enable malicious social engineering [S38]. Moreover, research indicates that compromising privacy can result in negative long-term economic effects [35].

Despite these benefits, and while consumers emphasize that privacy is important to them, they are typically not willing to take small additional efforts or pay for privacy [36]. This has become well-known as the so-called privacy paradox. Thus, in the past decades, governments have enacted rules and laws to protect consumers against violations of their privacy, specifically in the context of data that is captured through advanced ICTs such as personal computers, the world wide web, or smartphones. Examples comprise the European Data Protection Directive in 1995, the HIPAA Privacy and Security Rule in 1996, the APEC cross-border privacy rules in 2011, the GDPR in 2016, and the Consumer Privacy Act in 2020 in the USA, which comprises of Acts such as the CCPA of 2018.

2.2. Data Markets

According to F. Stahl et al. [3], there is a misconception in everyday language between the terms “market” and “marketplace”: A marketplace is the implementation of a market in terms of infrastructure, time, and location (virtual or physical) where the participants transact. Markets are the environments where buyers and sellers set the price and quantity of a particular good. Marketplaces have evolved over millennia; however, the most drastic changes arguably have happened in the past few decades. ICT has driven the costs of instant and ubiquitous communication to an often negligible amount, which has led to the digitization of many existing transaction-based ecosystems, including marketplaces [3]. Moreover, ICT has enabled the creation of virtual marketplaces that did not exist before [37]; potentially the most prominent example being e-commerce. In

this context, data have emerged as goods on their own [3]. Data markets incentivize institutions to collect more data and to profit from trading. In turn, the resulting improvements and innovations benefit the public good [38].

Beyond the above formal definition, from the selected studies, we can carve out several characteristics of data marketplaces: Y. Li et al. [S1] indicate that most of the data marketplaces in operation are centralized, where the platform is run by either a trusted third party (a broker) that coordinates buyers and sellers or by the data owner (e.g., a large institution) who is also selling the data. Their contribution is a decentralized data market to counter the drawbacks of centralized architectures, e.g., trading fairness or vulnerability to a malicious central broker; another 15 selected studies also propose decentralized architectures employing distributed ledger technology. On the other hand, Z. Guan et al. [S46] take another perspective, characterizing data trading platforms depending on the number and type of data domains: general platforms include data from any source type, while specialized platforms focus on one domain, e.g., financial, healthcare, or social media. C. Perera et al. [S27] identify two categories for data markets based on the type of participant: companies or private individual customers, e.g., owners of a smart home. Together, we distill three dimensions for characterizing data markets: (i) the degree of centralization, (ii) the types and number of data domains, and (iii) the types of sellers and consumers. These dimensions that permeate most of the identified solutions, however, all exhibit individual privacy trade-offs that practitioners need to be aware of (see Section 7).

2.3. The Internet of Things

The IoT is considered a network of physical devices that leverage sensors to measure and collect information from the real world and support the access and exchange of this data via the Internet instantly and ubiquitously [39]. IoT devices are considered essential to gather big data [40], which in turn brings new opportunities such as targeted advertisement, predictive maintenance, and quality improvements. Consequently, many companies have introduced the IoT in their strategy for participating in the data economy [39] and make substantial investments in the technologies that make them possible: sensors, wireless networks, and cloud computing infrastructure [40]. In this SLR, the definition of the IoT includes any device with a CPU that is connected to the Internet, including sensors in factories, supply chains, or vehicles, and devices such as smartphones, wearables, and computers that people use daily. These devices act as data collectors and as the gateway to a plethora of applications that collect users’ actions and behavior, such as browsers, social media, e-commerce, or media entertainment, as well as sensor data that is generated in business processes like manufacturing and predictive maintenance, and try to use them for analyses and predictions.

The design and implementation of data markets are dependent on the IoT. The ubiquity of IoT devices generates many constellations for different degrees of decentralization, with a myriad of possible sources and prosumer types. Furthermore, while such ubiquity will likely boost the data economy and its

products and services, IoT devices also permeate many aspects of an individual's life, e.g., dealing with highly sensitive health-care data, or capture information that is sensitive from a business perspective. The sensitivity of the data gathered from IoT devices hence calls for the implementation of PETs.

3. Research process

3.1. Goal and research questions

We employed the Goal-Question-Metric paradigm [41] to formulate the focus of this study as follows: we systematically analyze peer-reviewed literature to provide an overview of the state-of-the-art concerning available research and trade-offs on privacy-enhancing data markets for the IoT as well as potential research gaps from the point of view of both scholars and practitioners. Based on this paradigm, the research questions (RQs) we pursued were:

RQ1. *What relevant PETs enable IoT data markets?*

By answering this RQ, we aim to reveal, describe, and classify PETs in the context of data markets for the IoT both in their fundamentals and application to give researchers and practitioners an overview of the PETs researched and employed so far.

RQ2. *What challenges and trade-offs hinder privacy-enhancing IoT data markets?*

Through answering this RQ, we account for explicit and implicit challenges depicted and tackled in existing work so that researchers may quickly identify pain points in the field and focus their research.

3.2. SLR execution

We opted to conduct a SLR based on the guidelines of B. A. Kitchenham and D. Budgen [42]. SLRs aim to collect, structure, and summarize the existing evidence and gaps of a particular research field to pave the way for future research. Furthermore, SLRs need to provide a rigorous and auditable methodology that can be reviewed and replicated [43]. In contrast to conventional reviews, SLRs define research questions, and a set of predefined inclusion and exclusion criteria that assess potentially relevant primary studies to answer them [44][45]. Table F.11 of the Appendix contains the criteria for this SLR related to focus, quality, and accessibility.

To conduct the *study search*, we identified the most relevant publications in the field of privacy-enhancing data markets for the IoT to answer our research questions [46][47]. To obtain a corpus of high-quality publications, we defined a search strategy based on the work of H. Zhang et al. [47]. Accordingly, our strategy consisted of three phases:

(i) A preliminary search of the base literature. The base literature includes representative papers (8) in the field of privacy-enhancing data markets for the IoT known to the researchers before the SLR, and some other publications found manually in the digital library of the university, which also complied with the criteria described in Table F.11. We created preliminary search strings based on identified keywords and synonyms that

we found in the base literature and research questions. Afterward, we parsed our base literature with a tool to analyze frequent phrases and keywords. Using the results of this analysis, we refined our search terms [48]. Finally, we clustered the search terms into three strings based on the field of this SLR: Privacy, data markets, and IoT. Altogether, we composed the following search strings:

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.*

Consequently, we defined our final search string as **C1 AND C2 AND C3**.

(ii) The main search. Since no single source may contain all the high-quality, relevant publications [49][50], we selected seven electronic databases (see Table F.10 that focus on computer science or software engineering and, according to L. Chen et al., cover the most relevant databases in these fields [51]). The time frame that we specified covers any publication included in the selected digital libraries before the 13th of July 2020. With the defined search string, time period, databases, and in the manner Fig. 1 depicts, we collected 1291 studies (1136 after duplicates removal), which two researchers filtered independently and redundantly by title (119 selected out of 1136), abstract (79 selected out of 119), and body (37 selected out of 79) following the predefined inclusion and exclusion criteria of Table F.11 to reduce bias. After each of the three filtering phases, both researchers resolved conflicts in an informed discussion and attended to the criteria.

(iii) A backward search of the references of the 37 studies resulting from the main search. After filtering by title, abstract, and body, considering our inclusion and exclusion criteria, we included another 11 studies in our corpus. The process resulted in a total of 50 studies from which we subsequently extracted and synthesized data. Hence, the SLR yielded a considerable but not excessive number of results. Furthermore, thanks to the multiple synonyms in the search string, the 37 studies only missed two studies from the base literature. Moreover, the backward search only added a modest number of new works (11). Thus, the process suggests that the choice of search terms was suitable.

To answer the research questions, we performed a *data extraction* of key information from the 50 publications in a structured manner [52]. To reduce the degree of bias, two scientists defined and independently followed an extraction card,

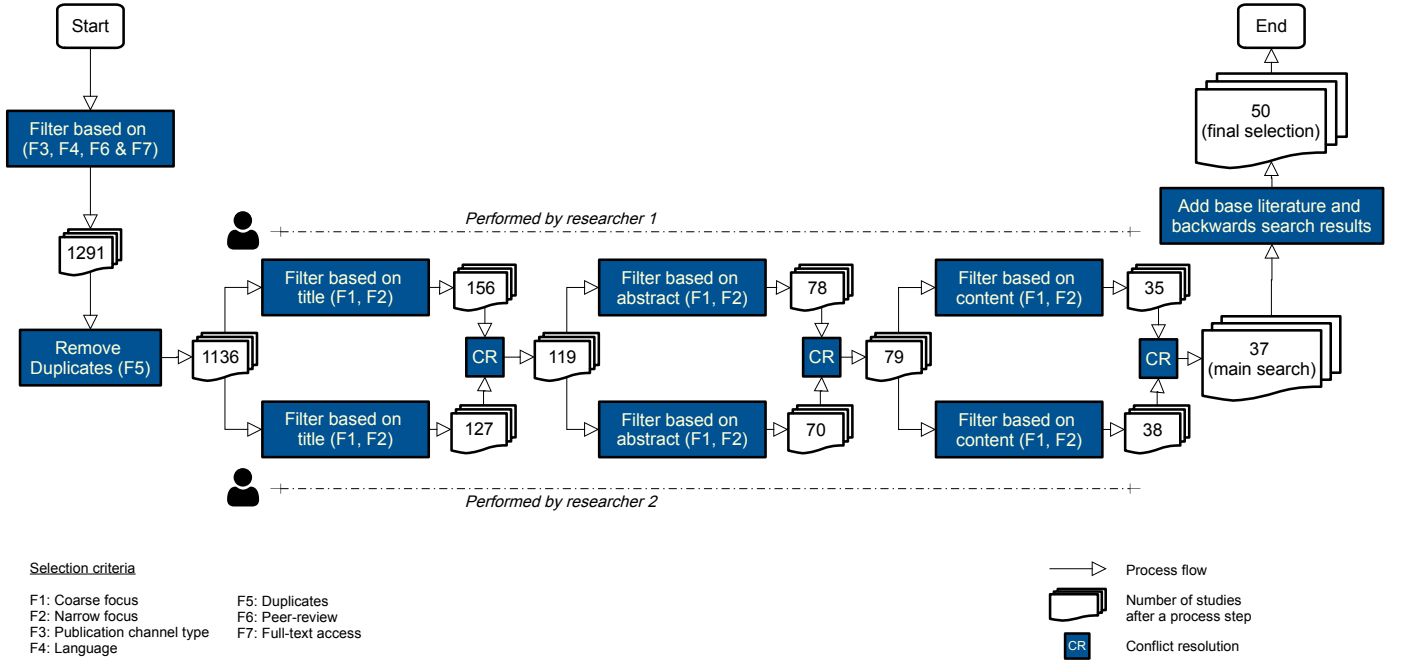


Figure 1: Study selection process.

which contained the following twenty fields: Authors, cite count, year, country, publication channel, publication type, publication source, research type, research approach, contribution type, tags, topic, subtopic, sub-subtopic, research goal, research questions, study findings, privacy-enhancing architecture or technologies, challenges, and future work. After the two scientists completed the data extraction, they held an informed discussion to resolve any possible conflicts on the extracted information. For the *data synthesis* necessary to answer RQ1 and RQ2, we adapted the “narrative synthesis” method described by B. A. Kitchenham and D. Budgen [42] and performed the following synthesis procedure: (i) we developed a preliminary synthesis of the findings, followed by (ii) exploring relationships in the data and (iii) refining the preliminary synthesis with the newly acquired knowledge. After the refinement, we returned to the second step until we deemed the RQs answered.

4. Related work

We found nine secondary studies in our SLR conducted in the context of privacy enhancement for IoT data markets [S4][S8][S14][S19][S23][S27][S35][S38][S48], which we summarize in Table B.3. Some of these studies focus on privacy-related challenges in data markets for the IoT [S14][S19][S27][S38], a second set concentrates on PETs from a technical perspective and discuss their challenges and opportunities [S23][S48][S35]. Moreover, [S8] analyzes users’ preferences in privacy-enhancing data markets, and [S4] lists technical design choices for data markets for the IoT.

These secondary studies provide valuable contributions and build the foundation of our work; however, they focus on different aspects of IoT data markets and, therefore, lack depth in

the concepts we present in this study. [S23][S35][S48] briefly present blockchain technology, secure and outsourced computation, k-anonymity, and differential privacy and sketch their implications while not considering other PETs that we identified in our SLR. Furthermore, although [S19] provides an overview of the available technologies and their challenges, the authors do not discuss PETs in detail, e.g., the study mentions anonymization but does not distinguish between k-anonymity and differential privacy. Furthermore, the authors only discuss a proper subset of the challenges we found: the recursive enforcement problem, the utility and privacy trade-off, and attacks on privacy. J. Pennekamp et al. [S19] base their results on observations from exemplary use cases and, therefore, cannot provide the scientific rigor and comprehensiveness of a SLR for a secondary study. The remaining secondary studies focus on privacy strategies instead of technology, discuss digital rights, describe challenges at a high level, or provide a user-centric view of data markets.

Overall, none of the related work conducts a SLR to establish a holistic view of the body of scientific knowledge developed so far in the field of privacy-enhancing IoT data markets, which is necessary to enhance the academic rigor and reduce the bias of a secondary study [42]. Furthermore, none of the found studies provides a fine-grained analysis and classification of technologies and challenges that have been suggested by research on the context of privacy-enhancing IoT data markets (see Section 6). Lastly, we also provide a mapping of technologies, IoT data market layers, and challenges in Table 2 that not other study has presented in such detail.

5. Terminology

To help the reader with following this SLR, we first provide some terminology. These definitions are the distillation of the concepts found in the 50 selected studies and other seminal studies regarding utility and integrity [53], and confidentiality and privacy [54]. When we use the word *assure*, a technology fully guarantees a quality of the data or computation. In contrast, when we use the term *enhance*, a technology improves a quality of the data or computation to some extent. These qualities concern with *authenticity*, *integrity*, *confidentiality*, *privacy*, and *utility*. In line with the definition of privacy of Section 2.1 in the context of computer science, we define these qualities by the absence of an attack against them, if applicable.

Data authenticity is preserved when a malicious entity has not tampered with the truthfulness of the original data; truthfulness covers both provenance and integrity. In the context of PETs, the degree of authenticity of data can be reduced to enhance privacy. Correspondingly, *identity authenticity* is preserved when a malicious entity has not impersonated another entity. In the context of PETs, the identity of an entity can be concealed to enhance privacy. *Data integrity* is preserved if data that has been copied and stored or is in motion is equal to the original [53]. In practical scenarios where data is exchanged, if *integrity* is not preserved, then *data authenticity* is also inherently violated. *Computational integrity* is preserved when, even in the presence of malicious entities, the output of an algorithm that runs on data is computed correctly. In the context of PETs, the computation can be concealed to enhance privacy. Furthermore, some technologies *enhance confidentiality*, i.e., ensure that data or specific properties thereof are only shared with the intended parties. Furthermore, we refer to *utility* as a measure of the usefulness of data for the successful completion of a task; it is high when the data is *authentic*. However, *utility* can also be improved through PETs, as they may help to facilitate the sharing of data that without applying PETs would have been too sensitive.

We briefly give some illustrative examples for the interplay of concepts: A digital signature can *assure data integrity* provided the corresponding private key is not accessible to an adversary, and *identity authenticity* if the signature includes a digital certificate that a trusted third party issued; otherwise, digital signatures *cannot* assure or enhance *identity authenticity*. Distributed ledgers can *assure data and computational integrity* by replicated storage and computation [55] but these ledgers cannot enhance *data authenticity*; additionally, replication is often problematic regarding *confidentiality* and, hence, *privacy* [56]. Furthermore, zero-knowledge proofs can provide evidence for *data and identity authenticity and computational integrity* without violating *privacy*, and truth discovery can *enhance* these qualities independent of *privacy* considerations. Moreover, privacy-preserving data mining can *enhance* privacy; however, if the right PETs are not employed, qualities such as *computational integrity* may not be *enhanced* or *assured*. As last examples, anonymization technologies such as differential privacy or *k*-anonymity *enhance* privacy by reducing *data authenticity*, and onion routing or ring signatures *enhance* privacy by forgoing

or reducing *identity authenticity*, respectively. Anonymization technologies consequently reduce *data utility* to some extent in exchange for *privacy*.

Lastly, we are mindful of the term *tackling*, which refers to technology directly and fully or partially solving a current challenge in the context of privacy enhancement, e.g., the copy problem or the recursive enforcement problem (REP) (see challenges in Section 6.2). We use the term *circumvent* when a technology bypasses a problem, i.e., the technology does not directly address the issue. However, still, the entities that leverage the technology are not affected by the problem. For example, obscuring the data and computation in a third-party server with homomorphic encryption (HE) does *not tackle* the REP; instead, HE *circumvents* such problem because the third-party server cannot see the contents. On the other hand, distributed ledger technology *tackles* the REP with a redundant and hence tamper-evident storage and execution.

6. Study results

6.1. Identified technologies that enable IoT data markets (RQ1)

This Section describes the PETs and other technologies that we identified in our SLR. Furthermore, we provide a new categorization of technologies that we justify based on the characteristics emphasized in the corresponding selected publications and the technical properties described in this Section.

Some selected papers from our SLR already provide frameworks to classify PETs. For example, S. Sharma et al. [S48] consider two categories for classification: *Outsourced computations*, where the service provider does not handle sensitive data, and *information sharing*. On the other hand, an *Internet of production* study [S19] gives a more fine-grained classification framework with five layers: *data security*, *data processing*, *proving support*, *platform capabilities*, and *external measures*. Furthermore, outside of our SLR, a notable framework developed by A. Trask et al. [10], which was heavily inspired by H. Nissenbaum’s work on contextual integrity [57], dissects an information flow into input, computation, and output, and in each of these three steps, the authors assess privacy and verifiability. Furthermore, they wrap the framework with flow governance, which contains the information flow rules upon which participants agree.

To classify the PETs identified in our SLR, we drew from some of the components of these two studies. While we mapped

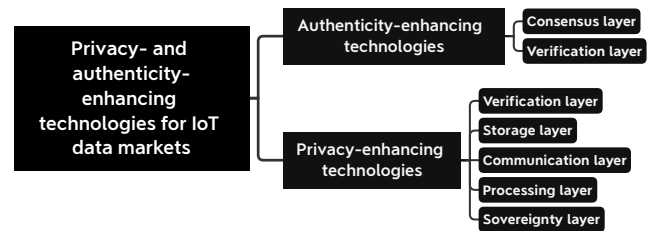


Figure 2: Overview of the categories of our classification of the identified technologies among the selected studies. Note that some PETs also enhance authenticity.

parts of the more general framework of [10] to our *structured* classification of PETs, some of the identified categories of [S19] overlap, such as *platform capabilities* and *external measures*. Therefore, we placed some of the technologies in those categories under the umbrella of *sovereignty*: policies and smart contracts. Furthermore, [S19] and [10] do not map PETs to the privacy aspects of IoT data markets. To tackle this research gap, we distilled from the architectures and concepts of the 50 selected papers the set of layers necessary to build a privacy-enhancing IoT data market: *verification*, *storage*, *communication*, *processing*, and *sovereignty*. In our reference model of Fig. 10, we also include other layers that do not require PETs.

Not all of the technologies that the authors included in [S19] enhance privacy, e.g., version control and most distributed ledger technologies. Therefore, we have introduced another branch for technologies that focus on authenticity in IoT data markets, which we call authenticity-enhancing technologies (AET). Note that some PETs accomplish data authenticity or integrity while enhancing privacy or confidentiality, e.g., zero-knowledge proofs, homomorphic encryption, or some digital signatures (see terminology in Section 5). Specifically, some PETs can also be AETs, but AETs are not always PETs and vice versa. Lastly, the identified AETs can be classified into a *verification* and a *consensus layer* associated with distributed ledger technology, as it deals with coordinating ecosystems while also providing tamper-resistance. We display this classification framework in Fig. 2. Accordingly, we structure this Section into (i) privacy- and (ii) authenticity-enhancing technologies. Within these two groups, we delve into the different layers where we allotted the specific technologies. The technologies included in the mappings are jointly employed by the authors of the selected studies to create holistic or parts of data market architectures for the IoT; the most salient architectures are described in Tables D.4, D.5, D.6, D.7, and D.8.

6.1.1. Privacy-enhancing technologies

Processing layer

The PETs we included in data processing aim to enhance the privacy of either data inputs, outputs, the intermediate steps of a computation, or a combination thereof while maintaining a high degree of utility. This Section follows the structure of Fig. 3.

Processing layer: Secure and outsourced computation

Secure and outsourced computation comprises PETs that enhance privacy through confidentiality. Furthermore, if the PET also employs digital signatures and their cryptography primitives, then the PET can also assure the integrity of the data and computation and identity authenticity in the presence of a digital certificate.

Zero-knowledge proofs (ZKPs). With ZKPs, a technology firstly conceived in the 1980s by S. Goldwasser et al. [21], a *verifier* can verify the authenticity of the data and the integrity of a computation conducted by a *prover* without the need to access the data or replicate the computation itself [22]. If the

statement that is proven is about claims attested in a digital certificate signed by a trusted entity (e.g., age over 18), ZKPs can verify identity authenticity while keeping the information leaked about the identity minimal.

Specifically, ZKPs exhibit zero-knowledgeness, i.e., the *verifier* learns nothing new from the *prover* beyond the correctness of their statement, completeness, i.e., the prover can convince the verifier of a correct statement with high probability, and soundness, i.e., the *prover* cannot convince the *verifier* of a wrong statement with high probability [58][S43]. Furthermore, there are interactive and non-interactive ZKP protocols. With the latter, there is no need to engage in sequential message exchange, and the prover can convince multiple parties of a claim with a single, potentially short, message [58]. These characteristics make non-interactive ZKPs highly attractive for use in blockchains [56]. ZKPs are also the building block of many anonymous credentials schemes, which allow the verification of information in a digital certificate without disclosing any unnecessary data, including the highly correlating value of the signature. Anonymous credentials were initially proposed in 1985 by D. Chaum [59], and developed further with ZKPs and blind signatures [60] chiefly by J. Camenisch and A. Lysyanskaya [61] [62] and by S. A. Brands [63].

Within our SLR in IoT data markets, V. Koutsos et al. [S43] employ non-interactive ZKPs to verify the correct computation of outputs, which, in turn, unlocks the payment from a smart contract in the Agora blockchain, eliminating a third-party verifier. While ZKPs have their limitations due to computational complexity, typically for the prover, and there is still a considerable gap between cryptographers and software engineers [64], we expect to see more publications such as V. Koutsos et al.'s [S43]. This projection is justified by the significant improvements in ZKPs' performance, and ease of use in recent years [65][66] and the availability of an increasing variety of domain-specific programming languages to implement ZKPs, such as bellman or circom in combination with snarkjs. Recently, first research has emerged that uses ZKPs to prove that a machine learning model was trained correctly on specific data [67], and there are many opportunities to leverage them in data markets, such as demonstrating that the input data of a computation was signed by a sensor that has had received a certificate from a trusted third party without revealing the sensor's identity or the data. In this case, the digital signature and certificate can be regarded as AETs, while their verification inside a ZKP enhances privacy and, hence, qualifies ZKPs as a PET.

Secure multiparty computation (SMC). In broad terms, SMC enables multiple parties to exchange information obliviously and jointly compute a function without revealing individual inputs to each other [20][68]. The SMC implementations that we observed in our SLR employ either secret sharing [S5][S10][S34] or garbled circuits [S34]. In secret-sharing-based SMC, each party first obfuscates the input by splitting it into shares. Secondly, this party distributes the shares among the other computing parties. Afterward, each party executes arithmetic operations independently on these shares, and finally, all parties share the outputs to reconstruct the result [20].

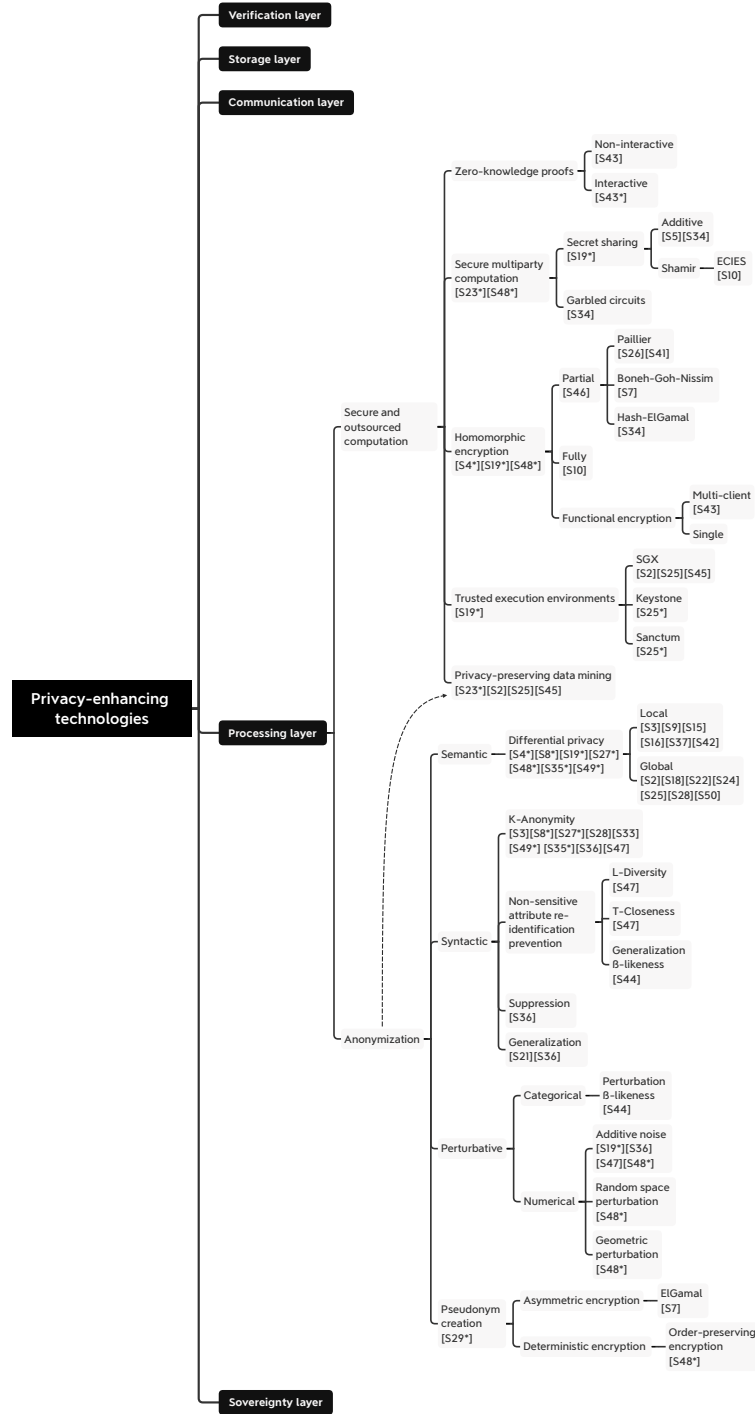


Figure 3: Classification of PETs employed for data processing. Any other privacy approach encountered in the SLR without explicit inclusion of the underlying technology was either not included in a leaf node but in a parent node or completely dismissed if too vague. *The publication reviews or briefly comments on the technology without delving in-depth or using it as a building block of the architecture concept, e.g., included for future work.

In Shamir’s scheme [69], one can specify a minimum of shares that the recipient needs to reconstruct the output, and any combination of fewer shares does not reveal anything about the secret to the receiving entity [S19] [70][71][S48]. On the other hand, in additive secret sharing, all the shares are needed. Outside SMC, Shamir’s scheme has been commonly used for key management schemata for cryptographic systems so that if some shares that represent a private key are lost, one can still reconstruct the key with the remaining shares [69]. On the other hand, SMC can also be implemented by garbled circuits [72], for only two [73] or multiple [74] parties. Garbled circuits are protocols that enable secure computation by using functions translated into Boolean circuits, i.e., a sequence of basic logic gates such as AND, XOR, and OR that may be combined to construct any function [72][75][S48]. Garbled circuits make use of oblivious transfer [76], which in turn utilizes asymmetric encryption, and of symmetric encryption for encrypting and decrypting each gate’s truth table. Lastly, there are SMC hybrids that combine these approaches [77].

SMC allows computing functions without revealing the inputs to other participating parties. SMC protects inputs against brute force attacks and it is to date considered less computationally expensive than alternatives such as fully homomorphic encryption [78]. Drawbacks of SMC include its sensitivity to network latency, which can considerably decrease the performance [S48][S9][S5]. Additionally, SMC protocols often need to be supplemented by mechanisms that prevent collusion [S5]. Moreover, since the individually provided inputs are only locally available, one cannot stop malicious entities from jeopardizing the authenticity of the input with false inputs; SMC can only prevent curious entities from learning information. A countermeasure for this reduction in accountability is zero-knowledge proofs to enforce the authenticity of participants’ local computations while maintaining them confidential [79].

Three of the papers in our SLR implement SMC in their architecture: [S5] uses additive secret sharing, [S10] employs Shamir’s secret sharing, and [S34] leverages a combination of garbled circuits and additive secret sharing. Additionally, other publications acknowledge the importance of SMC schemata by including them in their review [S19][S23][S48]; more details are provided in Table D.4. While several frameworks for SMC are available, the SMC solutions employed by these three publications were handcrafted; this may indicate that the integration of SMC into existing systems requires features that are not yet available with generic tools, such as performance aspects.

Homomorphic encryption (HE). HE allows performing operations on encrypted data (ciphertext) as if they were not encrypted. After the computation, the entities with the corresponding secret key can decrypt the output [80]. There are variations of HE depending on the diversity of operations it can perform [16][17]: Fully homomorphic encryption (FHE) schemata support addition and multiplication, while partially homomorphic encryption (PHE) schemata allow for only one of these alternatives; typically in exchange for drastically improved performance. Any other schema in-between is called somewhat homomorphic encryption [S48].

Five out of the six studies that use HE in our SLR use a PHE variation [S41][S26][S7][S34][S46]. Each of the former four specifies the name of the employed schema, namely the Paillier cryptosystem in the first two [18], Boneh-Goh-Nissim [81], and Hash-ElGamal [82]. The latter study only briefly mentions the additive homomorphic property of their handled data. On the other hand, [S10] uses FHE with a schema called fully homomorphic non-interactive verifiable secret sharing [83]. Several other articles in our SLR underline the importance of HE [S4][S19][S48]. On the other hand, V. Koutsos et al. [S43] suggest the use of multi-client functional encryption [84][85] instead of HE so that the scheme combines data from some individuals with others, and, in turn, malicious entities cannot trace back the output of the computation to a single user, as it may happen in HE. Furthermore, there is a related scheme called functional encryption [S43] that allows to retrieve a pre-specified function executed on a set of ciphertexts [84], e.g., decrypt only the mean of a set of encrypted numbers by deriving a function-specific decryption key from the secret key that was used for encrypting the data. A summary of papers from our SLR that mention or use HE is given in Table D.4.

The major limitation of FHE is its high computational complexity and the comparatively large storage needs of its ciphertext, which poses a significant challenge for its use and is aggravated in the context of IoT devices’ limitations [S41][S7][S48][S24]. Therefore, the approach adopted by most authors is the use of PHE instead of FHE [86], which, while still not as efficient as other PETs, consumes significantly more computing resources than PHE [87].

As observed for the case of MPC, while there exist generic frameworks such as SEAL, HELib, or TFHE, the authors of the publications in our SLR utilized handcrafted solutions, which may indicate the lack of framework versatility or performance. Overall, practitioners and companies may use HE to perform lightweight functions on data privately on non-local resources, e.g., computing in the cloud, which otherwise would be too expensive to maintain in-house. SMC would usually be preferred over HE when the inputs to the function belong to multiple parties. Nonetheless, some selected publications also employ HE in these cases, e.g., when data brokers determine the winner of an auction [S26][S34][S41].

Trusted execution environments (TEE). TEEs were first defined in 2009 by the Open Mobile Terminal Platform as “hardware and software components providing facilities necessary to support applications” that are secure against attacks that aim to retrieve cryptographic key material or other sensitive information. These features include defense against more sophisticated hardware attacks such as probing external memory [19] or measuring execution times and energy consumption. Moreover, TEEs defend against adversaries who are legitimate owners of the hardware or remote access to the operating system that can run the code themselves. TEEs allow a user to define secure areas of memory (“enclaves”) that enhance confidentiality and assure data and computation integrity of the code and data loaded in the TEE [S2], i.e., any other program outside the enclave cannot act on the data. Specifically, TEEs associate

unique encryption keys to computer hardware, making software tampering at least as hard as hardware tampering and certifying the computation results within the TEE. The reason is that the only way to hacking a TEE is physical access to the hardware and, consequently, performing manipulations so that the hardware provides false certifications to bypass remote attestation and sealed storage [S45]. Seal-stored data may not be accessed unless the user employs the correct hardware and software, and remote attestation is a process whereby a trusted third-party assures that the execution of a program in a specific piece of hardware is correct [S45].

Four of the selected papers in our SLR leverage TEEs [S2][S12][S25][S45], and a review mentions their importance [S19]. [S2], [S25] and [S45] propose TEEs to confidentially train and evaluate machine learning models on data available through a data market. While the role of TEEs in data markets overlaps with the use of HE and SMC, authors have preferred the latter technologies to enhance confidentiality in auctions and data processing, which may be due to the limited memory TEEs offered at the time. Furthermore, while there are open-source proposals for TEEs, e.g., Sanctum [88] and Keystone [89], because of their lack of maturity [S25], the reviewed four studies use Intel’s Software Guard Extension (SGX) [90], where Intel is the trusted third party, and, therefore, the single point of failure.

[S2], [S25] and [S45] use SGX for its primary purpose, i.e., confidential computing, and L. Ruinian et al. [S12] employ SGX for their blockchain architecture to perform “Proof of Useful Work”, a paradigm for resource-efficient mining [91] where nodes perform useful computations instead of just computing hashes like in Bitcoin or Ethereum. Moreover, N. Hynes et al. [S2] decided to use TEEs to enhance data and computation confidentiality for machine learning algorithms because of the low performance of SMC and HE on machine learning [92].

Privacy-preserving data mining (PPDM). J. Du et al. [S23] describe PPDM as a means to enhance privacy while extracting useful information from data mining. PPDM is achieved by performing the computation where the data reside, protecting the computation with cryptographic or data perturbation means, or a combination thereof. For example, if the data and computation are cryptographically protected, the computation can run anywhere; this is accomplished by deploying a ML model and input data in a trusted execution environments (TEE) or implementing a ML model using SMC or HE.

Furthermore, one may deploy ML models where the data is stored, using technologies such as federated learning [93][94][95], which trains a ML model across distributed data assets independently and finally aggregates the weights to form a unique model. Alternatively, split learning approaches [96][97] decompose neural networks’ layers into elements and, thus, the input data and labels do not need to be within the same machine. Split learning presents advantages when the local hardware for computations belongs to different network speeds or hardware configurations [98]. PPDM can also be achieved by perturbing input data or weights of the ML model with anonymization techniques such as differ-

ential privacy (DP), resulting in other technologies such as DP-stochastic-gradient-descent [99], where the weight updates are perturbed with noise and, therefore, one may not reconstruct the inputs based on the outputs, which may happen in federated learning [92].

With PPDM, individuals may enjoy a higher degree of privacy than outsourcing the computation transparently to a trusted third party. Data markets can offer an infrastructure leveraged by PPDM, where data prosumers and consumers only need to provide the input data and ML model, much like the studies in our SLR propose [S2][S25][S45] using TEEs to train models with DP. Like data, trained ML models could also be exchanged in markets.

Processing Layer: Anonymization

While the previously presented PETs hide sensitive data from unsolicited parties and thus provide confidentiality while enhancing or assuring data and computation integrity, the authorized receiver of the plaintext outputs may conduct reverse engineering and try to correlate data records with individuals (re-identification attack). Consequently, in general, employing only the secure and outsourced computation PETs is insufficient to provide the required degree of privacy. To tackle re-identification, anonymization technologies can help through protecting non-explicit identifiers and sensitive attributes [S47][S44]; the cost of this protection is forgoing data authenticity and thus decreasing utility. Given the frequency of re-identification attacks, anonymization should be a critical element of any survey or modern online application, and in particular IoT data markets [S8]. Fundamentally, one may observe that anonymization technologies rely on statistics, probability theory, and heuristics, while secure and outsourced computation usually employ cryptography and trusted hardware.

In this sub-section, we describe our findings for the most employed anonymization techniques identified in our SLR, categorizing in two groups[100]: syntactic technologies, which enforce a property on the anonymized data using, e.g., k -anonymity; and semantic technologies, which enforce a property in the anonymization mechanism itself using, e.g., differential privacy. Semantic technologies have an advantage over syntactic technologies, as semantic technologies can also enhance the privacy of individuals that are not included in released datasets, i.e., prevent background knowledge from being used for re-identification, and have a mathematical guarantee of privacy. Additionally, we cover other anonymization technologies not covered in these two groups, namely noise perturbation and pseudonym creation.

Syntactic technologies. In the syntactic technologies category, we identify the implementation of k -anonymity and its variations l -diversity and t -closeness, a newly proposed model called β -likeness, and also their building-blocks, generalization, and suppression. The most frequently utilized model for syntactic anonymization in our SLR is k -anonymity [S3][S28][S33][S36][S47], which was also reviewed or highlighted by [S8][S27][S49][S35]. K -anonymity is a privacy model that guarantees any individual in a dataset to

be indistinguishable from at least $k - 1$ others. K -anonymity achieves this by clustering a set of sensitive attribute values into equivalence classes of size k . However, finding an optimal value of k for minimum information loss is NP-hard; thus, researchers have proposed alternative heuristics [101]. Nonetheless, some of the selected studies use the building blocks of k -anonymity in its simplest form: generalization [S21][S36] and suppression [S36]. Suppression deletes selected data points, while generalization substitutes data points for others that belong to a higher level in a manually pre-defined hierarchy, e.g., substituting a city by a country to make the location less detailed.

The selected studies [S28][S33] apply k -anonymity to aggregate data from a set of entities. M. A. Alsheikh et al. [S47] go a step further by applying l -diversity to have at least l different values in sensitive attributes, and t -closeness, so that the distribution of the sensitive attributes within each equivalence class is at most at a distance t from the overall dataset distribution of that attribute. These two extra steps added to k -anonymity prevent against homogeneity and external knowledge attacks (l -diversity), and skewness and similarity attacks (t -closeness) [100]. Furthermore, D. Sánchez et al. [S36] tailor the use of k -anonymity based on record history, privacy policies, and disclosure context. With their new approach, they prevent a significant decrease of the data utility compared to homogeneously applying k -anonymity to all individuals' records equally.

Nonetheless, there are detractors of syntactic technologies in data markets because of the need for a centralized intermediary that sees and aggregates the data in an, e.g., k -anonymous fashion [S18]. However, practitioners could circumvent the need for a trusted third party by applying HE, SMC, or TEEs that use a semantic technology before the aggregated data is released. Moreover, J. Cao et al. [S44] state that these conventional syntactic approaches are not sufficient because they lack an attacker perspective in the model. For this reason, they design a novel, more complex model called β -likeness that explicitly bounds the additional knowledge that an adversary gains from seeing the released data.

In the context of this SLR, k -anonymity is employed to create synthetic data, before sharing them in an IoT data market. However, researchers also employ k -anonymity for privacy-enhancing location-based services that exchange location data in IoT data markets, whereby similar fake locations hide the real ones. This type of approach fits well with IoT devices embedded in phones, vehicles, laptops, among other mobile *things*. Some of the approaches named by [S3] are *cloaking*, which consists of sending a more extensive region containing the real one; and *geomasking*, whereby the real location is randomly displaced outside of an inner circle but within an outer one. D. Lopez et al. [S3] adopt *geomasking* for situations where low accuracy is sufficient and a high degree of privacy is required. While there exist more recent technologies that generate potentially more accurate and private synthetic data, e.g., employing generative adversarial networks (GANs) [102], or GANs with differential privacy [103], they are more complex than using k -anonymity.

Semantic technologies. The quintessential semantic technology is differential privacy (DP), which appears in its pure form or one of its flavors in 13 of the 35 studies that propose a solution in our SLR. Furthermore, another seven studies refer to DP to underline its importance or drawbacks. The potential reasons behind the high number of references and use of DP are multifaceted. While HE or SMC may protect the inputs and computations' confidentiality, they do not protect against reverse engineering the outputs (re-identification attack). Moreover, As syntactic technologies or other conventional anonymization technologies, e.g., additive noise, lack a mathematically proven guarantee of privacy, DP has become the de facto standard to address privacy for many practitioners [S37].

In broad terms, DP guarantees that the outputs of an analysis (a statistical query or a ML model) on a dataset are "essentially" identical irrespective of the presence or absence of an individual in the dataset. Additionally, DP is agnostic to auxiliary information available in the present or the future. DP is achieved by adding noise randomly sampled from a probability density function, like the Laplacian or the Gaussian. Specifically, the noise limits the output difference of a DP analysis executed on two datasets (one *with* and one *without* an individual) to be no greater than an upper bound, making the outputs "differentially" indistinguishable.

The set of selected studies of our SLR that use DP in their proposed solutions are [S2][S3][S9][S15][S16][S18][S22][S24][S25][S28][S37][S42][S50]. Tables D.5, D.6, and D.7 summarize their proposed architectures. Six of these studies employ DP locally [S3][S9][S15][S16][S37][S42], i.e., the noise is added to the data of an individual. In contrast, the rest of the studies apply DP globally, i.e., on aggregated data. Furthermore, we can cluster the studies into those that focus on a DP data trading design for data markets [S22][S37][S15][S50], crowdsensing data markets [S42][S9][S16][S24], and architectures that host a data market in an attempt to achieve end-to-end privacy [S2][S3][S18][S25][S28].

While DP offers a mathematical guarantee of privacy, however, DP is not a panacea. DP still holds flaws in its real-world implementations [104] that must be addressed by the research community and practitioners in the industry, and DP's combination with ML needs further improvements [?]. In our SLR, S. Sharma et al. [S48] and D. Sánchez and A. Viejo [S36] identify two specific problems with DP: firstly, DP cannot be used when a high level of accuracy is required [S48], e.g., analyzing data from the brakes of vehicles to improve safety. Secondly, with current approaches, releasing an entire dataset with DP is troublesome. Despite these challenges, the authors of [S2][S50] argue that the benefits of DP predominate, as DP can adapt to many use-cases and allows a practitioner to fine-tune the added noise to enhance privacy.

As we already noted with ZKP, SMC and HE, the authors of the publications that use DP in our SLR do not employ open-source DP libraries such as Smartnoise, Google-DP, diffprivlib, diffpriv, or Chorus despite the maturity of the former two; they instead use handcrafted implementations of DP. Aside from syntactic and semantic technologies, other anonymization

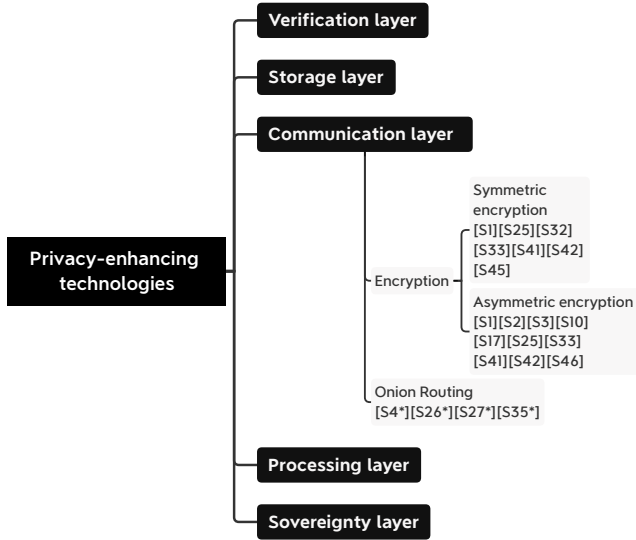


Figure 4: Classification of PETs employed for communication.

technologies are simpler to implement, e.g., sampled data release, character masking, truncation, rounding, top and bottom coding, data swapping, randomization, creating pseudonyms, character scrambling, microaggregation, or noise perturbation [23][105]. The two technologies employed by three studies found in our SLR were noise perturbation [S36][S47] and pseudonym creation [S7].

Perturbative anonymization. Perturbation relies on the use of noise to obfuscate sensitive information. One of the simplest forms of perturbation is additive noise, employed in [S36] and [S47]. Additive noise consists of adding to a deterministic value a random value sampled from a uniform distribution whose bounds are set by a specific percentage of the deterministic value. Furthermore, [S48] reviews two novel perturbative technologies: random space perturbation [106], which strives to protect the query privacy of cloud-stored data by utilizing a confluence of order-preserving encryption, dimensionality expansion, random noise injection, and projection; and geometric perturbation [107], which is motivated by the idea of protecting the geometric transformations that a machine learning model may perform on a dataset rather than the data itself. While perturbative technologies aim to tackle the same problems, unlike DP, they do not provide mathematical guarantees of privacy, even though they are also based on noise addition.

Pseudonym creation. Pseudonym creation is applied to direct identifiers, e.g., names or social security numbers, to enhance privacy while uniquely identifying each record. Practitioners create pseudonyms by hashing or deterministically encrypting an identifier, e.g., using order-preserving encryption [S48], or by applying asymmetric key encryption like ElGamal [S7]. However, researchers have demonstrated that pseudo-anonymization falters against some attacks like profiling, task tracing, or re-identification [S35].

Communication layer

The PETs included in this Section enhance the confidentiality of data in transit or of the sender’s identity (see Fig. 4). These PETs rely on cryptography.

Encryption. Encryption is one of the most fundamental technologies to enhance confidentiality [S19] because after encrypting a piece of data (cipher), only the anointed holders of a decryption key can decipher such data. We underline that encryption cannot guarantee privacy because nothing stops an intended receiver from publicly sharing the decrypted message; this emphasizes the importance of employing anonymization PETs. Encryption may be symmetric (one key to both encrypt and decrypt data) or asymmetric, known as public-key cryptography (two keys, a public key to encrypt, and a private key to decrypt, or vice versa). Encryption is the building block of virtually every secure communication established through a network and takes a key role in digital signatures.

While some publications from our SLR employ asymmetric encryption for the confidential communication of data [S1][S2][S3][S10][S17][S25][S33][S41][S42][S46] (most of these publications employ asymmetric encryption for digital signatures as well, hence the high frequency of digital signatures in Fig. G.14), other publications such as [S1][S25][S32][S33][S41][S41][S45] employ symmetric encryption to also confidentially store data. Naturally, encryption is also a building block for the privacy-enhancing *storage layer* (see Fig. 5).

Onion routing. Onion routing, the backbone of the P2P network resulting from the Tor project [108], consists of a series of re-transmission steps through the network’s nodes. A sender’s message is encrypted once for each step. The intermediary relays decrypt only its appointed encryption layer; thus, the node only knows the immediate sender and receiver, but not the origin of the chain of messages. As the messages are encrypted, the nodes cannot see the contents either. Overall, onion routing renders one’s messages unreadable and untraceable. The paper that suggests employing onion routing in a data market context is [S26], which some of the identified reviews equally appreciate [S4][S27][S35].

However, some drawbacks exist. Implementations backed by Tor have high-latency and redundant communication that challenges bandwidth, which can be challenging to align with a high transactional environment, such as IoT data markets. Moreover, if an architecture decides to use Tor, the network is often blocked by IT departments within organizations or even subject to state-level censorship by some governments [108]. Therefore, to enhance entities’ privacy in a network in these contexts, alternative technologies such as using a VPN can be used. However, these centralized alternatives offer lower privacy guarantees.

As there is not a central authority to set unilaterally privacy policies, one must bear in mind that onion routing enhances privacy only when malicious entities employ networking data such as IP addresses to identify users; onion routing will not help if

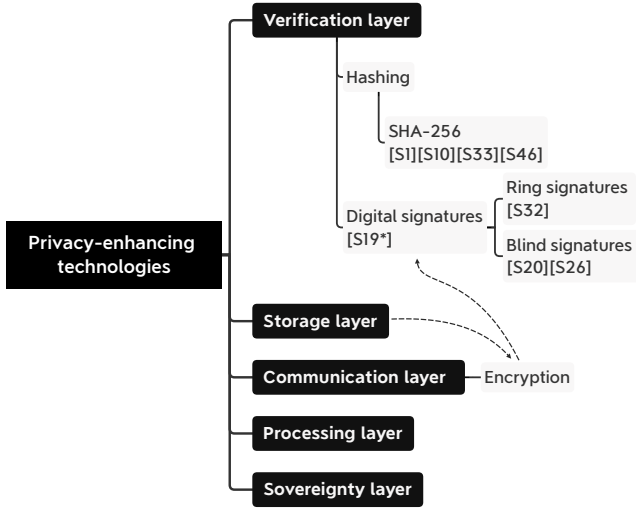


Figure 5: Classification of PETs employed for verification.

the data that users submit to the network is intrinsically sensitive or correlating. To tackle these limitations, practitioners may use onion routing as a building block of a more extensive privacy-enhancing system that leverages other PETs [S26].

Storage layer

The authors of the selected papers that propose confidential storage functionality in their architecture leverage symmetric encryption, mostly AES [S1][S25][S32][S33][S41][S41][S45] (encryption is described in the *communication layer*).

Verification layer

Some of the PETs that support data processing cannot verify the authenticity of data, identities, or the integrity of data [S19] by themselves. The PETs we include in this Section accomplish these verifications with different levels of privacy enhancement. The data processing PETs that can assure identity authenticity and data integrity use the digital signatures of the *verification layer* and the encryption technologies of the *communication layer* as building blocks. Furthermore, the credibility associated with verifying the information exchanged, analysis outputs, and identities can increase the willingness of users to share data [S8]. To navigate this Section, we refer to Fig. 5.

Privacy-enhancing digital signatures (DSs). DS schemata assure data integrity and identity authenticity if accompanied by a digital certificate. As a consequence, DSs also provide with non-repudiation [S1], i.e., actions that an entity cannot deny later. The steps that usually constitute a DS scheme are private and public key generation, encrypting a digest of data with a private key, and a signature verifier that employs the public key to check whether the sender signed the data with the respective private key.

DSs and the encryption primitives of the communication layer are so fundamental that one of the selected studies solely relies on HTTPS for their data market architecture [S17]. However, this architecture does not consider privacy beyond data

in transit; hence, most selected studies rely on multiple PETs. Moreover, although not all of the selected studies explicitly mention DSs, we can safely assume that since DSs are already a living part of virtually any enterprise IT system, most selected studies employ them in their architectures (hence the high frequency of DS utilization in Fig. G.14). Nonetheless, while DSs allow verifying the integrity of data or the authentic identity of the sender, users still need to trust the sender with the authenticity of the data.

So far, we have only described DS as an AET; however, some of the studies selected in this SLR employ two DS schemata based on asymmetric encryption primitives that make DSs privacy-enhancing:

- **Ring signatures** [S32], whereby any party within a pre-defined set of parties could have been the signer of a message; thus, the identity of the authentic signer is kept hidden [109][110].
- **Blind signatures** [S20][S26] are used so that the signer does not have access to the content being signed [60]. It is possible to use blind signatures in combination with zero-knowledge proofs to convince the signer that the content to be signed has the expected properties. Also, one can make an entity sign multiple contents and allow for spot checks to detect fraud. The latter procedure has been employed in the first approaches toward privacy-enhancing payments [59].

Hashing. Hashing is a conventional tool to deterministically map data of an arbitrary length to a fixed output length. In the context of privacy and verification, and specifically aligned with some of the selected studies [S1][S10][S33][S46], hashing is used to verify the integrity of transferred data by hashing the data and making the hash public before transferring the data. In this manner, the recipient can verify the integrity of the confidentially transferred data by comparing the hash of the received data with the previously published hash. Provided the entropy of the data is sufficiently high, nobody except the intended recipient can determine the data from the published hashed value. Hence, hashing can be considered a form of version control with a privacy component.

The hash function employed by the aforementioned publications was SHA-256; their authors commonly use the published hashed data on distributed ledger technologies to ensure its immutability and availability. In this setting, hashing enhances the confidentiality of the sender's data while the parties (network nodes) ensuring the integrity of the ledger (and inherently the persisted hash) cannot unveil the original data. The original data is only viewed by the intended receiver, which validates the integrity of the data received through another channel with the hash persisted in the ledger.

Sovereignty layer

The *sovereignty layer* deals with the concept of information control, the perceived ability to govern what is exposed from one's data [54]. Specifically, based on an entity's requirements,

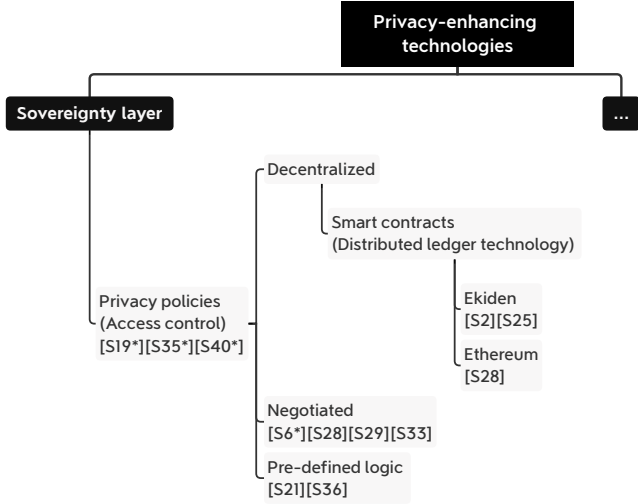


Figure 6: Classification of PETs used for sovereignty purposes.

this layer defines the entity’s rules and guidelines regarding ownership and management that can indirectly govern data processing, verification, and other IoT data market layers. Furthermore, to prevent entities’ violation of privacy, practitioners should map these rules to the PETs capable of fulfilling them. For example, GAIA-X’s high-level architecture contemplates privacy policies in their data *sovereignty layer* [5]. This Section is structured as per Fig. 6.

Privacy policies and privacy by design. Privacy policies embody the requirements and guidelines of a data governance model and are meant to be part of any privacy-enhancing application. To define them, given the regulatory and human aspect of privacy policies, it is also helpful to adopt perspectives from definitions beyond computer science, such as the one underlined in Section 2. A. F. Westin [111] indicates that the privacy requirements depend on the recipient of the information, e.g., an individual can have different reservations when disclosing information to a family member than to the government. Privacy policies should reflect this definition, which means that individuals should express the privacy policies that they expect. Several studies in our SLR explicitly propose policies as part of their solution [S6][S21][S25][S28][S29][S33][S36], while many others review privacy policies [S19][S35][S40] or mention similar ideas. For example, E. M. Schomakers et al. [S8] do not provide a concrete implementation or explicitly name privacy policies; however, they mention that in data sharing scenarios, the data owners should be able to control some fundamental aspects: data types to share, with whom to share, the required degree of trust in another party, the purpose of sharing, and for which benefit.

Among the publications discussing privacy policies, there is a discernible classification. Four of these publications [S6][S28][S29][S33] consider privacy policies as a negotiation between the user and a third party, such as the data consumer or data broker. S. Spiekermann et al. [S6] provide a set of legal requirements and high-level technical solutions that facilitate the introduction of policies in international data mar-

kets, e.g., write policies in a standard language. On the other hand, S. Duri et al. [S29] present an approach in which the data owner can choose among a set of four privacy policies, which includes how data is aggregated, and S. Kiyomoto et al. [S33] relies on a *privacy policy manager* that acts as a gatekeeper and manages the privacy settings from a set of users. Furthermore, another set employs a pre-defined logic to execute PETs based on the desires and track record of the data shared by the individual [S21][S36], while the last set relies on smart contracts for decentralized pre-defined [S2][S25] or negotiated [S28] policies.

Nonetheless, the implementation of policies faces challenges. Firstly, there may be multiple colliding policies, i.e., applications must prioritize policies depending on the context [S21]. Secondly, there is no uniformly accepted global standard for electronic privacy policies [S6]. C. Perera et al. [S40] investigate how practitioners model privacy policies in different domains, focusing on IoT applications. They point to the lack of a uniform standard and propose to utilize ontology-based privacy-knowledge modeling. Thirdly, policy enforcement also causes overhead and an increase in latency due to the need for compliance checks and a lack of automation [S21]. Furthermore, conventional users should include their privacy preferences with minimal manual effort; otherwise, they could be overwhelmed. C. Perera et al. [S40] suggest using recommender systems based on similar users’ data to address this issue; however, this solution may incur a biased recommendation. Moreover, data acquisition costs for privacy policies should not be too costly regarding computational resources since they should scale to a growing number of transactions [S40].

Privacy policies are crucial to protect users’ privacy; however, they are not enough. Organizations must consider privacy issues at each stage of the data pipeline, contemplating aspects that escape user-defined or mutually-agreed policies; and take into account that typically, neither users nor data brokers will be privacy experts. If a user does not know the potential harms of sharing sensitive information such as DNA data, a data consumer may take advantage of the user. Therefore, while privacy policies are a stepping stone towards end-to-end privacy, practitioners must develop systems with a privacy-by-design philosophy [S36][S4].

Privacy by design is a term coined in the ’90s by the former information and privacy commissioner for the Canadian province of Ontario, A. Cavoukian, who created seven principles [112]. Privacy by design claims that privacy goes beyond current regulations and must be an ever-present concern in the minds of organizations [112]. Following privacy-by-design principles entails, for example, preventing sensitive information extraction by default [S36], minimizing the amount of shared data at each exchange [113], and increasing the price of large data packages [113], among others. However, adopting these design principles comes with effort, as it forces developers to adapt their design patterns. For example, current homomorphic encryption techniques force data scientists to express their analysis in terms of additions and multiplications, and differential privacy requires new software engineering design patterns that track the privacy budget of individuals or data scientists.

Smart contracts (SCs). A SC alone is mainly equivalent to conventional scripts; however, because SCs are executed in distributed-ledger-technology-based architectures (DLT), SCs inherit from DLT their enhanced availability and integrity guarantees [55]. DLTs execute SCs synchronously on every node of a P2P network if an arbitrary transaction demands a function's execution. Once deployed, no one can change the script, not even the creators (unless there is an intended call of the script that enables modification), and the script will remain in the network as long as the network exists unless specified differently (e.g., through a self-destruct call). This inherited integrity property of SCs makes them a unique tool to specify and enforce policies between parties or for any other process where no trusted third party is available.

Within this review, all the studies that used the Ethereum, Quorum, Hyperledger Iroha, or Ekliden blockchains rely on SCs to declare privacy policies [S2][S25] (Ekliden) [S28] (Ethereum), fair auctions [S3] (Hyperledger Iroha), or payments or incentives [S32] (Ethereum) [S43] (Agora) [S3] (Hyperledger Iroha). However, while SCs ease verification and enable democratic proposals of privacy policies, SCs also inherit the privacy flaws of DLT, i.e., SCs by default imply the disclosure of data and computations to all DLT network nodes [56]. For example, the architecture from R. Cheng et al. [S25] employs SCs to set user-defined policies, yet it relies on trusted execution environments to enforce them. SCs alone cannot enforce privacy policies without relying on other PETs; the only privacy-related feature that a SC can offer to an IoT data market is declaring privacy policies.

Data access control. Data access control refers to allowing an organization or an individual to choose *who* has access to *which* data. In data markets, access control represents a subset of privacy policies and may utilize different PETs to enforce access rights. While access control is a long-established approach, R. Cheng et al. [S25] propose a novel method, using a key-rotation system [114] in combination with a key manager. Thus, the potential impact of a leaked key is only temporal, with the downside of shorter access permissions.

6.1.2. Authenticity-enhancing technologies

The authenticity-enhancing technologies (AET) included in this subsection focus on enhancing the authenticity of data and identities, also covering data integrity as described in Section 5. Some of the AET that we describe incorporate privacy-enhancing features, while others do not address or even aggravate privacy protection issues and thus need to be combined with PETs.

Consensus layer

Distributed ledger technology (DLT). While DLT may take different forms, most architectures follow the blockchain design pattern, except for IOTA, which uses the so-called Tangle [115]. A blockchain is a tamper-proof distributed database whose state

is stored, synchronized, and replicated by nodes in a P2P network following a consensus algorithm [55]. By its distributed nature, the shared ledger becomes a medium to verify claims, data, payments, or contracts, as once an entity writes something on the ledger, it is practically impossible to modify or erase this record in the future. This property makes blockchain a decentralized and highly reliable alternative to conventional auditing methods like version control [S19].

Benefits of DLT in IoT data markets are the ability to represent the governance, distribution, and roles of authorities on a technical basis [S3], and the enforcement or transparent storage of pre-defined rules by the architects of the respective platform [S25]. Other benefits include eliminating the need for a trusted third party, which removes a single point of failure, improves censorship resistance, and provides more robust data and computational integrity guarantees. DLTs also enable payments through their often built-in cryptocurrencies or other payment systems implemented via smart contracts [S32][S46].

However, some of the studies in our SLR also point at the challenges of current DLT designs: IOTA fails to deliver regarding throughput [S28], is still centralized [S20], and provably has security flaws [116]. Furthermore, blockchains exhibit low transaction throughput [S25], high latency [S11], limited storage [S1] and scalability [S11][S25], computational overhead [S25], high energy consumption [S12], and, most importantly, excessive information exposure that can entail a privacy violation [117]. However, some of these aspects can be mitigated. For example, the energy consumption issue only concerns proof-of-work blockchains [118], and performance can be improved to some extent by private permissioned blockchains that restrict participation in consensus and read access to a small number of nodes in a consortium [119].

Despite the possible operational improvements, employing a DLT for a privacy-enhancing IoT data market needs in-depth consideration. Firstly, through highly replicated storage, a DLT is not suitable to store large amounts of data produced by IoT devices, not even in a privacy-compliant manner. Consequently, most architectures of the selected studies transfer data through interplanetary file systems [S28], employ a hashing verification approach as described in the *communication layer* [S1][S10][S33] or use Merkle trees [S46]. Secondly, while DLT allows for disintermediation and verification in a trust-less manner, it exposes to the network whatever information someone writes on the ledger for as long as the network exists, which may, among others, violate GDPR's Article 17 "*Right to be forgotten*" for personally identifiable information [120]. Lastly, even if an organization uses a DLT only for the matching and clearing steps of an auction, potentially sensitive business information such as turnover can become available to other network participants, which can conflict with antitrust regulation.

Despite the privacy and performance issues of DLT, 31 % of the papers included in our SLR implement a DLT as the backbone of IoT data market architectures, employing the Ethereum blockchain [S1][S13][S20][S28][S32][S45][S46], Quorum [S18], the Agora blockchain [S43], Hyperledger Iroha [S3], Hyperledger Fabric [S10][S33], IOTA [S20][S28][S30],

Intel’s TEE-based consensus Rem [S12], and Ekiden [S2][S25]; while a few only consider them agnostically [S11][S49] or in a review [S19][S23]. The most salient architectures are described in Tables D.7, and D.8.

Some of the selected studies include privacy-enhancing features in their stack. For example, Quorum supports private transactions and private contracts through a public-private state separation and P2P encrypted message exchange for the direct transfer of private data [S18]. However, the interaction between the private and public ledgers is thus naturally limited and cannot be directly applied, for example, to an on-chain payment system. Another example is Ekiden, which offers a horizontally scalable blockchain potentially capable of hosting end-to-end privacy-enhancing applications through key management protocols and Intel’s TEEs [S25] (Note that [S12] uses these TEEs only for consensus, not privacy). Like the solution that W. Dai et al. [S45] present, Ekiden allows for smart contracts to execute data analysis in TEEs. However, it is essential to note that these DLTs accomplish the described privacy and integrity functionalities not because of the DLT characteristics but by leveraging the PETs described throughout Section 6.1.

Verification layer

This verification layer corresponds to AETs that can be employed for the verification of data and identities. We structure this Section according to Fig. 7.

Truth discovery (TD). TD encompasses algorithms that aim to find the authentic value if different data sources provide conflicting information. As a consequence, TD enhances data and computation integrity and can also enhance identity authenticity, e.g., through reputation systems [S9]. In our SLR, we found that TD takes different forms. For example, the survey by J. Du et al. [S23] mentions a mechanism called peer-prediction-based trustable data aggregation [121] in which the system administrator rewards participants for predicting outcomes of arbitrary events based on other participants’ data. This design creates incentives for honest reports and therefore enhances data correctness, resulting in almost all participants choosing to report their bids truthfully [S23]. Moreover, J. Du et al. [S23] also propose mutual validation in which an IoT device compares its data with that of other nearby IoT devices. However, this only applies to specific measurements that are positively correlated for neighboring devices, e.g., temperature, speed of a vehicle, or location in particular settings; and it seems challenging to establish generic handling of differences.

Other TD approaches are majority voting, implemented by Y. Li [S9] in their crowdsourcing architecture and by W. Dai [S45] in their data processing-as-a-service model. The last identified approach creates a reputation system around the application; system designers may use it not only for the quality of service but also for the quality of data, which Y. Li [S9] also employs. While TD leverages transparency to enhance data and identity authenticity and data and computation integrity, TDs are flexible to include PETs such as ZKPs, SMC, HE, TEEs, and DP, which could add privacy features. Furthermore, TD can also tackle the oracle problem of DLT, i.e., nodes within

the network cannot assure the authenticity of data from outside the network, e.g., the price of a physical asset or the result of an election. For example, ChainLink [122] is an initiative that utilizes incentives to create a trusted oracle network and incorporates many of the principles of TD.

Digital signatures (DS). While DS¹ schemata are commonplace for authentication purposes in today’s IT architectures, we have found within the selected studies the use of two notable public key cryptography (PKC) schemata that offer some convenience-related advantages over conventional PKC systems:

- **Identity-based** [S7], where a key generation center (KGC) creates a secret key in a way that the entity’s public key can be a publicly available unique string, e.g., the entity’s email address. The KGC must be trusted because it holds the master secret key from which all parties’ secret keys can be derived. This digital signature assures identity authenticity as the signature is digitally certified by the KGC.
- **Certificateless** [S12] DS schemata are a special form of identity-based PKC whereby an entity’s private key is generated by both the entity and a KGC so that the KGC is not aware of the private key of the entity. However, the entity can prove that the KGC was involved in the key generation [123]. This approach assures the authenticity of an entity while tackling the single point of failure of the KGC.

Decentralized identifiers (DIDs). Identifiers, such as mobile phone numbers, ID cards, user names, or emails, can link an entity electronically across multiple IT systems. These links are sometimes but not always unique and are facilitated by identity providers that centrally host registries of these identifiers’ [124]. In contrast, DIDs are globally unique (with certainty through publishing them on a DLT or probabilistically through randomized generation) identifiers decoupled from centralized registries. DIDs essentially correspond to URLs linked to a file containing one or several public keys and associated metadata that specifies the policies of controlling or interacting with the associated identity. There are two studies in our SLR that employ DIDs in combination with DLT in their conceptual frameworks [S3][S20], described in Tables D.7 and D.8, respectively.

Digital fingerprints (DF). DFs are unique physical identifiers that can be attached to or are inherent of items, and thus, one can be sure to interact with, e.g., the right IoT device [S19]. At a high level, DFs can be seen as a form of version control. However, attaching an identifier securely to a physical object is difficult unless it has a unique property, e.g., unique metal patterns in the soldering of a chip. However, even if the attachment is relatively tamper-proof, e.g., with a crypto-chip, the same

¹We introduced the fundamentals of DSs in the *verification layer* within the PETs branch.

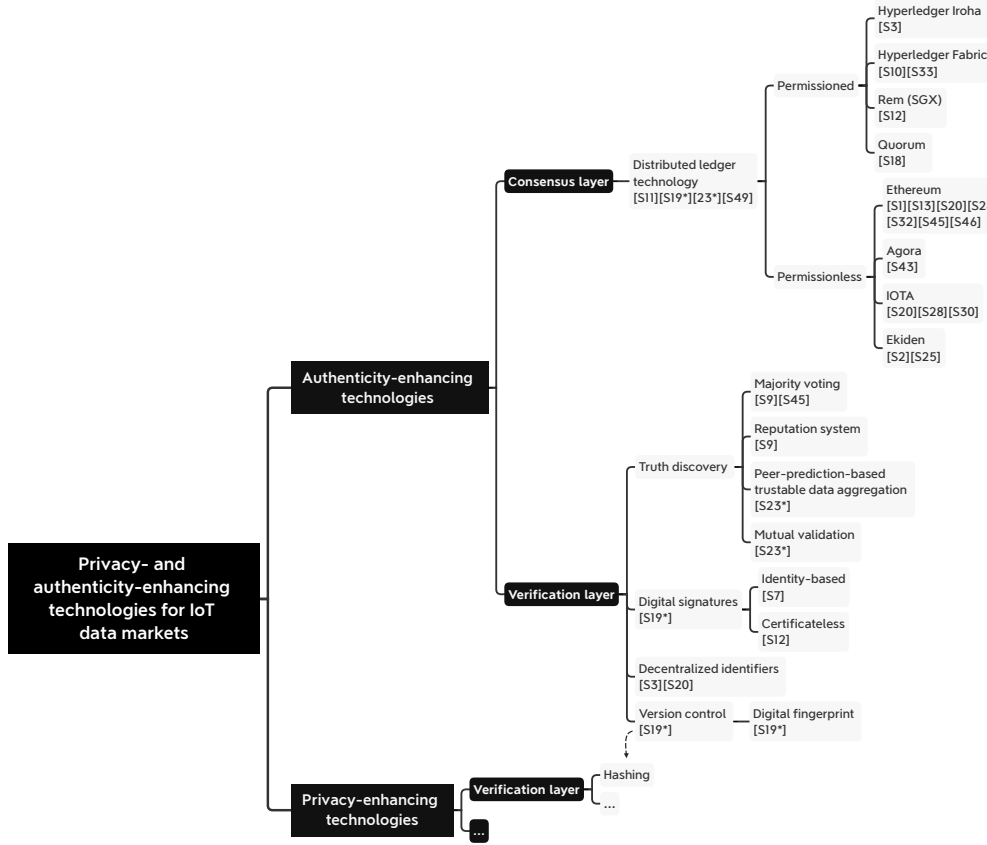


Figure 7: Classification of the identified authenticity-enhancing technologies in the selected studies of this SLR.

problem also pervades the items that interact with the digital fingerprinted item, e.g., tracking scanners. Therefore, despite the authenticity assurance of DFs, their authentication can only be as truthful as the honesty of the devices that scan the DF.

6.2. Challenges and trade-offs that PETs face in IoT data markets (RQ2)

This Section distills the implicit and explicit challenges unveiled in our SLR and other seminal studies [10][25] concerning privacy in the context of IoT data markets. We further classify them into narrow and broad challenges depending on the scope of their definition. Fig. 8 summarizes and outlines the structure of this Section.

Narrow challenges

Aside from the inherent complexity and low maturity of some PETs and the compatibility issues with legacy systems [25], we identified another specific set of challenges tackled or circumvented by the selected studies.

The trade-off between utility and privacy. Practitioners working with personal data face the challenge of balancing the enhancement of individuals' privacy with the preservation of data's utility [S14]; this challenge is explicitly mentioned by some of the selected studies [S6][S9][S13] and implicitly tackled by others [S2][S24][S25][S35][S47]. This dichotomy is the underlying reason behind the tension between data owners and

consumers; the former aim to maximize privacy while the latter intend to maximize utility, which, in turn, is frequently determined by data authenticity (see terminology in Section 5). Furthermore, privacy officers should consider balancing this trade-off at each stage of an information flow [10]: input, computation, output, in transit, and at rest, which increases the complexity of the task. On the other hand, decision makers' or data scientists' quality of judgment depends on computational integrity and the authenticity of data and identities, which is affected by the privacy-utility trade-off.

While PETs from the secure and outsourced computation category seem to circumvent the utility-privacy trade-off by concealing inputs, computation, and outputs, the anointed recipients of these outputs can still perform a re-identification attack. Thus, anonymization PETs, such as differential privacy, should also be included in the stack as they lower the success of re-identification attacks [10].

Data and identity authenticity and accountability encompass another notable conundrum from the utility-privacy trade-off. Some PETs, namely anonymization technologies such as differential privacy, increase plausible deniability at the expense of reducing authenticity and, therefore, accountability [S14][S43]. If the data is fuzzy, the data owner may claim that such a result is not resembling the truth, which is favorable for individual users. However, in some contexts, such protection is not beneficial for society, e.g., in criminal contexts. Regarding authentic data from fuzzy identities, if an authority cannot trace data back

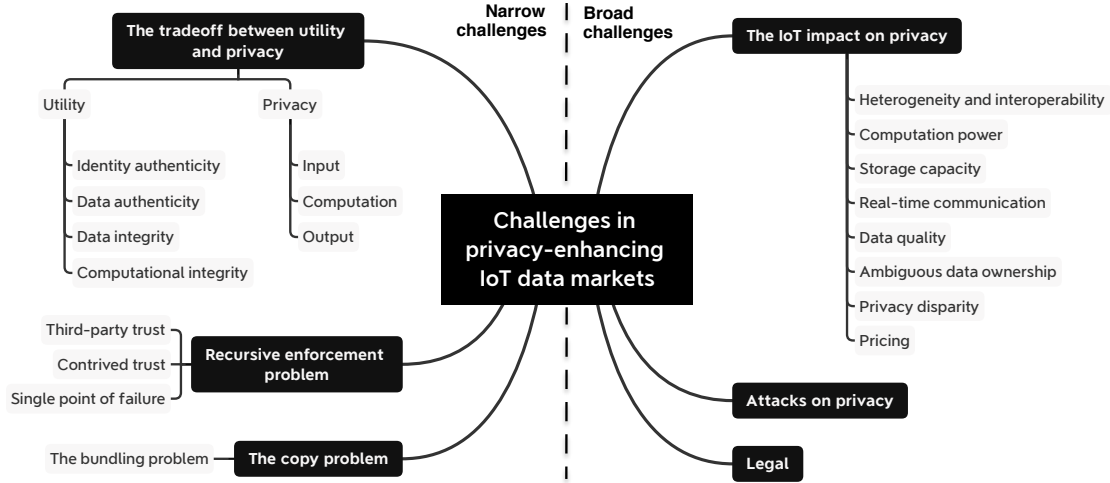


Figure 8: Overview of the narrow and broad challenges facing privacy-enhancing IoT data markets.

to the origin, an individual could try to claim plausible deniability, which would hinder processes such as tracking COVID-19 patients to improve pandemic countermeasures. For practitioners to find a balance in these contexts, the requirements of any application or platform should first define the accountability of the involved entities to strike an optimal balance between utility and privacy.

Regarding accountability in data markets specifically, mechanisms to punish misbehavior, such as banning an entity for re-selling or not selling authentic data [S7], can be beneficial to enhance the utility of the market. While AETs such as truth discovery, e.g., majority voting [S9] [S45] or reputation systems [S9], incentivize market participants to report honestly about the exchanged data, their identity, and computation integrity, among others, the privacy of the entities is not necessarily enhanced. Additional related issues that may arise are verifying the purchased data’s authenticity without violating the individuals’ privacy [S7]. For example, if a data broker sells analysis outputs (insights) and not the (privacy-enhanced) original data, the original data owner’s digital signature is not valid to authenticate the insights [S7]. Nevertheless, for example, systems could use zero-knowledge-proof-based authentication of data and computation, coupled with value deposits locked in smart contracts to hold participants accountable through enforcing reimbursement across intermediaries. Moreover, other nascent solutions exploit the ubiquity and proximity of IoT devices in specific contexts because the data gathered is likely to be correlated, which allows for mutual data authenticity verification among IoT devices [S23]. However, exploiting correlation can only be used for specific measurements, e.g., weather conditions, vehicle speed, location, among others.

The recursive enforcement problem (REP). A. Colman et al. [S49] define trust as a “[...] *directional relationship between two entities – a trustor and a trustee – where a trustor trusts a trustee to perform a specific action within a given scope.*” Therefore, following this definition, users automatically become trustors of the application owners when they engage in digital activity. However, the number of data breaches [9] and

privacy scandals such as Cambridge Analytica indicate that this trust is not always deserved. The REP encompasses the underlying problem of third-party trust with more nuance: Given a third-party authority (*A*), there ought to be another authority (*B*) to supervise *A*, so that *A* can be trusted; and there should be yet another authority *C* to supervise *B* [10], and so forth.

The REP is a significant challenge that has been covered and tackled implicitly by some of the studies in our review [S1][S3][S4][S7][S10][S16][S25][S41]. Additionally, others tackle a sub-set of the REP, which is the single point of failure of trusting a unique third party [S6][S12]. According to E. Schomakers et al. [S8], the hesitation in trusting third parties is one of the main reasons for the slow adoption of IoT data markets. Indeed, it is hard to technically ensure and prove that the third party will not use one’s data for purposes other than those agreed [S8]. Additionally, adoption is further slowed down because the use of third parties to supervise other parties incurs costs [S46]. Furthermore, users’ daily interactions with “trusted” third parties can be regarded as a product of *contrived trust*, another form of the REP. For instance, applications from large service providers with negligible competitors push users to accept the sometimes poor privacy conditions, e.g., GPS apps. Note that *contrived trust* is different from the trust users have on cryptography, open-source code, or consensus mechanisms that the broad scientific community has audited over the years.

Tackling the REP requires reducing the power and the responsibility of the third party in a particular aspect of a specific service by, e.g., distributing such responsibility among other parties or distributing the power among multiple parties that enforce rules on each other. These measures can ease the hesitation to trust a single third party, tackle contrived trust, and reduce the single point of failure, because the third party would be supervised and held accountable by other third parties in a flat hierarchy. Fortunately, the PETs included in this study can also circumvent — only onion routing can tackle — the REP, which, in turn, reduce the need for third-party trust and, therefore, reduce contrived trust and a single point of failure. Ad-

ditionally, 5 of the 7 AETs included in this study can tackle the REP, primarily distributed ledger technology, whose architecture was purposefully built to tackle the byzantine generals problem [125][126], a manifestation of the REP.

The copy problem (CP). Once an entity releases data freely or for profit-seeking, the data is no longer under the original owner's control. Consequently, the recipients of such data can *copy* and, e.g., re-sell or use the entity's data for a non-agreed purpose without informing or acknowledging the original owner [10]. Beyond the privacy threats the CP entails for users of service providers under poor privacy conditions, the CP is a major obstacle for organizations to engage in data markets, which some of the selected studies implicitly tackle [S2][S25][S45]. The CP leads companies to either hoard or sell data as fast as organizations obtain the data, lest its value drops [10]. Nonetheless, secure and outsourced computation PETs such as trusted execution environments or homomorphic encryption can tackle the CP by allowing other entities to extract value without ever losing control over the input data beyond the specific information sold, such as the result of an algorithm evaluated on this data. This paradigm is profound because tackling the CP makes data scarce to some extent, as the original data is not shared, and the data owner would not allow a non-agreed computation. Thus, selling the *access* to data can be more attractive to companies, as data would preserve their value longer than releasing the data.

A subset of the CP is the *bundling problem* (BP) [10], which is an attack vector different to re-identification that occurs when an entity requests actively or passively more data than strictly needed to (i) prove a claim or (ii) perform an analysis. Harvesting more information than needed worsens data breaches' consequences for individuals and companies and indicates questionable business ethics. For instance, (i) to prove one's age with an ID card, the prover usually shares all the information contained in the ID instead of only the age and proof of the card's authenticity. The BP is a subset of the CP because if one tackles the CP, neither necessary nor additional information is released beyond the required computation. For example, tackling the CP by restricting verification and processing to a trusted execution environment also tackles the BP. In this setting, the data consumers cannot copy the necessary data or metadata for other unsolicited analyses, despite being able to process metadata to verify the authenticity of the data and the integrity of the computation and obtaining the desired outputs of the analysis. Additionally, (ii) anonymization-based PETs such as differential privacy or k-anonymity reduce data authenticity to tackle the BP. For instance, in a demographic analysis that only requires the first digits of the ZIP code to perform clustering, data curators can generalize the ZIP codes with k-anonymity, so only the strictly necessary information is revealed to data scientists. Nonetheless, anonymization can suffer from background-knowledge-based attacks [S48][127] and does not solve the CP because the data consumers can replicate the privacy-enhanced data.

Broad challenges

The IoT impact on privacy. The paradigm brought by the IoT brings significant amounts of data to markets; however, this paradigm also bears some of the shortcomings of IoT devices [128]. Table 1 contains an overview of these challenges and briefly discusses their impact on privacy. In summary, privacy is always affected by the context and employed technologies, which underlines the importance of adhering to privacy-by-design principles [112] and the need for practitioners in other fields such as software engineering, economics, law, and politics to tackle together the diverse issues that IoT entails for privacy.

Attacks on privacy. Adversaries can be malicious, actively trying to breach users' privacy through hacking, or honest but curious, passively gathering data from users to reveal hidden insights [S35]. Both of these entities can carry re-identification attacks with the collected information. Within the context of the IoT, the list of security and privacy attacks is extensive (sniffing, cache poisoning, DoS/DDoS, sinkhole attacks, replay attacks, among others) [131]. Furthermore, within our SLR, W. Dai et al. [S45] discuss some of the additional attack vectors these malicious or curious entities may execute in the context of IoT data markets to learn sensitive information from users. Notable ones include: *Data forwarding*, which is one way the *copy problem* materializes; *roles collision*, where data brokers and buyers may be the same or collaborating entities, and, therefore, the broker could rig the auction for its benefit and access the sold data; and *side channel attacks*, where attackers exploit the physical properties of the hardware or its power consumption to extract knowledge from the hidden computations (trusted execution environments suffer mainly from this attack).

Such attacks make the possession of data intrinsically risky because if attackers are successful, data re-identification is possible [S38], even if data have undergone some form of privacy enhancement [132]. There are common attacks used to re-identify data, e.g., reconstruction, tracing, or linkage attacks [104][133]. Some of the most famous re-identification *white-hat* attacks involve A. Narayanan and V. Shmatikov [127] who deanonymized the Netflix Prize dataset with IMDB's public dataset in 2008, M. Archie et al. [134] who performed the same feat with Amazon's public review data, and L. Sweeney et al. [6] (the inventor of k-anonymity) who re-identified participants within a genome sequence dataset in 2013. Furthermore, in 2014, X. Gao et al. [7] tracked drivers with home address and vehicle speed as inputs, and in 2020, D. Kondor et al. [132] matched users with large-scale mobility datasets from a mobile network operator and transportation smart card usage.

Overall, the fact that IoT data markets will facilitate access to large quantities of data from different domains, including biometrics, will increase the impact of these attacks and the potential harms to individuals, e.g., insurance, employment, or price discrimination. Therefore, IoT data markets require a more robust adoption of PETs and security standards.

Legal challenges. Progressively along the past decades, governmental institutions have released laws to protect the privacy

Table 1: Overview of challenges brought by the IoT paradigm into data markets explicitly covered by some of the studies included in this SLR.

Challenge	Studies	Description	Impact on privacy
Heterogeneity and Interoperability	[S36] [S39]	The IoT consists of billions of IoT devices from different manufacturers, running different software on different local networks and geographic regions, with different computation power and storage capacity [129]. Furthermore, different communications standards, connectivity and availability aggravate the interplay of IoT devices.	An IoT data market should be agnostic to these differences and minimize any additional requirements; however, it is unclear how global data markets should harmonize data coming from different jurisdictions with different privacy regulations and how an IoT device can interact with another whose, e.g., verification schemata are inadequate from its perspective. In addition to these obstacles, a lack of interoperability may restrain PETs that involve the communication between many devices, e.g., SMC.
Computation power	[S5] [S11] [S48]	Manufacturers produce many IoT devices designed to consume low energy and require minimal volume, limiting these IoT devices to the core functionalities of monitoring and communication [S11].	Any additional computation requires a higher investment in resources and manufacturing, and running some PETs becomes infeasible without this extra investment. Consequently, a set of PETs is excluded without more computation power, e.g., cryptography-based PETs such as HE, SMC, ZKP, some digital signatures, or consensus algorithms. This limitation, however, may only apply to contexts where it might not be possible to connect IoT devices acting as clients with proprietary or trusted third-party nodes where these PETs are executed.
Storage capacity and real-time communication	[S1] [S5] [S11] [S17] [S29] [S48]	Minimizing the physical volume of an IoT device reduces their price but limits their storage capacity, forcing IoT devices to transmit the data to a data warehouse or a data market as quickly as possible. This tendency intensifies in some IoT applications where the time delay tolerance is low to enhance the utility of real-time information. [S17]	Processing time constrains the number of usable PETs, excluding those that require long execution times, such as fully homomorphic encryption or creating a zero-knowledge proof.
Data quality	[S14]	An unreliable IoT design may afflict thousands of IoT devices mass-produced by a manufacturer, which at deployment may lead to millions of unreliable data points. Furthermore, networks may also be unreliable, further worsening the quality [S14].	The impact may seem beneficial in terms of privacy; however, unreliable data leads to verification and secure computation schemata to fail and anonymization technologies to over-perturb the data as the underlying data is not entirely truthful.
Ambiguous data ownership	[S4] [S10] [S14]	When purchasing a device or a cluster of IoT devices, e.g., a phone, consumers also expect to own the data they are generating. However, the phone manufacturer and service providers expect to receive parts of this data nowadays with meager consent. In addition to this clash of interests, there are scholars that ponder whether data belongs to anyone in the first place, like L. Determann [130].	Having unclear data ownership leads to a misguided deployment of PETs, which may cause detrimental consequences if the privacy measures fall short. On the other hand, if the practitioner knows who has the right to the data and what the owner is reticent to share with a third party, then selected PETs and their privacy tuning can be optimized accordingly.
Privacy disparity	[S4] [S40]	Depending on the IoT devices' deployment location, the degree of privacy measures should be higher or lower, e.g., sensors in vehicles, smart homes, phones, and wearables. Furthermore, IoT deployments should adapt the monitoring time to an adequate amount depending on the context [S40].	Some PETs, such as semantic and syntactic technologies, allow adjusting the degree of privacy; however, others are more rigid. Selecting and adapting a PET to the IoT devices' deployment context requires expertise.
Pricing	[S4] [S14] [S17] [S50]	There are multiple variables imposing the price of data aside from supply and demand: the truthfulness, the source, either purchasing the data or the access [S17], and the privacy level. These factors add additional complexity to pricing, e.g., the sources have become disparate with the IoT, which drives pricing to a more granular task than before, when aggregated data could be sold as a unit [S17].	Aside from payment enforcement mechanisms, pricing involves negotiations, which frequently must ensure privacy. This adds an extra layer of complexity to the deployment of PETs. Furthermore, as data markets trade with more granular data points, PETs that need aggregation might be excluded in some contexts, e.g., syntactic technologies such as k-anonymity.

of their citizens (see Section 2.1). These laws also refer to an individual's and businesses' right to exploit their data commercially, which provides leeway for data markets [S28] and aims to uncover the untapped potential of data for innovations.

Nonetheless, research points out the sometimes unrealistic expectation to monitor the entirety of the Internet for privacy violations [S45], the dexterity of hackers to find novel deception methods [S3], and that laws are more reactive than preventative. Well-known networks of illegal proprietary digital asset exchanges, e.g., scientific works, and how users of digital services give away data tacitly provide testimony of the failure of data-related legal measures today, and the problems will likely increase with the accruing number of IoT devices [S38]. Moreover, privacy regulation can strangle free markets and innovations if they are too stringent [S45].

Aligned with these deficiencies, J. Henrik et al. [S38] introduced privacy regulation pitfalls that the IoT unfolds in data markets in 2013. They note that (i) definitions of personally identifiable information will be deprecated as unprecedented amounts of data can be aggregated, easing re-identification, (ii) the development and audit of PETs is costly, which may limit business models and potentially make disregarding privacy regulation profitable [135], (iii) privacy violations result on small fines or remain unpunished, (iv) technology tends to outpace regulation, and (v) the ubiquity of IoT devices will yield more illegal secondary personal data markets. After almost a decade of further research, (i) seems to be valid, at least in some scenarios. The ambiguity of privacy regulation is a barrier in some cases, as practitioners may default to weaker forms of privacy if their architecture appears to comply, leading to a more likely re-identification [127]; an attack that is more probably with the increasing number of IoT devices. However, in defense of these practitioners, while PETs have improved since 2013, some PETs that offer better privacy enhancements are still complex and not yet performant in 2021.

With the prior arguments, pitfall (ii) seems to hold; however, (iii) is no longer a strong pitfall. Since the enforcement of GDPR [120] in 2018, GDPR has punished multiple corporations with considerable fines ranging between €20 million and up to 4 % of a corporation's annual worldwide turnover of the preceding financial year. As of the writing of this publication, GDPR has harvested considerable fines assigned to Google in France on two occasions [136], Amazon [137], H&M [138] or the telecommunications operator TIM [139]. These fines alone accumulate to €282 million. These statistics are a sign that PETs are not appropriately introduced in production applications even by *big* technology companies and that not complying with privacy regulation in an IoT data market has dire economic consequences. While these fines could indicate how profitable it still is to violate privacy regulation (iii), one can no longer strongly defend (iii). Pitfall (iv) seems to materialize as long as the nature of law-making does not change. Lastly, pitfall (v) is concerning, given the existence of legal personal data markets that store up to 750 million user-profiles and trade 75 million online auctions daily like BlueKai [24], whose data could leak to the increasing number of illegal *shadow markets* [S6].

7. Discussion

This Section presents a set of key findings distilled from the two research questions answered in Sections 6.1 and 6.2, the content and metadata of the 50 publications included in our SLR, and other seminal studies that we encountered throughout our SLR but which do not necessarily address IoT data markets directly.

The attention of scientists towards privacy-enhancing technologies in the field of data markets for IoT devices has increased notably in recent years. The selected publications are modern, as 49 of the 50 studies were published between 2012 and 2020, and 34 of them (68 %) were published either in 2018, 2019, or during the first half of 2020. While the absolute number of publications in 2020 is lower than in 2019 because we captured only the first seven months of 2020, Fig. 9 illustrates the arguably accelerating trend of the cumulative curve of publications in the field of privacy-enhancing IoT data markets.

The most frequent research type (design and creation) and least common research contribution (lessons learned) suggest that privacy-oriented IoT data markets are still maturing and have not faced many production-grade implementations yet. According to Fig. G.13, around 76 % of the publications use a *design and creation* research approach, while only 4 % perform a case study. A further indication of field novelty is that only one out of the 50 publications had the contribution type *lessons learned* [S40]. Furthermore, while 35 studies (70 %) were of research type *solution proposal*, to the best of the author's knowledge, only one solution appears to have an implemented system that is applied in production [S25].

The selected studies rarely leverage existing libraries that provide PETs and often only build upon architectures developed in previous work to a small degree. Therefore, to gain more practical relevance, it may be beneficial for researchers to improve and extend existing work instead of reinventing the wheel. The research community and industry have developed many open-source libraries to employ zero-knowledge proofs, homomorphic encryption, secure multiparty computation, or differential privacy (see Section 6.1); however, none of the studies have indicated their use. Furthermore, studies often do not build upon each other, leading to further overlap. For example, W. Gao et al. [S41] and Z. Chen et al. [S34] both showcase an auction that obscures the bids by employing partially homomorphic encryption. However, W. Gao et al. only refers to the work from Z. Chen et al. in one line, noting that “[...] *there is only few literatures on designing privacy-preserving schemes in data market auctions.*” Moreover, [S42] builds upon [S27], and [S2] upon [S25], but each of these two sets belongs to the same group of researchers. In conclusion, it may be beneficial for researchers to incorporate building blocks from previous data market architectures to advance privacy-oriented research.

Moreover, many studies included in Tables D.7 and D.8 aim to create an IoT data marketplace employing distributed ledger technology (DLT). However, there seems not to be a consensus about which DLT to use for IoT data markets, as the authors build upon Ethereum, IOTA, Hyperledger Iroha, Fabric, Agora,

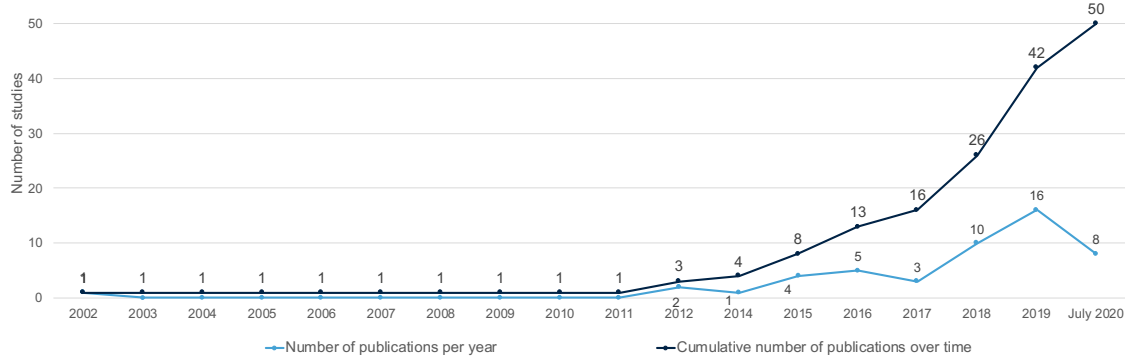


Figure 9: Publications in the field of privacy-enhancing data markets for the IoT from January 2002 to July 2020.

or Quorum, among others. Specifically, as an example, [S32] uses Ethereum smart contracts for payments while [S28] only uses these contracts for whitelisting and employs IOTA for payments instead.

The content of the selected studies can be categorized into two main orthogonal research streams within the context of privacy-enhancing IoT data markets: architectures and data trading schemata. The first research stream is dedicated to the design of privacy-enhancing architectures for the exchange of data in IoT data markets (25 studies, 50 %), and the second one focuses on the design of privacy-enhancing data trading such as auctions (12 studies, 24 %). The remaining studies can be associated with domains like legal [S6], user preferences [S8], or IoT data market challenges [S14][S19]. The selected studies, and also international initiatives such as the European GAIA-X [5], hence envision data markets beyond matchmaking and auction capabilities. Specifically, the studies that we analyzed structure the software, hardware, abstract entities, and their coordination, data processing, storage, communication, and the offered services to build a holistic or part of a privacy-enhancing IoT data market that includes PETs to tackle some of the challenges described in Section 6.2.

Despite the acknowledged need for combining of anonymization and secure and outsourced computation techniques, none of the researchers behind the 12 studies proposing data trading schemata, and only two publications out of the 25 designing data market architectures employ both PET categories in combination. Although PETs such as homomorphic encryption (HE) or secure multiparty computation conceal inputs and computation, the outputs can leak information about the underlying data and hence may be exposed to re-identification attacks [10]. Combining secure and outsourced computation techniques with anonymization-based PETs like differential privacy (DP) can help to make the outputs less sensitive. Moreover, leveraging only anonymization PETs does not sufficiently address the copy problem.

Within the 12 selected studies focused on data trading, DP is the most frequently used PET in auctions to enhance the privacy of the exchanged data. [S16][S22][S24][S37] employ DP in various forms to set the privacy levels and, subsequently, the price of the traded IoT data. Researchers might choose

DP over other anonymization technologies because DP is the only PET with a mathematical guarantee of privacy [13]. At the same time, partially HE (PHE) is the PET of choice to enhance the privacy of the bidding process. A group of authors [S26][S34][S41] chose PHE primarily for hiding the bids, confidentially computing the winner, and only revealing the output to the auction’s winner. Researchers might decide to use PHE over other forms of HE, secure multiparty computation, or trusted execution environments (TEEs) despite PHE’s significantly less general scope because PHE has relatively high performance and is conceptually simple.

Together, DP and PHE can enhance the privacy of auctions holistically, which is a contribution we have not found in this review. S. Sharma et al. [S48] emphasize that some HE schemata, such as Paillier’s, must complement other methods to guarantee more protection. Even more, while HE protects the input and the computation itself, if the intended recipients of the decrypted output are malicious, they may reverse engineer the output to learn properties about the input. An additional modification employing, e.g., differential privacy, of inputs or in the decrypted outputs before sharing may help prevent this attack in exchange for accuracy and thus utility. The same argument applies to other methods of secure and outsourced computation when being used in isolated form. We consequently point to a lack of combination in the research stream of data market architectures, except for two publications from the same group of researchers [S2][S25], which use TEEs to train machine learning models with DP.

The selected studies employ three dimensions to characterize data markets that entail privacy concerns: (i) the degree of centralization, (ii) the types and number of data domains, and (iii) the types of sellers and consumers. Each of these dimensions, as for example characterized by [S1][S27][S46] respectively, brings privacy concerns. (i) Data may be stored by the seller, the platform provider, or a decentralized platform using, e.g., a combination of commercial cloud storage, interplanetary file systems, or blockchains. Depending on the degree of decentralization and replication, practitioners need to consider different risks of leakage. In particular, if the architecture relies on a blockchain, PETs are particularly important [117]. (ii) An increase in the number and types of data domains opens up additional attack vectors and more possibilities for malicious en-

tities to link an individual’s data across databases. This hyper-connectivity between datasets can render the definitions of de-identified data, such as HIPAA’s, obsolete and suggests that privacy enhancements in the data economy should be defined globally and not locally. (iii) The degree of privacy enhancement should depend on the type of seller and consumer, e.g., consumers may expect higher privacy guarantees when a health insurance company gathers their data than when the collector is a renowned health research institution.

Based on our classifications in Section 6 and inspired by a set of seminal selected studies, we have created a reference model for the design of IoT data markets in Fig. 10, and detailed in Table 2. Most of the studies included in this SLR proposed solutions without following a reference model, except for D. López and B. Farooq [S3] and C. Niu et al. [S7], who developed their own without a systematic research. C. Niu et al. [S7] condense their architecture into two layers: data acquisition and trading. On the other hand, D. López and B. Farooq [S3] present a more holistic view of privacy-enhancing IoT data markets with six layers (*identification, privacy, contractual, communication, consensus, and incentive*) inspired by the Open System Interconnection model and heavily conditioned by the use of blockchain technology. This model, however, lacks essential steps of an IoT data market that several publications in our SLR focused on, namely storage [S1][S12][S25][S28] and processing [S7][S19][S20][S45]. Furthermore, the *identification layer* [S3] can be regarded as a subset of verification, which also includes data verification. Other studies base their market design on the type of participants [S15][S24][S27][S33][S43], e.g., sellers, aggregators, brokers, among others, and the type of data domain [S46], e.g., health, financial, or a combination. However, these categories cannot be transferred to other contexts as easily as a reference model that is agnostic to entity and data domain types.

Our reference model hence combines and generalizes some of the layers from [S3] and [S7] and complements them with additional layers such as the data auction, storage, verification, processing, and sovereignty layer (see Fig. 10). Most of these layers need multiple PETs, as there is no “one-size-fits-all” technology to enhance privacy. To navigate these layers in detail, refer to Table 2 and Fig. C.11. Furthermore, we distinguish between a contractual and sovereignty-related design to separate formal agreements from privacy and ownership policies. Furthermore, given the distinct purpose and implementation that auction schemata play in a data market, they should be respected by a unique IoT data market layer (auction dedicated studies: [S16][S22][S24][S37], among others). Lastly, incentives are necessary to encourage behavior that preserves the pre-defined qualities of the IoT data market, e.g., optimized prices [S3][S16], data authenticity [S16][S23], or maintaining the infrastructure like a permissionless DLT.

Aside from the ubiquity of digital signatures in IT system, in this SLR, Distributed ledger technology (DLT) is most frequently employed as the backbone of IoT data market design (see Fig. G.14), despite the lack of consensus on its use and

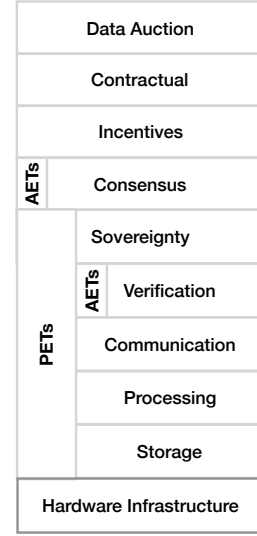


Figure 10: Reference model for the layers of a privacy-enhancing IoT data market.

DLT-based applications in production.. Although centralized systems seem more efficient and easier to deploy, and despite the seemingly few industrial applications that are running on blockchain today, many researchers in this SLR still advocate for distributed systems using DLT. Within the 35 solution proposals, around 31 % chose permissionless DLT, 14 % consortium DLTs, and the authors of the remaining 55 % either reviewed DLT, implemented a centralized solution or focused on designing a narrow feature. However, we noted that within the 45 % of DLT-based designs, many authors still relied on single entities for data processing or storage. Specifically, only one of the 50 studies [S25] has a public blockchain-based ecosystem in production, yet without a real-world use case running. These statistics indicate a lack of adoption despite substantial research efforts.

Furthermore, while blockchains enhance authenticity, assure integrity, and enable payments without the need for a trusted third party, blockchains are limited in storage capacity [S1], computation power [S25], and can exacerbate privacy issues because of their inherent data and computation replication [S6][119]. Consequently, almost all of the studies that include blockchain technology to support an IoT data market require PETs to protect users’ data and identities. These studies go as far as creating innovative privacy-enhancing blockchain architectures with other PETs as building blocks, e.g., trusted execution environments [S25][S45], or adding a privacy layer to their market design based on differential privacy [S3]. However, within the literature, there are also questionable statements such as “[...] researchers and technologists have found that blockchain can be a potential solution to the privacy problem by decentralizing information [...] Blockchain can be used to securely share private information [...]” [S3], “Blockchain-based approaches provide decentralized security and privacy [...]” [S11], or “Blockchain has been proven to possess security, immutability, and privacy properties, which has caused a lot of researchers to introduce it into the privacy and security con-

cerned IoT” [S23]. These statements, coupled with the current excitement around blockchain, can lead practitioners in the industry to wrongfully push blockchain for “privacy”. Therefore, the community would benefit from clear explanations of why authors employ blockchain and clearly state the need for other technologies to enhance privacy.

8. Conclusion

With this review, we reveal the landscape of PETs in the field of data markets for the IoT. To accomplish this task, we have conducted a systematic literature review (SLR) to identify and filter the studies aiming to solve this landscape’s challenges. Consecutively, we formulated terminology to dissect the selected studies’ architectures and findings and identified the PETs that related work uses and which specific challenges they address.

The benefits of data trading have become evident to institutions and businesses [33], but the increasing number of data breaches also demonstrates that the privacy threat is real [9]. Society’s progress thus partly depends on the former, but it is hindered by the latter. This conundrum has sparked the interest of researchers around the globe. The authors of the selected studies in this SLR have devised proposals for privacy-enhancing IoT data marketplaces to comply with privacy requirements while striving to maintain the utility, profitability, and fair and seamless exchange of data. Since this is a relatively new, multidisciplinary research field, the optimal combination of technologies and theoretical foundations employed in these proposals is still in the development phase. Therefore, no proposal seems to have established itself as canonical yet, and our review aims to provide a holistic overview of tools and challenges for further research to build on. We have distilled our research into a set of key findings and that data authenticity and privacy must be optimally balanced for data markets to flourish. However, the field of PETs needs more maturity to impact data markets for the IoT positively in practice. Additionally, to realize the ownership economy, we encourage researchers to solve the copy problem and improve privacy-enhancing verification. We also discovered that research on privacy-oriented data markets could benefit from increased reuse of components from previous articles and existing open-source libraries and a more explicit description of critical objectives. For example, the benefits of utilizing distributed ledger technology (DLT) in data markets for IoT architectures often remain unclear, and authors do not sufficiently consider DLT’s lack of maturity and its inherent privacy challenges.

Additionally, the IoT’s particular characteristics bring a new set of challenges for privacy enhancement, most notably, the consequences of a lack of interoperability, computation and storage constraints, and the privacy disparity across jurisdictions. Furthermore, institutions that incorporate PETs early may have a competitive advantage [4], as proficiently analyzing privacy-enhanced data might become more important than having the most data. We have also observed the importance of first determining the sovereignty layer in data market design because the ownership and management rules for the participants’ data impact the PETs selection of the rest of the layers. We also

must underline that there is no “one-size-fits-all” PET; only a combination may tackle the various privacy challenges facing data markets for the IoT. Lastly, we recommend that institutions invest resources in the research and adoption of PETs to remain competitive in the advent of a more privacy-enhancing IoT.

Limitations. Even though we have adopted a rigorous research design and paid particular attention to the selection and analysis of published studies, SLRs have limitations that may have undermined its effectiveness. These threats include (i) incompleteness of study search, (ii) bias on study selection, and (iii) inaccuracy of data extraction.

(i) Some relevant publications might be absent. To mitigate this limitation, we searched in several highly reputed digital libraries, performed a preliminary search to determine suitable search strings, conducted a backward search to identify additional related work, and included studies in advance that met the standards and filters of this SLR. These measures reduce the probability of missing relevant publications. (ii) The experience and knowledge of the researchers may drive the study selection with an inherent bias. Nonetheless, following Kitchenham [44], we aimed to create a set of explicit inclusion and exclusion criteria to maximize the degree of objectivity. To mitigate different appreciations of these criteria, we carried out a preliminary search to ensure researchers have a consistent understanding of the requirements. Furthermore, two researchers conducted the selection process independently and resolved the conflicts between their decisions interactively. (iii) There might be a bias in selecting the extracted data, which may affect the classification results of the selected studies. To mitigate this potential limitation, the two researchers specified a set of data extraction cards (see Section 3.2) to eliminate any misalignment in the data extraction process results.

Future work. The opportunities and need for future work highlighted by the selected studies resonate with the challenges covered in Section 6.2. Most notably, there is a need to solve the copy problem [10][S4][S17] and to lessen IoT devices’ limitations regarding computation [S5][S11], storage and capacity [S29][S48] to tackle or circumvent the constraints PETs may induce. Moreover, to decrease the probability of re-identification attacks, further work is needed to advance the maturity of PETs and combine them, e.g., bringing together differential privacy and secure and outsourced computation efficiently. Additional research is also necessary to create standards for data markets such as a language to describe privacy requirements, universal APIs to interact between different IoT devices with various degrees and techniques for privacy protection, and machine-readable definitions of privacy, e.g., using ontologies [S40].

If society considers privacy a necessity, it should be enhanced by default and optimally in any system without attaching price tags to one’s privacy, as some of the selected studies pursued suggest [S16][S15][S22]. We find this posture a

Table 2: A mapping of privacy- and authenticity-enhancing technologies (PET and AET) to the narrow challenges using the terminology defined for this review. The extent of enhancement of privacy, utility, and characteristics of the different PETs and AETs varies from significantly increasing ++, over +, +-, - to significantly decreasing --. na denotes *not applicable*. w/ denotes *with*. *Considering a digital certificate when using digital signatures, if applicable. The privacy column assumes data and identity are authentic.

Layer	Technology category	Technology	Narrow challenges						
			Privacy-utility tradeoff					Recursive enforcement problem	Copy problem [Bundling problem (BP)]
			Privacy [Confidentiality (Conf)]			Utility			
			Input [Data (D)] [Identity (I)]	Computation	Output	Authenticity [Data (D)] [Identity (I)]	Integrity [Data (D)] [Computation (C)]		
PET									
Processing	Secure and outsourced computation (SOC)	Zero knowledge proofs for computational integrity	++ D	++	++	++ D	++ D, C	Circumvents	Tackles (BP)
		ZKPs of anonymous credentials	++ I	++	++	++ I	++ D, C	Circumvents	Tackles (BP)
		Trusted execution environments	++ D	++	+ Conf	na	++ D, C	Circumvents	Tackles
		Partially homomorphic encryption	++ D	++	+ Conf	na	++ D, C	Circumvents	Tackles
		Fully homomorphic encryption	++ D	++	+ Conf	na	++ D, C	Circumvents	Tackles
		Secure multiparty computation	++ D	++	+ Conf	na	++ D, C	Circumvents	Tackles
		Privacy-preserving data mining	W/ SOC (w/ AN) Federated (w/ AN)	++ D (+ D)	++ (na)	+ Conf (+)	na (+- D) na (+- D)	++ D, C (na) na (na)	Circumvents
	Anonymization (AN)	Differential privacy (DP)	+ D	na	+	+ D	na	na	Tackles (BP)
		K-anonymity	+ D	na	+ -	- D	na	na	Tackles (BP)
		Perturbative	+ D	na	+ -	- D	na	na	Tackles (BP)
		Pseudonym creation	+ I	na	+ -	- I	na	na	Tackles (BP)
	Storage	Encryption		Storage layer: ++ D, Conf Communication layer: ++ D, Conf			na	++ D	Circumvents
Communication	Onion routing		++ I D, Conf	na	na	na	++ D	Tackles	na
Verification	Hashing		++ D, Conf	na	na	na	++ D	Circumvents	na
	Ring digital signatures		+ I	na	na	- I	++ D	na	Tackles (BP)
	Blind digital signatures		+ I D	na	na	+ I na D	++ D	na	Tackles (BP)
Sovereignty	Smart contracts (for privacy policies)		Characteristics and challenges are pegged to distributed ledger technology.						
	Privacy policies (Access control)		Characteristics and challenges are pegged to the selected PETs and AETs employed to fulfil the privacy requirements.						
AET									
Consensus	Distributed ledger technology (permissioned)		- I D, Conf	+ Conf	+ Conf	+ I na D	+ D, C	Tackles	na
	Distributed ledger technology (permissionless)		+ I D	-	-	+ I na D	++ D, C	Tackles	na
Verification	Truth discovery		- I D	-	-	+ I + D	++ D, C	Tackles	na
	Decentralized identifiers		+ I	na	na	++ I	na	Tackles	na
	Digital fingerprint (Version control)		+ I D	na	na	+ I + D	++ D	na	na
	Identity-based digital signatures		+ I	na	na	+ I	++ D	na	na
	Certificateless digital signatures		+ I	na	na	+ I	++ D	Tackles	na

worthy research endeavor and encourage researchers to ponder whether monetizing privacy in a competitive market would ultimately benefit society. Furthermore, legal practitioners have ample ground to develop legislation specifically around privacy in IoT data markets and for economists to delve into data pricing and decentralized market interactions. Legal researchers could investigate how stringent privacy regulations should be,

as heavy regulation may strangle free markets and innovations [S45]. Additionally, the legal, pricing, and privacy aspects hinge around data sovereignty. As long as ownership is ambiguous, researchers' efforts will struggle to have maximum impact. Furthermore, the relevance of our results may reach beyond IoT data markets, as the analysis of PETs and derived insights, e.g., how IoT impact privacy, can permeate other research ar-

eas such as privacy-by-design software engineering, policy-making, data governance, politics, and economics. Moreover, most PETs have specific performance-, complexity-, or utility-related shortcomings (which we describe in Section 6.1) that researchers can address.

Lastly, we recommend researchers derive decision trees based on Table 2 to enhance the decision-making of privacy officers beyond our work. Moreover, so far we could not find any formulation of an information-theoretic quantification of the data leaked from a data market. We also encourage social scientists to focus on questions related to data sovereignty. To realize a vision of data markets that benefit society, researchers will need to concentrate on roadblocks such as the copy problem. Finally, institutions should consider updating their privacy-enhancing processes to participate in IoT data markets effectively.

Acknowledgements

We would like to sincerely thank The BMW Group for making this publication possible. We also wish to thank the Bayerisches Forschungsinstitut für Digitale Transformation for supporting our research on differential privacy.

Selected Studies

- [S1] Y. N. Li, X. Feng, J. Xie, H. Feng, Z. Guan, Q. Wu, A decentralized and secure blockchain platform for open fair data trading, *Concurrency Computation* 32 (7) (2019) 1–11. doi:10.1002/cpe.5578.
- [S2] N. Hynes, D. Dao, D. Yan, R. Cheng, D. Song, A demonstration of sterling: A privacy-preserving data marketplace, *Proceedings of the VLDB Endowment* 11 (12) (2018) 2086–2089. doi:10.14778/3229863.3236266.
- [S3] D. López, B. Farooq, A multi-layered blockchain framework for smart mobility data-markets, *Transportation Research Part C: Emerging Technologies* 111 (June 2019) (2020) 588–615. doi:10.1016/j.trc.2020.01.002.
- [S4] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, W. Zhao, A Survey on Big Data Market: Pricing, Trading and Protection, *IEEE Access* 6 (May) (2018) 15132–15154. doi:10.1109/ACCESS.2018.2806881.
- [S5] D. Bogdanov, R. Jagomägis, S. Laur, A Universal Toolkit for Cryptographically Secure Privacy-Preserving Data Mining, *LNCS 7299 - Intelligence and Security Informatics 7299* (2012). URL <https://link.springer.com/content/pdf/10.1007/978-3-642-30428-6.pdf>
- [S6] S. Spiekermann, A. Novotny, A vision for global privacy bridges: Technical and legal measures for international data markets, *Computer Law and Security Review* 31 (2) (2015) 181–200. doi:10.1016/j.clsr.2015.01.009. URL <http://dx.doi.org/10.1016/j.clsr.2015.01.009>
- [S7] C. Niu, Z. Zheng, F. Wu, X. Gao, G. Chen, Achieving Data Truthfulness and Privacy Preservation in Data Markets, *IEEE Transactions on Knowledge and Data Engineering* 31 (1) (2019) 105–119. arXiv:1812.03280, doi:10.1109/TKDE.2018.2822727.
- [S8] E. M. Schomakers, C. Lidynia, M. Zieffle, All of me? Users' preferences for privacy-preserving data markets and the importance of anonymity, *Electronic Markets* (2020). doi:10.1007/s12525-020-00404-9.
- [S9] Y. Li, C. Miao, L. Su, J. Gao, Q. Li, B. Ding, Z. Qin, K. Ren, An efficient two-layer mechanism for privacy-preserving truth discovery, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2018) 1705–1714. doi:10.1145/3219819.3219998.
- [S10] L. Zhou, L. Wang, T. Ai, Y. Sun, BeeKeeper 2.0: Confidential blockchain-enabled IoT system with fully homomorphic computation, *Sensors (Switzerland)* 18 (11) (2018). doi:10.3390/s18113785.
- [S11] A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram, Blockchain for IoT security and privacy: The case study of a smart home, *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017* (2017) 618–623. doi:10.1109/PERCOMW.2017.7917634.
- [S12] R. Li, T. Song, B. Mei, H. Li, X. Cheng, L. Sun, Blockchain for Large-Scale Internet of Things Data Storage and Protection, *IEEE Transactions on Services Computing* 12 (5) (2019) 762–771. doi:10.1109/TSC.2018.2853167.
- [S13] J. Wei, M. Sabonuchi, R. Roche, Blockchain-enabled peer-to-peer data trading mechanism, *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)* 1349–1354. doi:10.1109/Cybermat.ics.
- [S14] Z. Zheng, W. Mao, F. Wu, G. Chen, Challenges and opportunities in IoT data markets, *SocialSense 2019 - Proceedings of the 2019 4th International Workshop on Social Sensing* (2019) 1–2. doi:10.1145/3313294.3313378.
- [S15] M. M. Khalili, X. Zhang, M. Liu, Contract design for purchasing private data using a biased differentially private algorithm, *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.
- [S16] L. Yang, M. Zhang, S. He, M. Li, J. Zhang, Crowd-empowered privacy-preserving data aggregation for mobile crowdsensing, *Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)* (2018) 151–160. doi:10.1145/3209582.3209598.
- [S17] K. Mišura, M. Žagar, Data marketplace for Internet of Things, *Proceedings of 2016 International Conference on Smart Systems and Technologies, SST 2016* (2016) 255–260. doi:10.1109/SST.2016.7765669.
- [S18] X. Zheng, Data trading with differential privacy in data market, *ACM International Conference Proceeding Series* (8) (2020) 112–115. doi:10.1145/3379247.3379271.
- [S19] J. Pennekamp, M. Henze, S. Schmidt, P. Niemietz, M. Fey, D. Trauth, T. Berge, C. Brecher, K. Wehrle, Dataflow Challenges in an Internet of Production, in: *ACM Workshop on Cyber-Physical Systems Security & Privacy (CPS-SPC'19)*, November 11, 2019, London, United Kingdom. ACM, 2019, pp. 27–38. doi:10.1145/3338499.3357357.
- [S20] Z. J. Wang, C. H. V. Lin, Y. H. Yuan, C. C. J. Huang, Decentralized Data Marketplace to Enable Trusted Machine Economy, *2019 IEEE Eurasia Conference on IOT, Communication and Engineering, ECICE 2019* (2019) 246–250. doi:10.1109/ECICE47484.2019.8942729.
- [S21] M. Guerriero, D. A. Tamburri, E. Di Nitto, Defining, enforcing and checking privacy policies in data-intensive applications, *Proceedings - International Conference on Software Engineering* (2018) 172–182. doi:10.1145/3194133.3194140.
- [S22] M. Shi, Y. Qiao, X. Wang, Differentially private auctions for private data crowdsourcing, *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) 1–8. doi:10.1109/ISPA-BDCLOUD-SustainCom-SocialCom48970.2019.00013.
- [S23] J. Du, C. Jiang, E. Gelenbe, L. Xu, J. Li, Y. Ren, Distributed Data Privacy Preservation in IoT Applications, *IEEE Wireless Communications* 25 (December) (2018) 68–76. doi:10.1109/MWC.2017.1800094.
- [S24] G. Gao, M. Xiao, J. Wu, S. Zhang, L. Huang, G. Xiao, DPDT: A Differentially Private Crowd-Sensed Data Trading Mechanism, *IEEE Internet of Things Journal* 7 (1) (2020) 751–762. doi:10.1109/JIOT.2019.2944107.
- [S25] R. Cheng, F. Zhang, J. Kos, W. He, N. Hynes, N. Johnson, A. Juels, A. Miller, D. Song, Ekiden: A platform for confidentiality-preserving, trustworthy, and performant smart contracts, *Proceedings - 4th IEEE European Symposium on Security and Privacy, EURO S and P 2019* (2019) 185–200. doi:10.1109/EuroSP.2019.00023.
- [S26] T. Jung, X. Y. Li, Enabling privacy-preserving auctions in big data, *Proceedings - IEEE INFOCOM 2015-Augus (BigSecurity)* (2015) 173–178. arXiv:1308.6202, doi:10.1109/INFCOMW.2015.7179380.
- [S27] C. Perera, R. Ranjan, L. Wang, End-to-end privacy for open big data markets, *IEEE Cloud Computing* 2 (4) (2015) 44–53. doi:10.1109/MC.2015.78.
- [S28] M. Zichichi, M. Contu, S. Ferretti, V. Rodríguez-Doncel, Ensuring personal data anonymity in data marketplaces through sensing-as-a-service and distributed ledger technologies, *CEUR Workshop Proceedings* 2580 (2020). URL https://www.researchgate.net/publication/340183476_Ensuring_Personal_Data_Anonymity_in_Data_Marketplaces_through_Sensing-as-a-Service_and_Distributed_Ledger
- [S29] S. Duri, M. Gruteser, X. Liu, P. Moskowitz, R. Perez, M. Singh, J. M. Tang, Framework for security and privacy in automotive telematics, *Proceedings of the ACM International Workshop on Mobile Commerce* (2002) 25–32. doi:10.1145/570709.570711.
- [S30] P. Tzianos, G. Pipelidis, N. Tsiamitros, Hermes: An open and transparent marketplace for IoT sensor data over distributed ledgers, *ICBC 2019 - IEEE International Conference on Blockchain and Cryptocurrency* (2019) 167–170. doi:10.1109/BLCC.2019.8751331.
- [S31] K. Li, L. Tian, W. Li, G. Luo, Z. Cai, Incorporating social interaction into three-party game towards privacy protection in IoT, *Computer Networks* 150 (2019) 90–101. doi:10.1016/j.comnet.2018.11.036.
- [S32] Y. Zhao, Y. Yu, Y. Li, G. Han, X. Du, Machine learning based privacy-preserving fair data trading in big data market, *Information Sciences* 478 (2019) 449–460. doi:10.1016/j.ins.2018.11.028.
- [S33] S. Kiyomoto, M. S. Rahman, A. Basu, On Blockchain-Based Anonymized Dataset Distribution Platform, *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)* (2017) 85–92. URL <https://ieeexplore.ieee.org/document/7965711>
- [S34] Z. Chen, L. Chen, L. Huang, H. Zhong, On Privacy-Preserving Cloud Auction, *Proceedings of the IEEE Symposium on Reliable Distributed Systems* (2016) 279–288. doi:10.1109/SRDS.2016.045.

- [S35] L. Pournajaf, D. A. Garcia-Ulloa, L. Xiong, V. Sunderam, Participant Privacy in Mobile Crowd Sensing Task Management, *ACM SIGMOD Record* 44 (4) (2016) 23–34. doi:10.1145/2935694.2935700.
- [S36] D. Sánchez, A. Viejo, Personalized privacy in open data sharing scenarios, *Online Information Review* 41 (3) (2017) 298–310. doi:10.1108/OIR-01-2016-0011.
- [S37] K. Jung, S. Park, Privacy Bargaining with Fairness: Privacy-Price Negotiation System for Applying Differential Privacy in Data Market Environments, 2019 IEEE International Conference on Big Data (2019) 1389–1394doi:10.1109/BigData47090.2019.9006101.
- [S38] J. H. Ziegeldorf, O. G. Morchon, K. Wehrle, Privacy in the internet of things: Threats and challenges, *Security and Communication Networks* 7 (12) (2014) 2728–2742. doi:10.1002/sec.795.
- [S39] C. Perera, C. McCormick, A. K. Bandara, B. A. Price, B. Nuseibeh, Privacy-by-design framework for assessing internet of things applications and platforms, *ACM International Conference Proceeding Series* 07-09-Nove (2016) 83–92. doi:10.1145/2991561.2991566.
- [S40] C. Perera, C. Liu, R. Ranjan, L. Wang, A. Zomaya, Privacy-knowledge Modeling for the Internet of things: A look back, *Computer* 49 (12) (2016) 60–68. doi:10.1109/MC.2016.366.
- [S41] W. Gao, W. Yu, F. Liang, W. G. Hatcher, C. Lu, Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption, *IEEE Transactions on Network Science and Engineering* 7 (2) (2020) 776–791. doi:10.1109/TNSE.2018.2846736.
- [S42] S. Park, K. Park, J. Lee, K. Jung, PRIVATA: Differentially private Data market framework using Negotiation-based Pricing mechanism, *Proceedings of ACM CIKM conference (CIKM'19)*, November 3–7, 2019, Beijing, China. (2019) 156–157doi:10.1007/978-3-663-10915-0_47.
- [S43] V. Koutsos, D. Papadopoulos, D. Chatzopoulos, S. Tarkoma, P. Hui, *Agora: A Privacy-Aware Data Marketplace* (2020) 13. URL <https://eprint.iacr.org/2020/865.pdf>
- [S44] J. Cao, P. Karras, Publishing microdata with a robust privacy guarantee, *Proceedings of the VLDB Endowment* 5 (11) (2012) 1388–1399. arXiv:1208.0220, doi:10.14778/2350229.2350255.
- [S45] W. Dai, C. Dai, K. K. R. Choo, C. Cui, D. Zou, H. Jin, SDTE: A Secure Blockchain-Based Data Trading Ecosystem, *IEEE Transactions on Information Forensics and Security* 15 (2020) 725–737. doi:10.1109/TIFS.2019.2928256.
- [S46] Z. Guan, X. Shao, Z. Wan, Secure, Fair and Efficient Data Trading without Third Party Using Blockchain, 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) 1349–1354doi:10.1109/Cybermatics.
- [S47] M. A. Alsheikh, Y. Jiao, D. Niyato, P. Wang, D. Leong, Z. Han, The Accuracy-Privacy Trade-off of Mobile Crowdsensing, *IEEE Communications Magazine* 55 (6) (2017) 132–139. arXiv:1702.04565, doi:10.1109/MCOM.2017.1600737.
- [S48] S. Sharma, K. Chen, A. Sheth, Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems, *IEEE Internet Computing* 22 (2) (2018) 42–51. doi:10.1109/MIC.2018.112102519.
- [S49] A. Colman, M. J. M. Chowdhury, M. Baruwat Chhetri, Towards a trusted marketplace for wearable data, *Proceedings - 2019 IEEE 5th International Conference on Collaboration and Internet Computing, CIC 2019 (Cic)* (2019) 314–321. doi:10.1109/CIC48465.2019.00044.
- [S50] Z. Cai, Z. He, Trading private range counting over big IoT data, *Proceedings - International Conference on Distributed Computing Systems* 2019-July (2019) 144–153. doi:10.1109/ICDCS.2019.00023.

References

- [1] IDC, Open Evidence, European data market smart (feb 2017).
URL https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=44400
- [2] A. R. Miller, C. Tucker, Health Information Exchange, System Size and Information Silos (2013) 29.
URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1457719
- [3] F. Stahl, F. Schomm, G. Vossen, L. Vomfell, A classification framework for data marketplaces, *Vietnam Journal of Computer Science* 3 (3) (2016) 137–143. doi:10.1007/s40595-016-0064-2.
- [4] McKinsey & Company, Four ways to accelerate the creation of data ecosystems.
URL <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/four-ways-to-accelerate-the-creation-of-data-ecosystems>
- [5] G. Eggers, B. Fondermann, B. Maier, K. Ottradovetz, J. Pformmer, R. Reinhardt, H. Rollin, A. Schmieg, S. Steinbuß, P. Trinius, A. Weis, C. Weiss, S. Wilfling, GAIA-X: Technical Architecture.
URL https://www.data-infrastructure.eu/GAIA-X/Redaktion/EN/Publications/gaia-x-technical-architecture.pdf?__blob=publicationFile&v=5
- [6] L. Sweeney, A. Abu, J. Winn, Identifying Participants in the Personal Genome Project by Name, *SSRN Electronic Journal* (2013). doi:10.2139/ssrn.2257732.
- [7] X. Gao, B. Firner, S. Sugrim, V. Kaiser-Pendergrast, Y. Yang, J. Lindqvist, Elastic pathing: your speed is enough to track you, in: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, ACM Press, Seattle, Washington, 2014, pp. 975–986. doi:10.1145/2632048.2632077.
- [8] A. Sunyaev, N. Kannengießer, R. Beck, H. Treiblmaier, M. Lacity, J. Kranz, G. Fridgen, U. Spankowski, A. Luckow, Token economy, *Business & Information Systems Engineering* (2021).
URL <https://link.springer.com/article/10.1007/s12599-021-00684-1>
- [9] IBM Security and Ponemon Institute LLC, 2018 cost of a data breach study: Global overview (Jul 2018).
URL <https://www.intlxolutions.com/hubfs/2018-Global-Cost-of-a-Data-Breach-Report.pdf>
- [10] A. Trask, E. Bluemke, B. Garfinkel, C. G. Cuervas-Mons, A. Dafoe, Beyond privacy trade-offs with structured transparency (2020). arXiv:2012.08347.
URL https://www.researchgate.net/publication/347300876_Beyond_Privacy_Trade-offs_with_Structured_Transparency
- [11] R. Hes, J. J. Borking, Netherlands, I. a. P. Commissioner/Ontario (Eds.), *Privacy-enhancing technologies: the path to anonymity*, rev. ed Edition, no. 11 in *Achtergrondstudies en verkenningen*, Registratiekamer, The Hague, 1998.
URL https://www.researchgate.net/publication/243777645_Privacy-Enhancing_Technologies_The_Path_to_Anonymity
- [12] R. Oppliger, Privacy-enhancing technologies for the world wide web, *Computer Communications* 28 (16) (2005) 1791–1797. doi:10.1016/j.comcom.2005.02.003.
- [13] C. Dwork, A. Roth, The Algorithmic Foundations of Differential Privacy, *Foundations and Trends® in Theoretical Computer Science* 9 (3-4) (2013) 211–407. doi:10.1561/04000000042.
- [14] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, M. Naor, Our data, ourselves: Privacy via distributed noise generation, in: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 4004 LNCS, 2006, pp. 486–503. doi:10.1007/11761679-29.
- [15] P. Samarati, L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression 19.
URL <https://epic.org/privacy/reidentification/Samarati-Sweeney-paper.pdf>
- [16] M. A. Will, R. K. Ko, A guide to homomorphic encryption, Elsevier Inc., 2015. doi:10.1016/B978-0-12-801595-7.00005-7.
- [17] P. Chaudhary, R. Gupta, A. Singh, P. Majumder, Analysis and Comparison of Various Fully Homomorphic Encryption Techniques, 2019 International Conference on Computing, Power and Communication Technologies, GUCON 2019 (2019) 58–62.
URL <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8940577>
- [18] P. Paillier, Public-key cryptosystems based on composite degree residuosity classes, *Eurocrypt* (1999). doi:10.1007/3-540-48910-X_9.
- [19] OMTP, Advanced trusted environment: Omtp tr1 (May 2009).
URL <http://www.gsma.com/newsroom/wp-content/uploads/2012/03/omtpadvancedtrustedenvironmentomtptr1v11.pdf>
- [20] A. C. Yao, Protocols for secure computations, in: 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), 1982, pp. 160–164. doi:10.1109/SFCS.1982.38.
- [21] S. Goldwasser, S. Micali, C. Rackoff, The knowledge complexity of interactive proof systems, *SIAM J. Comput.* 18 (1) (1989) 186–208. doi:10.1137/0218012.
- [22] O. Goldreich, Y. Oren, Definitions and properties of zero-knowledge proof systems, *Journal of Cryptology* 7 (1) (1994) 1–32. doi:10.1007/BF00195207.
- [23] G. Bondel, G. M. Garrido, K. Baumer, F. Matthes, Towards a Privacy-Enhancing Tool Based on De-Identification Methods 8.
URL <https://aisel.aisnet.org/pacis2020/157/>
- [24] S. Spiekermann, R. Böhme, A. Acquisti, K.-L. Hui, Personal data markets, *Electronic Markets* 25 (2) (2015) 91–93. doi:10.1007/s12525-015-0190-1.
- [25] P. B. Anne Zöll, Christian M. Olt, Privacy-sensitive Business Models: Barriers of Organizational Adoption of Privacy-Enhancing Technologies (2021) 22.
URL https://aisel.aisnet.org/ecis2021_rp/34/
- [26] A. F. Westin, *Privacy and Freedom*, IG Publishing, New York, 1967.
URL <https://scholarlycommons.law.wlu.edu/wlulr/vol125/iss1/20/>
- [27] G. A. Fink, H. Song, S. Jeschke (Eds.), *Security and privacy in cyber-physical systems: Foundations, principles, and applications*, first edition Edition, Wiley IEEE Press, Hoboken, NJ, 2018.
URL <https://ieeexplore.ieee.org/servlet/opac?bknumber=8068866>
- [28] K. Renaud, D. Galvez-Cruz, Privacy: Aspects, definitions and a multifaceted privacy preservation approach, 2010, pp. 1 – 8. doi:10.1109/ISSA.2010.5588297.
- [29] D. J. Solove, The meaning and value of privacy, in: B. Roessler, D. Mokrosinska (Eds.), *Social Dimensions of Privacy*, Cambridge University Press, Cambridge, 2015, pp. 71–82. doi:10.1017/CB09781107280557.005.
- [30] F. T. Wu, Defining privacy and utility in data sets, 84 *University of Colorado Law Review* 1117 (2013); 2012 TRPC (2012) 1117–1177doi:10.2139/ssrn.2031808.
- [31] M. Deng, K. Wuyts, R. Scandariato, B. Preneel, W. Joosen, A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements, *Requirements Engineering* 16 (1) (2011) 3–32. doi:10.1007/s00766-010-0115-7.
- [32] R. Garratt, M. R. v. Oordt, Privacy as a public good: A case for electronic cash, *Journal of Political Economy* (2018). doi:10.1086/714133.
- [33] N. Kaaniche, M. Laurent, Attribute-based signatures for supporting anonymous certification, in: I. Askoxylakis, S. Ioannidis, S. Katsikas, C. Meadows (Eds.), *Computer Security – ESORICS 2016*, Springer International Publishing, Cham, 2016, pp. 279–300.
URL <https://www.semanticscholar.org/paper/Attribute-Based-Signatures-for-Supporting-Anonymous-Kaaniche-Laurent-Maknavicius/3b0624ff32b9258ca2351c894d320d83a546fcd6>
- [34] J. E. Campbell, M. Carlson, Panopticon.com: Online surveillance and the commodification of privacy, *Journal of Broadcasting & Electronic Media* 46 (4) (2002) 586–606. doi:10.1207/s15506878jobem4604_6.
- [35] A. Lichter, M. Löffler, S. Siegloch, The Long-Term Costs of Government Surveillance: Insights from Stasi Spying in East Germany, *Journal of the European Economic Association* 19 (2) (2020) 741–789. arXiv:https://academic.oup.com/jeea/article-pdf/19/2/741/37108669/jvaa009.pdf, doi:10.1093/jeea/jvaa009.

- URL <https://doi.org/10.1093/jeea/jvaa009>
- [36] S. Kokolakis, Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon, *Computers & Security* 64 (2017) 122–134. doi:<https://doi.org/10.1016/j.cose.2015.07.002>.
 - [37] J. Coppel, E-Commerce: Impacts and Policy Challenges, OECD Economics Department Working Papers 252, series: OECD Economics Department Working Papers Volume: 252 (Jun. 2000). doi:10.1787/801315684632.
 - [38] J. Kennedy, Big data's economic impact, [Online]. Available: <https://www.ced.org/blog/entry/big-datas-economic-impact>, [Accessed on 04 Jul. 2021].
 - [39] A. M. Oberländer, M. Röglinger, M. Rosemann, A. Kees, Conceptualizing business-to-thing interactions – a sociomaterial perspective on the internet of things, *European Journal of Information Systems* 27 (4) (2018) 486–502. doi:10.1080/0960085X.2017.1387714.
 - [40] I. Lee, K. Lee, The Internet of Things (IoT): Applications, investments, and challenges for enterprises, *Business Horizons* 58 (4) (2015) 431–440. doi:10.1016/j.bushor.2015.03.008.
 - [41] V. Basili, G. Caldiera, D. Rombach, The goal question metric approach, *Encyclopedia of Software Engineering* (1994) 528–532. URL <http://www.cs.toronto.edu/~sme/CSC444F/handouts/GQM-paper.pdf>
 - [42] B. A. Kitchenham, D. Budgen, Evidence-based software engineering and systematic reviews, Chapman and Hall/CRC, 2015. URL <https://dl.acm.org/doi/book/10.5555/2994449>
 - [43] B. Kitchenham, Procedures for Performing Systematic Reviews, Joint Technical Report (2004). doi:10.5144/0256-4947.2017.79.
 - [44] D. C. B. Mariano, C. Leite, L. H. S. Santos, R. E. O. Rocha, R. C. de Melo-Minardi, A guide to performing systematic literature reviews in bioinformatics (2017). arXiv:1707.05813.
 - [45] T. Dybå, T. Dingsøyr, G. Hanssen, Applying Systematic Reviews to Diverse Study Types: An Experience Report, *Proceedings - 1st International Symposium on Empirical Software Engineering and Measurement, ESEM 2007 (7465)* (2007) 126–135. doi:10.1109/ESEM.2007.59.
 - [46] O. Dieste, A. Grimán, N. Juristo, Developing search strategies for detecting relevant experiments, *Empirical Software Engineering* 14 (5) (2009) 513–539. doi:10.1007/s10664-008-9091-7.
 - [47] H. Zhang, M. A. Babar, P. Tell, Identifying relevant studies in software engineering, *Information and Software Technology* 53 (6) (2011) 625–637. doi:10.1016/j.infsof.2010.12.010.
 - [48] A. Kilgariff, V. Baisa, J. Bušta, M. Jakubček, V. Kovár, J. Michelfeit, P. Rychlý, V. Suchomel, The Sketch Engine: ten years on, *Lexicography* (2014). URL https://www.researchgate.net/publication/271848017_The_Sketch_Engine_Ten_Years_On
 - [49] O. P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, *Journal of Systems and Software* 80 (4) (2007) 571–583. doi:10.1016/j.jss.2006.07.009.
 - [50] B. A. Kitchenham, O. P. Brereton, A systematic review of systematic review process research in software engineering, *Information and Software Technology* 55 (12) (2013) 2049–2075. doi:10.1016/j.infsof.2013.07.010.
 - [51] L. Chen, M. A. Babar, H. Zhang, Towards an Evidence-Based Understanding of Electronic Data Sources (January 2015) (2010). doi:10.14236/ewic/ease2010.17.
 - [52] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in software engineering, Springer Science & Business Media, 2012. doi:10.1007/978-3-642-29044-2.
 - [53] R. Y. Wang, D. M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems* 12 (4) (1996) 5–33. doi:10.1080/07421222.1996.11518099.
 - [54] T. Dinev, H. Xu, J. H. Smith, P. Hart, Information privacy and correlates: an empirical attempt to bridge and distinguish privacy-related concepts, *European Journal of Information Systems* 22 (3) (2013) 295–316. doi:10.1057/ejis.2012.23.
 - [55] B.-J. Butijn, D. A. Tamburri, W.-J. v. d. Heuvel, Blockchains: a systematic multivocal literature review, *ACM Computing Surveys (CSUR)* 53 (3) (2020) 1–37. URL <https://dl.acm.org/doi/abs/10.1145/3369052>
 - [56] R. Zhang, R. Xue, L. Liu, Security and privacy on blockchain, *ACM Computing Surveys (CSUR)* 52 (3) (2019) 1–34. URL <https://dl.acm.org/doi/10.1145/3316481>
 - [57] H. Nissenbaum, Privacy in Context: Technology, Policy, and the Integrity of Social Life, Stanford University Press, 2009. URL <https://www.sup.org/books/title/?id=8862>
 - [58] G. I. Simari, A Primer on Zero Knowledge Protocols (2002) 12. URL <http://cs.uns.edu.ar/~gis/publications/zkp-simari2002.pdf>
 - [59] D. Chaum, Security without identification: transaction systems to make big brother obsolete, *Communications of the ACM* 28 (10) (1985) 1030–1044. doi:10.1145/4372.4373.
 - [60] J. L. Camenisch, J.-M. Piveteau, M. A. Stadler, Blind signatures based on the discrete logarithm problem, in: A. De Santis (Ed.), *Advances in Cryptology — EUROCRYPT'94*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1995, pp. 428–432. URL <https://link.springer.com/chapter/10.1007/BFb0053458>
 - [61] J. Camenisch, A. Lysyanskaya, Dynamic Accumulators and Application to Efficient Revocation of Anonymous Credentials, in: G. Goos, J. Hartmanis, J. van Leeuwen, M. Yung (Eds.), *Advances in Cryptology — CRYPTO 2002*, Vol. 2442, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 61–76, series Title: Lecture Notes in Computer Science. doi:10.1007/3-540-45708-9_5.
 - [62] J. Camenisch, T. Groß, Efficient Attributes for Anonymous Credentials (2010) 29. URL <https://eprint.iacr.org/2010/496.pdf>
 - [63] S. A. Brands, Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy, MIT Press, Cambridge, MA, USA, 2000. URL <https://direct.mit.edu/books/book/1912/Rethinking-Public-Key-Infrastructures-and-Digital>
 - [64] E. Bangerter, S. Barzan, S. Krenn, A.-R. Sadeghi, T. Schneider, J.-K. Tsay, Bringing Zero-Knowledge Proofs of Knowledge to Practice (2009) 12. URL <https://eprint.iacr.org/2009/211.pdf>
 - [65] M. Hoffmann, M. Klooß, A. Rupp, Efficient Zero-Knowledge Arguments in the Discrete Log Setting, Revisited, in: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ACM, London United Kingdom, 2019, pp. 2093–2110. doi:10.1145/3319535.3354251.
 - [66] T. Nakanishi, H. Yoshino, T. Murakami, G.-V. Policharla, Efficient Zero-Knowledge Proofs of Graph Signature for Connectivity and Isolation Using Bilinear-Map Accumulator, in: *Proceedings of the 7th ACM Workshop on ASIA Public-Key Cryptography*, ACM, Taipei Taiwan, 2020, pp. 9–18. doi:10.1145/3384940.3388959.
 - [67] Y. Zhang, Zero-knowledge proofs for machine learning, in: *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, Association for Computing Machinery, 2020, p. 7. URL <https://doi.org/10.1145/3411501.3418608>
 - [68] M. Stadler, Publicly verifiable secret sharing, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1070 (1996) 190–199. doi:10.1007/3-540-68339-9_17.
 - [69] A. Shamir, How to share a secret, *Commun. ACM* 22 (11) (1979) 612–613. doi:10.1145/359168.359176.
 - [70] T. P. Pedersen, Non-interactive and information-theoretic secure verifiable secret sharing, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 576 LNCS (1992) 129–140. doi:10.1007/3-540-46766-1_9.
 - [71] A. Shamir, How to share a secret, *Publications of the ACM* (1979). doi:10.1007/978-3-642-15328-0_17.
 - [72] A. C. Yao, Protocols for secure computations, in: *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, IEEE, Chicago, IL, USA, 1982, pp. 160–164. doi:10.1109/SFCS.1982.38.
 - [73] Y. Lindell, B. Pinkas, A Proof of Security of Yao's Protocol for Two-Party Computation, *Journal of Cryptology* 22 (2) (2009) 161–188. doi:10.1007/s00145-008-9036-8.
 - [74] A. Ben-David, N. Nisan, B. Pinkas, FairplayMP: a system for secure multi-party computation, in: *Proceedings of the 15th ACM conference*

- on Computer and communications security - CCS '08, ACM Press, Alexandria, Virginia, USA, 2008, p. 257. doi:10.1145/1455770.1455804.
- [75] S. Yakubov, A Gentle Introduction to Yao's Garbled Circuits (2017). URL <https://web.mit.edu/sonka89/www/papers/2017ygc.pdf>
- [76] Z. A. Genç, V. Iovino, A. Rial, The simplest protocol for oblivious transfer, *Information Processing Letters* 161 (2020) 1–12. doi:10.1016/j.ipl.2020.105975.
- [77] P. Pullonen, S. Siim, Combining Secret Sharing and Garbled Circuits for Efficient Private IEEE 754 Floating-Point Computations, in: M. Brenner, N. Christin, B. Johnson, K. Rohloff (Eds.), *Financial Cryptography and Data Security*, Vol. 8976, Springer Berlin Heidelberg, Berlin, Heidelberg, 2015, pp. 172–183, series Title: Lecture Notes in Computer Science. doi:10.1007/978-3-662-48051-9_13.
- [78] Y. Yang, X. Huang, X. Liu, H. Cheng, J. Weng, X. Luo, V. Chang, A Comprehensive Survey on Secure Outsourced Computation and Its Applications, *IEEE Access* 7 (2019) 159426–159465. URL <https://ieeexplore.ieee.org/document/8884162/>
- [79] E. Boyle, N. Gilboa, Y. Ishai, A. Nof, Practical Fully Secure Three-Party Computation via Sublinear Distributed Zero-Knowledge Proofs, in: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, ACM, London United Kingdom, 2019, pp. 869–886. URL <https://dl.acm.org/doi/10.1145/3319535.3363227>
- [80] X. Chen, *Introduction to Secure Outsourcing Computation*, Morgan & Claypool publishers, 2016. URL https://www.researchgate.net/publication/295681472_Introduction_to_Secure_Outsourcing_Computation
- [81] D. Boneh, E.-J. Goh, K. Nissim, Evaluating 2-dnf formulas on ciphertexts, Vol. 3378, 2005, pp. 325–341. URL https://www.researchgate.net/publication/221354138_Evaluating_2-DNF_Formulas_on_Ciphertexts
- [82] V. Nikolaenko, S. Ioannidis, U. Weinsberg, M. Joye, N. Taft, D. Boneh, Privacy-preserving matrix factorization, in: *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, Association for Computing Machinery, 2013, p. 801–812. URL <https://doi.org/10.1145/2508859.2516751>
- [83] L. Zhou, L. Wang, Y. Sun, T. Ai, AntNest: Fully Non-Interactive Secure Multi-Party Computation, *IEEE Access* 6 (2018) 75639–75649. URL <https://ieeexplore.ieee.org/document/8550709/>
- [84] D. Boneh, A. Sahai, B. Waters, Functional Encryption: Definitions and Challenges, in: Y. Ishai (Ed.), *Theory of Cryptography*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 253–273. URL <https://eprint.iacr.org/2010/543.pdf>
- [85] J. Chotard, E. Dufour Sans, R. Gay, D. H. Phan, D. Pointcheval, Decentralized Multi-Client Functional Encryption for Inner Product, in: T. Peyrin, S. Galbraith (Eds.), *Advances in Cryptology – ASIACRYPT 2018*, Springer International Publishing, Cham, 2018, pp. 703–732. URL <https://eprint.iacr.org/2017/989.pdf>
- [86] Z. Brakerski, C. Gentry, V. Vaikuntanathan, (leveled) fully homomorphic encryption without bootstrapping, *ACM Trans. Comput. Theory* 6 (3) (Jul. 2014). doi:10.1145/2633600.
- [87] W. Wang, Y. Hu, L. Chen, X. Huang, B. Sunar, Exploring the Feasibility of Fully Homomorphic Encryption, *IEEE Transactions on Computers* 64 (3) (2015) 698–706. doi:10.1109/TC.2013.154.
- [88] V. Costan, I. Lebedev, S. Devadas, Sanctum: Minimal Hardware Extensions for Strong Software Isolation (2016) 19. URL <https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/costan>
- [89] D. Lee, D. Kohlbrenner, S. Shinde, K. Asanović, D. Song, Keystone: an open framework for architecting trusted execution environments, in: *Proceedings of the Fifteenth European Conference on Computer Systems*, ACM, Heraklion Greece, 2020, pp. 1–16. doi:10.1145/3342195.3387532.
- [90] I. Anati, S. Gueron, S. P. Johnson, V. R. Scarlata, Innovative Technology for CPU Based Attestation and Sealing (2013) 7. URL <https://software.intel.com/content/dam/develop/external/us/en/documents/hasp-2013-innovative-technology-for-attestation-and-sealing-413939.pdf>
- [91] F. Zhang, I. Eyal, R. Escriba, A. Juels, R. V. Renesse, REM: Resource-efficient mining for blockchains, in: *26th USENIX Security Symposium (USENIX Security 17)*, USENIX Association, Vancouver, BC, 2017, pp. 1427–1444. URL <https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/zhang>
- [92] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, Y. Wu, A framework for evaluating gradient leakage attacks in federated learning (2020). arXiv:2004.10397. URL <https://www.semanticscholar.org/paper/A-Framework-for-Evaluating-Gradient-Leakage-Attacks-Wei-Liu/9853a348f61aec83b410f307ab905a4ae001fcd4>
- [93] J. Konečný, B. McMahan, D. Ramage, Federated optimization: distributed optimization beyond the datacenter (2015). arXiv:1511.03575. URL <https://docplayer.net/15450695-Federated-optimization-distributed-optimization-beyond-the-datacenter.html>
- [94] T. Li, A. K. Sahu, A. Talwalkar, V. Smith, Federated learning: Challenges, methods, and future directions, *IEEE Signal Processing Magazine* 37 (3) (2020) 50–60. doi:10.1109/MSP.2020.2975749.
- [95] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, H. Yu, Federated Learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13 (3) (2019) 1–207. doi:10.2200/S00960ED2V01Y201910AIM043.
- [96] P. Vepakomma, O. Gupta, T. Swedish, R. Raskar, Split learning for health: Distributed deep learning without sharing raw patient data (2018). URL https://aiforsocialgood.github.io/iclr2019/accepted/track1/pdfs/31_aig_iclr2019.pdf
- [97] O. Gupta, R. Raskar, Distributed learning of deep neural network over multiple agents, *Journal of Network and Computer Applications* 116 (2018) 1–8. doi:https://doi.org/10.1016/j.jnca.2018.05.003.
- [98] M. G. Poirot, P. Vepakomma, K. Chang, J. Kalpathy-Cramer, R. Gupta, R. Raskar, Split Learning for collaborative deep learning in healthcare 9. URL https://www.researchgate.net/publication/338228319_Split_Learning_for_collaborative_deep_learning_in_healthcare
- [99] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Oct 2016). doi:10.1145/2976749.2978318.
- [100] S. De Capitani Di Vimercati, S. Foresti, G. Livraga, P. Samarati, Data privacy: Definitions and techniques, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 20 (6) (2012) 793–817. doi:10.1142/S0218488512400247. URL <https://www.worldscientific.com/doi/abs/10.1142/S0218488512400247>
- [101] A. Meyerson, R. Williams, On the complexity of optimal k-anonymity, *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems* 23 (2004) 223–228. doi:10.1145/1055558.1055591. URL <https://dl.acm.org/doi/10.1145/1055558.1055591>
- [102] E. Dikici, L. M. Prevedello, M. Bigelow, R. D. White, B. S. Erdal, Constrained generative adversarial network ensembles for sharable synthetic data generation (2020). arXiv:2003.00086. URL https://www.researchgate.net/publication/339642358_Constrained_Generative_Adversarial_Network_Ensembles_for_Sharable_Synthetic_Data_Generation
- [103] A. Torfi, E. A. Fox, C. K. Reddy, Differentially private synthetic medical data generation using convolutional gans (2020). URL https://www.researchgate.net/publication/347624671_Differentially_Private_Synthetic_Medical_Data_Generation_using_Convolutional_GANs
- [104] C. Dwork, A. Smith, T. Steinke, J. Ullman, Exposed! A Survey of Attacks on Private Data, *Annual Review of Statistics and Its Application* 4 (1) (2017) 61–84. doi:10.1146/annurev-statistics-060116-054123.
- [105] ISO, Privacy enhancing data de-identification terminology and classification of techniques (2018).

- URL <https://www.iso.org/standard/69373.html>
- [106] H. Xu, S. Guo, K. Chen, Building confidential and efficient query services in the cloud with rasp data perturbation (2013). doi:10.1109/TKDE.2012.251.
- [107] K. Chen, L. Liu, Geometric data perturbation for privacy preserving outsourced data mining, *Knowledge and Information Systems* 29 (3) (2011) 657–695.
URL <http://link.springer.com/10.1007/s10115-010-0362-4>
- [108] R. Henry, A. Herzberg, A. Kate, Blockchain access privacy: Challenges and directions, *IEEE Security and Privacy* 16 (4) (2018) 38–45. doi:10.1109/MSP.2018.3111245.
- [109] R. Rivest, A. Shamir, Y. Tauman, How to Leak a secret, *Lecture Notes in Computer Science*, vol 2248. Springer, Berlin, Heidelberg. (2001).
URL <https://ieeexplore.ieee.org/document/6032224>
A <https://cryptoslate.com/ethereum-network-congestion-doubles-gas-fees-as-game>
- [110] M. Bellare, D. Micciancio, B. Warinschi, Foundations of Group Signatures: Formal Definitions, Simplified Requirements, and a Construction Based on General Assumptions, *Eurocrypt* 2656 (2003) 1–27.
URL <https://cseweb.ucsd.edu/~mihir/papers/gs.pdf>
- [111] A. F. Westin, *Privacy And Freedom* (1970).
URL <https://www.worldcat.org/title/privacy-and-freedom/oclc/792862>
- [112] A. Cavoukian, *The 7 Foundational Principles* (2011) 2.
URL <https://sites.psu.edu/digitalshred/2020/11/13/privacy-by-design-pbd-the-7-foundational-principles-cavoukian/>
- [113] S. Gürses, C. Troncoso, C. Diaz, Engineering: Privacy by design, *Science* 317 (5842) (2011) 1178–1179.
URL <https://www.esat.kuleuven.be/cosic/publications/article-1542.pdf>
- [114] M. Yung, S. Jarecki, H. Krawczyk, A. Herzberg, Proactive Secret Sharing Or : How to Cope With Perpetual Leakage, *Communication* (1995) 1–22.
URL https://www.researchgate.net/publication/221355399_Proactive_Secret_Sharing_Or_How_to_Cope_With_Perpetual_Leakage
- [115] IOTA-Foundation, About the Tangle (2020).
URL <https://legacy.docs.iota.org/docs/getting-started/1.1/the-tangle/overview>
- [116] E. Heilman, N. Narula, G. Tanzer, J. Lovejoy, M. Colavita, M. Virza, T. Dryja, Cryptanalysis of Curl-P and Other Attacks on the IOTA Cryptocurrency, *IACR Transactions on Symmetric Cryptology* (2020) 367–391 doi:10.46586/tosc.v2020.i3.367-391.
- [117] D. Wang, J. Zhao, Y. Wang, A Survey on Privacy Protection of Blockchain: The Technology and Application, *IEEE Access* 8 (2020) 108766–108781. doi:10.1109/ACCESS.2020.2994294.
URL <https://ieeexplore.ieee.org/document/9093015/>
- [118] J. Sedlmeir, H. U. Buhl, G. Fridgen, R. Keller, The energy consumption of blockchain technology: beyond myth, *Business & Information Systems Engineering* 62 (6) (2020) 599–608.
URL https://www.researchgate.net/publication/342313238_The_Energy_Consumption_of_Blockchain_Technology_Beyond_Myth
- [119] N. Kannengießer, S. Lins, T. Dehling, A. Sunyaev, Trade-offs between distributed ledger technology characteristics, *ACM Computing Surveys (CSUR)* 53 (2) (2020) 1–37.
URL <https://dl.acm.org/doi/10.1145/3379463>
- [120] European Parliament and Council of the European Union, Regulation (eu) 2016/679 directive 95/46/ec (general data protection regulation): General data protection regulation (4 May 2016).
URL <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679>
- [121] A. Ghosh, K. Ligett, A. Roth, G. Schoenebeck, Buying private data without verification, *EC 2014 - Proceedings of the 15th ACM Conference on Economics and Computation* (2014) 931–948 arXiv:1404.6003, doi:10.1145/2600057.2602902.
- [122] S. Ellis, A. Juels, S. Nazarov, Chainlink a decentralized oracle network (2021).
URL <https://link.smartcontract.com/whitepaper>
- [123] S. S. Al-Riyami, K. G. Paterson, Certificateless public key cryptography, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2894 (2003) 452–473.
URL <https://eprint.iacr.org/2003/126.pdf>
- [124] D. Reed, M. Sporny, D. Longley, C. Allen, A. Grant, M. Sabadello, Decentralized identifiers (dids) v1.0 (2021).
URL <https://w3c.github.io/did-core/>
- [125] L. Ismail, H. Hameed, M. AlShamsi, M. AlHammadi, N. AlDhanhani, Towards a Blockchain Deployment at UAE University: Performance Evaluation and Blockchain Taxonomy, in: *Proceedings of the 2019 International Conference on Blockchain Technology*, ACM, Honolulu HI USA, 2019, pp. 30–38. doi:10.1145/3320154.3320156.
- [126] L. Lamport, R. Shostak, M. Pease, *The Byzantine Generals Problem*, Association for Computing Machinery, New York, NY, USA, 2019, p. 203–226. doi:10.1145/3335772.3335936.
- [127] A. Narayanan, V. Shmatikov, Robust De-anonymization of Large Sparse Datasets, in: *2008 IEEE Symposium on Security and Privacy* (sp 2008), IEEE, Oakland, CA, USA, 2008, pp. 111–125, ISSN: 1081-6011. doi:10.1109/SP.2008.33.
- [128] C. Perera, R. Ranjan, L. Wang, S. U. Khan, A. Y. Zomaya, Big Data Privacy in the Internet of Things Era, *IT Professional* 17 (3) (2015) 32–39. doi:10.1109/MITP.2015.34.
- [129] C. Perera, C. H. Liu, S. Jayawardena, The Emerging Internet of Things Marketplace from an Industrial Perspective: A Survey, *IEEE Transactions on Emerging Topics in Computing* 3 (4) (2015) 585–598.
URL <https://ieeexplore.ieee.org/document/7004800>
- [130] L. Determann, No one owns data, *UC Hastings Law* 70 (2018) 44. doi:10.2139/ssrn.3123957.
- [131] L. M. Zagi, B. Aziz, Privacy Attack on IoT: A Systematic Literature Review, in: *7th International Conference on ICT for Smart Society: AIoT for Smart Society, ICISS 2020 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2020.
URL <https://ieeexplore.ieee.org/document/9307568>
- [132] D. Kondor, B. Hashemian, Y.-A. de Montjoye, C. Ratti, Towards Matching User Mobility Traces in Large-Scale Datasets, *IEEE Transactions on Big Data* 6 (4) (2020) 714–726. doi:10.1109/TBDATA.2018.2871693.
- [133] A. Wood, M. Altman, A. Bembeneck, M. Bun, M. Gaboardi, J. Honaker, K. Nissim, D. O'Brien, T. Steinke, S. Vadhan, Differential Privacy: A Primer for a Non-Technical Audience, *SSRN Electronic Journal* (2018). doi:10.2139/ssrn.3338027.
- [134] M. Archie, S. Gershon, A. Katcoff, A. Zeng, De-anonymization of Netflix Reviews using Amazon Reviews (2018) 5.
URL <https://www.readkong.com/page/de-anonymization-of-netflix-reviews-using-amazon-reviews-1439089>
- [135] T. W. S. Journal, Google to pay \$22.5 Million in FTC settlement (2012).
URL <https://www.wsj.com/articles/SB10000872396390443404004577579232818727246>
- [136] J. Porter, Google fined €50 million for GDPR violation in France (2019).
URL <https://www.theverge.com/2019/1/21/18191591/google-gdpr-fine-50-million-euros-data-consent-cnll>
- [137] N. Lomas, France fines Google \$120M and Amazon \$42M for dropping tracking cookies without consent (2020) 1–12.
URL <https://dataprotection.news/france-fines-google-120m-and-amazon-42m-for-dropping-tracking-cookies-without-consent/>
- [138] BBC, H&M fined for breaking GDPR over employee surveillance - BBC News (2020).
URL <https://www.bbc.com/news/technology-54418936>
- [139] Marketing : the Italian SA fines TIM EUR27.8 million (2020).
URL <https://edpb.europa.eu/news/national-news/2020/marketing-italian-sa-fines-tim-eur-278-million-en>
- [140] G. Goos, J. Hartmanis, J. van Leeuwen, D. Hutchison, T. Kanade, J. Kitzler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. P. Rangan, B. Steffen, *Lecture Notes in Computer Science* 556.
URL https://doi.org/10.1007/978-3-319-70139-4_56
- [141] P. Cramton, Y. Shoham, R. Steinberg, An overview of combinatorial auctions, *ACM SIGecom Exchanges* 7 (1) (2007) 3–14.
URL <https://dl.acm.org/doi/10.1145/1345037.1345039>

- [142] R. Dingledine, N. Mathewson, P. Syverson, Tor: The Second-Generation Onion Router., Tech. rep., Defense Technical Information Center, Fort Belvoir, VA (Jan. 2004).
URL <http://www.dtic.mil/docs/citations/ADA465464>
- [143] D. Bogdanov, S. Laur, J. Willemson, Sharemind: a framework for fast privacy-preserving computations 15.
URL https://link.springer.com/chapter/10.1007/978-3-540-88313-5_13
- [144] N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, G. Yu, Collecting and Analyzing Multidimensional Data with Local Differential Privacy (2019) 13.
URL <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8731512>
- [145] R. Bassily, A. Smith, Local, private, efficient protocols for succinct histograms, Proceedings of the forty-seventh annual ACM symposium on Theory of Computing (Jun 2015). doi:10.1145/2746539.2746632.
- [146] A. Herzberg, S. Jarecki, H. Krawczyk, M. Yung, Proactive Secret Sharing Or: How to Cope With Perpetual Leakage, in: D. Coppersmith (Ed.), Advances in Cryptology — CRYPTO' 95, Springer Berlin Heidelberg, Berlin, Heidelberg, 1995, pp. 339–352.
URL https://link.springer.com/chapter/10.1007/3-540-44750-4_27
- [147] A. Bellet, A. Habrard, M. Sebban, A survey on metric learning for feature vectors and structured data (2014). arXiv:1306.6709.
URL <https://hal.archives-ouvertes.fr/hal-01666935>
- [148] B. Poettering, D. Stebila, Double-authentication-preventing signatures, International Journal of Information Security 16 (1) (2017) 1–22.
URL <http://link.springer.com/10.1007/s10207-015-0307-8>
- [149] S. Yu, C. Wang, K. Ren, W. Lou, Achieving secure, scalable, and fine-grained data access control in cloud computing, Proceedings - IEEE INFOCOM (2010). doi:10.1109/INFCOM.2010.5462174.
- [150] R. Wieringa, N. Maiden, N. Mead, C. Rolland, Requirements engineering paper classification and evaluation criteria: A proposal and a discussion, Requirements engineering 11 (1) (2006) 102–107. doi:10.1007/s00766-005-0021-6.

Appendix A. Acronyms

AET	Authenticity-enhancing technologies
BP	Bundling problem
CP	Copy problem
DF	Digital fingerprint
DID	Decentralized identifier
DLT	Distributed ledger technology
DP	Differential privacy
DS	Digital signature
FHE	Fully homomorphic encryption
GAN	Generative adversarial networks
GDPR	General Data Protection Regulation
HE	Homomorphic encryption
ICT	Information and communication technology
IoT	Internet of things
KGC	Key generation center
ML	Machine learning
PET	Privacy-enhancing technology
PHE	Partially homomorphic encryption
PKI	Public key infrastructure
PPDM	Privacy-preserving data mining
REP	Recursive enforcement problem
RQ	Research question
SC	Smart contract
SLR	Systematic literature review
SMC	Secure multiparty computation
TD	Truth discovery
TEE	Trusted execution environment
ZKP	Zero-knowledge proof

Appendix B. Summaries of the selected secondary studies

Table B.3: Secondary studies on privacy in data markets.

Year	Study	Topic	Description
2020	[S8]	Privacy-enhancing design of data markets	Analyzes internet users' preferences for privacy in data sharing to uncover mental models of these preferences and their motives, barriers, and conditions for a privacy-enhancing data market. It provides a set of key findings, the two most notable ones being that the primary barrier to creating data markets is privacy and moral concerns and that the level of anonymization has the largest effect on the willingness to share.
2019	[S19]	Privacy and security data flow challenges in an internet of production	Introduces the internet of production and illustrates its inter-organizational data flows. It identifies security and privacy demands and challenges within these data flows, namely authenticity, data access scope, and anonymity. Furthermore, it provides a small survey of PETs to tackle these challenges: provide confidentiality, hide information during computation (data processing), verify the authenticity of information (providing support), deploy mechanisms that enforce rules (platform capabilities), and support approaches that focus on the security of data flows (external measures).
2019	[S14]	Challenges and research opportunities in data markets for the IoT	A short study that identifies three research opportunities in IoT data markets: Procurement, pricing, and privacy. The most significant identified challenges are: an ambiguity in data ownership that hinders trading, the difficulty to detect data piracy, and that privacy must be considered before trading.
2018	[S23]	Privacy enhancing in IoT applications	Introduces and surveys privacy-enhancing technologies in the processes of data aggregation, trading, and analysis; in particular, it discusses outsourced computation, data validation, and blockchain technology. Additionally, it describes types of privacy breaches and their countermeasures. Furthermore, it reviews relevant aspects of pricing procedures as well as game-theoretical approaches and auction schemes.
2018	[S4]	Pricing, trading, and protection in data markets for the IoT	Surveys the three fields of pricing models and strategies, design of platforms and data trading, and digital copyright mechanisms with a focus on privacy enhancement.
2018	[S48]	Privacy-enhancing analytics for IoT and cloud-based systems	Summarizes privacy-enhancing technologies in the specific use case of a health data collecting app in the health industry. More specifically, it separates privacy-enhancing technologies into two scenarios: Outsourced computation and information sharing.
2016	[S35]	Privacy enhancing in crowdsourcing task management	Surveys privacy-enhancing technologies and the challenges of crowdsourcing task management. The proposed technologies are anonymization, such as k-anonymity, spatio-temporal privacy approaches, such as spatial cloaking or aggregated location via differential privacy, and policy-based privacy preferences. The challenges that they present revolve around trust and credibility, reward-based tasking, utility, efficiency, enforcing privacy-enhancing technologies, and raise privacy awareness.
2015	[S27]	Privacy enhancing and challenges in data markets for the IoT	The study introduces privacy enhancing for data markets for IoT devices, focusing on sensing-as-a-service (data analysis of user-aggregated data). It identifies three challenges: Developing IoT middleware for data analysis and autonomous privacy enhancing, autonomous end-user consent acquisition and negotiation, and the autonomous modeling and negotiation of privacy risk and economic reward. The most prevalent privacy-enhancing technologies and strategies they introduce are personal information hubs, onion routing, and data aggregation via differential privacy or k-anonymity.
2014	[S38]	Privacy threats and challenges in the IoT	Classifies the threats and challenges that come along with privacy in IoT applications for individuals into seven categories: Re-identification of individuals through persistent pseudo-identifiers, localization and tracking, profiling for social engineering and price discrimination, information disclosure in life cycle transitions, information linkage of previously separated systems, inventory attacks, and the disclosure of private information to an uninvited audience.

Appendix C. Mappings of privacy- and authenticity-enhancing technologies

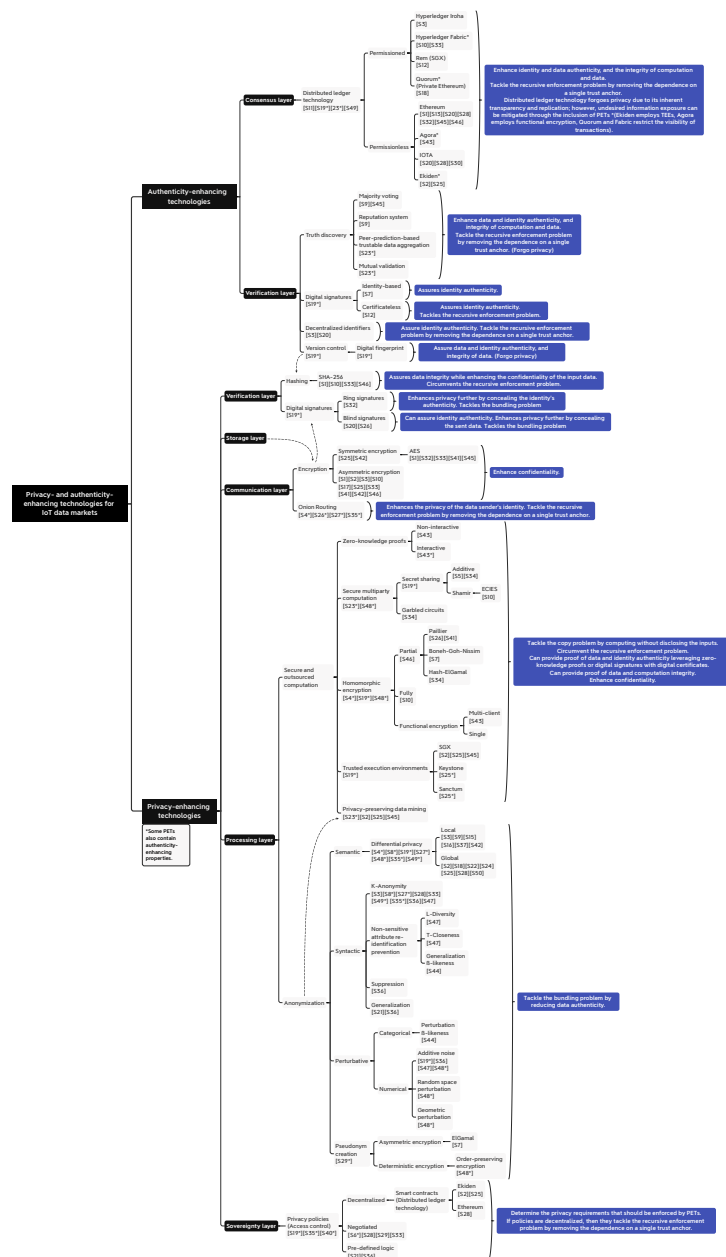


Figure C.11: Classification of the identified privacy- and authenticity-enhancing technologies in this SLR, together with the challenges they tackle. Any other privacy approach encountered in the SLR without a succinct inclusion of the underlying technology was either not included in a leaf node but in a parent node or completely dismissed if too vague. *The publication reviews the technology without delving into it in-depth or using it as a building block of the architecture concept, e.g., the technology is only mentioned in the opportunities for future work.

Appendix D. Summaries of the selected primary studies

Table D.4: Set of examples from the selected studies describing different data marketplace architectures and auctions based on PETs and AETs with a focus on secure computation technologies.

Year	Study	Privacy-enhancing approaches	Description
2020	[S41]	Partially homomorphic encryption, symmetric encryption, and digital signatures	Develops a privacy-enhancing auction for big data trading using Paillier’s cryptosystem [18] and a one-time pad. They consider four entities: sellers, buyers, an auctioneer, and an intermediary platform. A data auction is carried out without any entity seeing the data (except the auction winner) or the bid values, which are ordered obliviously by the auctioneer thanks to the homomorphic properties of the ciphertext. Furthermore, to efficiently encrypt the data, the authors use symmetric encryption (AES). Lastly, digital signatures are created with the same homomorphic cryptosystem, which the authors use to encrypt the symmetric keys.
2018	[S7]	Partially homomorphic encryption, digital signatures, and pseudonym creation	Implements a platform for data markets that facilitates data processing and outcome verification while enhancing the privacy of identities and their data. The authors consider four entities: data contributor, service provider, data consumer, and a registration center; in a two-layer system model: data acquisition and trading. Furthermore, the platform synchronizes data processing and signature verification into the same homomorphic ciphertext space (encrypt and then sign). Additionally, they tightly integrate data processing with outcome verification via a set of homomorphic properties. To achieve a trade-off between functionality and performance, they selected a partially homomorphic scheme called Boneh-Goh-Nissim cryptosystem [81].
2018	[S10]	Fully homomorphic encryption, secure multiparty computation, distributed ledger technology, and hashing	Provides a distributed outsourcing computation architecture, whereby data owners may request fully homomorphic computations with a schema called fully homomorphic non-interactive verifiable secret sharing [83]. Moreover, the proposed architecture allows transactions to be verified by the participants of the permissioned blockchain thanks to the immutability properties of the blockchain; Hyperledger Fabric was the selected blockchain architecture. Moreover, the hash value of the shared data is stored in the blockchain, for data recipients to verify the truthfulness of the received data. Furthermore, for secure multiparty computation, the authors implement Shamir’s secret sharing [69] with ECIES by leveraging the distributed nature of the blockchain. In this manner, the data owner may share verifiable pieces of information with a set of servers. Then, the servers execute the necessary computations, and when several verified responses are received by the agreed data consumer, the true result is recovered.
2016	[S34]	Partially homomorphic encryption, and secure multiparty computation	Develops an auction cloud-based framework that cryptographically hides the bids from all auction participants until a winner is determined. It achieves this by combining PHE based on the hashed scheme [82] of ElGamal [140], and secure two-party computation through garble circuits and additive secret sharing.
2015	[S26]	Partially homomorphic encryption, digital signatures, and onion routing	Proposes a combinatorial auction [141] mechanism that ensures the privacy of the bidders. The bidders bids are blindly signed through the third party [60] so that the third party does not learn the contents. Later, these signatures are used by the bidder to prove the authenticity of the bids. The winner is determined by the third party through a partially homomorphically encrypted computation using the Paillier cryptosystem [18]. Lastly, the identities of the bidders are enhanced by using onion routing [142].
2012	[S5]	Secure multiparty computation	Implements a privacy-enhancing data mining service market, whereby data donors distribute confidential data among a set of participants employing additive secret sharing. The miners collectively perform secure multiparty computation based on the author’s algorithm [143]. Finally, the results are in turn sent to the previously agreed analyst, who combines them to obtain the intelligible output.

Table D.5: Set of examples from the selected studies describing different data marketplace architectures and trading mechanisms based on PETs and AETs with a focus on anonymization technologies - Part I.

Year	Study	Privacy-enhancing approaches	Description
2020	[S24]	Differential privacy	Designs a privacy-enhancing crowd-sensed data trading mechanism. First, the data broker orchestrates an auction whereby data consumers bid in a differentially private manner for a data asset. Secondly, to form a data asset, the data broker creates a set of data generation tasks, some of which are fake to protect the privacy of the auction winner. Lastly, the data broker selects data owner outputs in a differentially private manner. More specifically, both the auction-based data pricing and the data collection are based on the differentially private exponential mechanism.
2019	[S37]	Differential privacy	Proposes a differentially private data market auction framework with a fair negotiation method to set the price and noise; this study is extended in [S42]. The entities involved are a data provider, a consumer, and a trusted market manager that matches providers with consumers and enforces Rubinstein bargaining. Firstly, the data provider and consumer enter a negotiation phase that involves the data query, the ϵ values, and the unit price for ϵ . Once the negotiation is over, the data provider answers the query with the agreed ϵ with local differential privacy.
2019	[S42]	Differential privacy and digital signatures	Proposes a differentially private data market framework. This study extends [S37] by specifying the type of differential privacy algorithm, and the digital signature schemata followed to deploy the framework in practice. The authors use local differential privacy for numeric [144] and for categorical [145] data types.
2019	[S15]	Differential privacy	Designs contracts for a data marketplace whereby a data broker matches the required accuracy from a data consumer with the degree of privacy that data owners desire. Furthermore, by handpicking the data sources, the differentially private algorithm incurs a bias that makes the output more accurate while maintaining the desired privacy from the data owners. Lastly, the authors derive an optimal data contract to minimize payment while satisfying accuracy and privacy.
2019	[S50]	Differential privacy	Proposes a framework for counting trading range query results, and designs a pricing approach for the traded results. Firstly, the framework calculates the range counts approximately, and secondly, it protects the results further by using differential privacy, while satisfying the accuracy demands of data consumers. The authors also design the pricing scheme in a way that prevents arbitrage.
2019	[S22]	Differential privacy	Designs an online auction with two stages, whereby a trusted auctioneer aggregates data from data owners and applies differential privacy before selling the data to consumers. In the first stage, the auctioneer selects data owners based on their privacy requirements to maximize profit. In the second stage, the auctioneer applies differential privacy to the aggregated data and subsequently sells the data to a consumer in an auction.
2018	[S2]	Differential privacy, distributed ledger technology, smart contracts, privacy policies, asymmetric encryption, and trusted execution environments	Implements an end-to-end privacy-enhancing decentralized data marketplace for data consumers to train machine learning algorithms. The authors achieve end-to-end privacy by protecting inputs with asymmetric encryption and differential privacy, and the execution with trusted execution environments. More specifically, differential privacy prevents the weights of machine learning algorithms from overfitting to the inputs. Because of the privacy limitations of current distributed ledger technology applications, the authors of this study and of the subsequent publication [S25] created a novel concept unlike any other blockchain-based system. For example, in principle, the smart contracts of their architecture may contain machine learning algorithms which may be executed in trusted execution environments, hold privacy policies and payment logic, and point to where encrypted data and decryption keys are stored privately. On the other hand, data consumers also deploy smart contracts that may interact with the data owners.

Table D.6: Set of examples from the selected studies describing different data marketplace architectures and trading mechanisms based on PETs and AETs with a focus on anonymization technologies - Part II.

Year	Study	Privacy-enhancing approaches	Description
2018	[S16]	Differential privacy	Develops an auction framework for privacy-enhancing data aggregation for mobile crowdsensing. The auctioneer chooses data owners based on their sensing capabilities, and the data owners apply differential privacy to their inputs sampling from a noise distribution tailored by the auctioneer for each data owner based on its qualities. Moreover, data owners are rewarded for providing accurate data. The goal of the platform is to optimize task allocation to a set of data owners while minimizing their payment, taking into account accuracy and privacy constraints.
2018	[S9]	Differential privacy and truth discovery	Designs two locally differentially private mechanisms for truth discovery in crowd sensing, so that the answers from edge devices are protected while being useful in aggregate. The second mechanism provides more utility for an equal degree of privacy, and consists on the users randomly selecting a probability distribution, and in turn, adding noise sampled thereof to their truthful answer.
2018	[S21]	Privacy policies and generalization	Proposes a data market framework that models and enforces privacy policies dynamically for data-intensive applications. More specifically, the authors implement a data-flow-focused system with a policy enforcement algorithm defined by users and a context. In data-flow computing, directed graphs embody the application, where edges represent data streams and nodes represent functional operators and data sources or sinks. The data is anonymized based on policies and enforced by generalization, e.g., substituting Munich with Germany. To formalize a language to model the privacy policies, the authors use metric first-order temporal logic.
2017	[S47]	K -anonymity and additive noise	Models a data marketplace in which groups of users may actively monetize their data through a mediator and a set of mobile crowd sensing service providers. The authors use a reverse auction, where users bid for performing sensing tasks. Individual users may set their own privacy preferences, and if they are a coalition of users, they are protected by k -anonymity, t -closeness, l -diversity and local noise addition approaches. 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.
2016	[S36]	K -anonymity, additive noise, and privacy policies	Designs a one-to-one privacy enhancing paradigm for a data market place in which privacy policies and data requirements are defined based on the publication record of the data owner. Because published records of a user aggregate over time and thus accrue privacy risk, the paradigm relies on privacy risk management, which is enforced by evaluating the risk associated with revealing yet another piece of information with regard to the privacy requirements. This evaluation is based on the preferences of the user, or if unfeasible, based on current regulation; furthermore, it is based on an assessment of the background information, achieved by semantically analyzing attributes that if released could be linked to externally available information. Ultimately, to privatise the data, the authors propose syntactic technologies such as k -anonymity, suppression, and generalization, and semantic ones like additive noise.

Table D.7: Set of examples from the selected studies describing different data marketplace architectures based on PETs and AETs with an underlying distributed ledger technology – Part I.

Year	Study	Privacy-enhancing approaches	Description
2020	[S31]	Distributed ledger technology, smart contracts, decentralized identifiers, digital signatures, k -anonymity, and differential privacy	Implements a framework for mobility data markets with six layers, each with a purpose and a technology to execute. Furthermore, the framework focuses on location-based services. The identity layer uses asymmetric identity keys, i.e., a key issued only to a real person, to verify that an entity is a real individual. The privacy layer leverages k -anonymity for Geomasking (low utility and high privacy), and when the service needs an exact location, differential privacy for Geo-Indistinguishability (high utility and low privacy). Moreover, the contract layer is based on smart contracts that enforce fair trade and the resolve disputes automatically. For the private communication layer, the authors use decentralized identifiers (DID) [124] issued by the device of a person itself. When devices communicate, the communication has a unique ID based on both of the DIDs, thus, even though the communication data is persisted in a blockchain, it is nontrivial to track the locations of a user. Consecutively, the incentive layer uses smart contracts and data brokers to promote data exchange for a profit; however, this architecture does not tackle the copy problem. The consensus layer is based on a consortium blockchain for distributed governance among non-anonymous honest entities. The blockchain selected was Hyperledger Iroha, chiefly because of its lightweight quality that couples with deployments in IoT devices.
2020	[S43]	Distributed ledger technology, smart contracts, functional encryption, and zero-knowledge proofs	Proposes a privacy-enhancing decentralized data marketplace employing the Agora blockchain with verification technology that enable data prosumers to monetize their data. The privacy-enhancing aspect is achieved by sending encrypted data to brokers employing a primitive called <i>multi-client functional encryption</i> [84][85], which ensures that the receiver may only decrypt the output of a formerly agreed-on function. Moreover, consumers may purchase these outputs, together with a proof of correctness from the broker by using non-interactive zero-knowledge proofs. For the decentralized architecture, the authors employ the Agora blockchain, and atomic payments are performed via smart contracts.
2020	[S45]	Distributed ledger technology, smart contracts, trusted execution environments, truth discovery, digital signatures	Proposes a data processing-as-a-service model based on a blockchain-based data trading ecosystem, whereby neither data brokers nor consumers have access to the raw data, only to the analysis. The use of a blockchain (Ethereum) prevents a single point of failure and allows for immutability and transparency in transactions. Furthermore, to protect the data, the analysis results, and the processing itself, the authors use Intel's SGX trusted execution environment [90], in addition to the symmetric encryption algorithm AES-256 to provide encryption and decryption within and outside the secure environment. The architecture uses the conventional Ethereum Virtual Machine (EVM) for traditional smart contracts, while the data analysis contracts are executed in a SGX-protected EVM where an initial key exchange is needed. Lastly, the nodes in the network form a compute market, i.e., multiple nodes execute the analysis and only the most frequent result is delivered to the data consumer, and the corresponding nodes are rewarded.
2020	[S28]	Distributed ledger technology, privacy policies, differential privacy, k -anonymity, and digital signatures	Proposes an architecture for a personal data marketplace in which personal data is stored decentraly in an allegedly GDPR compliant manner. To accomplish this, transactions and pointers to the data are encrypted and stored using a distributed ledger technology, namely IOTA. The data is stored either in an interplanetary file system, or in an IOTA-based storage format. In order to access such data, a data aggregator must request permissions through Ethereum-based smart contracts (whitelists) owned by data consumers. Once the permission has been granted, the trusted data aggregator, whose mutually agreed privacy policies are persisted in another Ethereum-based smart contract, waits until enough data owners exist to fulfill a particular analysis, so that the aggregator may perform k -anonymity. The data aggregator sells the anonymized data to consumers and remunerates data owners accordingly. However, the presence of a trusted aggregator defeats to some extent the purpose of a decentralized platform. Furthermore, the link between Ethereum and IOTA is carried out by trusted authentication services, which allow data aggregators to decrypt the data. Lastly, in order for data owners to grant access to their data, the authors recommend dynamic threshold encryption [146] over centralized forms of authentication services.
2020	[S18]	Distributed ledger technology, smart contracts, and differential privacy	Proposes a data trading approach in which privacy loss is publicly auditable and data owners set their privacy requirements on publicly available contracts. To accomplish this, the author uses a private Ethereum blockchain called Quorum that supports a set of built-in privacy measures, such as private transactions, messaging, and contracts; however, this design also restricts the interactions that are possible with smart contracts outside the private subset. Furthermore, the data owner applies differential privacy locally before sharing the data with the consumer.
2019	[S25]	Distributed ledger technology, trusted execution environments, smart contracts, privacy policies, digital signatures, and differential privacy	Implements an end-to-end privacy-enhancing decentralized data marketplace for data consumers to train machine learning algorithms, among other Turing-complete tasks. The architecture proposed is a mature version of [S2]. Containing all the features of [S2], R. Cheng, et al. [S25] improve the performance of a newly designed distributed ledger technology to allow for horizontal scaling, i.e., the more nodes are added to the network, the more performant the network is, unlike e.g., Bitcoin or Ethereum; furthermore, the authors tackle the problem of confidentiality by separating consensus from execution, in turn, computations are performed in a trusted execution environment. Horizontal scaling is achieved by allowing for parallel transaction execution, which is, in turn, accomplished by a set of transaction schedulers, and by creating dedicated committees for computation, storage, merging outputs, key management, and consensus. However, scalability through restricting the degree of redundancy entails a security/integrity tradeoff. Key management committees are necessary for the use of trusted execution environments to enable confidential computations. The architecture uses symmetric keys for state encryption and asymmetric encryption for concealing user inputs. The authors achieve end-to-end privacy by protecting inputs with asymmetric encryption and differential privacy, and the execution with trusted execution environments. More specifically, differential privacy prevents the weights of machine learning algorithms from overfitting to the inputs. Because of the privacy limitations of current distributed ledger technology applications, they created a new concept so that smart contracts allow for privacy-enhancing features; this concept was introduced by [S2].

Table D.8: Set of examples from the selected studies describing different data marketplace architectures based on PETs and AETs with an underlying distributed ledger technology – Part II.

Year	Study	Privacy-enhancing approaches	Description
2019	[S32]	Distributed ledger technology, smart contracts, and digital signatures	Proposes a privacy-enhancing fair data trading protocol. The protocol relies on the Ethereum blockchain to achieve a decentralized nature, however, the authors claim the protocol is blockchain agnostic. Nonetheless, despite using a decentralized network, the market manager holds non-negligible authority, as it may trace the identity of sellers so that they can be punished monetarily in case they misbehave. Furthermore, once the buyers have decided which data asset to purchase, the sellers use symmetric keys to encrypt data in chunks before sending it to the buyers. Upon receiving the data chunks, the buyer (i) challenges a set of data chunks, and upon verification of truthful data, (ii) employs similarity learning, a machine learning technology [147], to decide whether to finally purchase the data. Consequently, once the buyer decides to purchase the data, the seller and buyer interact via a payment smart contract and double-authentication-preventing signatures [148] to ensure payment and data decryption. Lastly, in order to enhance the anonymity of the actors, the protocol uses ring signatures [109][110].
2019	[S20]	Distributed ledger technology, smart contracts, digital signatures, and decentralized identifiers	Implements a decentralized data market architecture with secure data processing for the IoT. To achieve decentralization, the authors rely on the distributed ledger technology IOTA, and Ethereum-based smart contracts for subscribing to data streams. The constellation of actors consists of three entities: a data provider, a consumer, and a broker; the former two entities are included in a registry via decentralized identifiers [124]. The product that the consumers purchase is a key to a data stream for a predetermined period of time, created but not accessible by the data broker. For the consumer to attain data access in a private manner, blind signatures [60] are employed, which enable a data broker to verify stream access keys from the data provider without ever accessing these keys. More specifically, the data provider “blinds” the session key with the broker’s public key and sends the blinded key to the broker, consequently, the broker certifies the key with its signature and returns the signature to the provider who removes the blinding factor to access the stream. Lastly, to exchange stream data, an inter-planetary file system is employed.
2019	[S12]	Distributed ledger technology, trusted execution environments, and digital signatures	Proposes a distributed IoT data storage system and a data trading scheme. The authors use the blockchain for its distributed nature, immutability, and requester authentication; moreover, their solution is blockchain agnostic. However, for the consensus algorithm, the authors rely on Intel’s Software Guard Extension (SGX) [90] to deploy a trusted execution environment, to perform “Proof of Useful Work”. The blockchain only contains pointers (addresses) to a distributed hash table, where the data is stored off-chain by peers of the network. Only certified data consumers, e.g., other IoT devices, would be able to query addresses in the blockchain. Furthermore, the authors employ certificateless cryptography so that the key generation center of conventional identity-based encryption does not need to be trusted [123]. To perform the cryptographic operations, edge devices are deployed. Lastly, to share data with purchasers, the authors propose to use either asymmetric encryption or re-encryption [149].
2019	[S1]	Distributed ledger technology, digital signatures, and hashing	Prototypes a decentralized fair data trading platform. The authors rely on the Ethereum blockchain to avoid third-party data brokers and to leverage the ledger’s immutability properties. Moreover, data sellers utilize smart contracts to propose their data offers and to interact with sellers. Sellers include the hash of the data in the ledger so that the buyer may initiate a rebuttal if there is an expectation mismatch. Additionally, to ensure accountability the authors rely on digital signatures to verify that the data was encrypted using a specific key that belongs to the seller, and encrypt the data efficiently using symmetric encryption. The asset traded are decryption keys, and buyers may retrieve data as ciphertexts from untrusted storage.
2018	[S46]	Distributed ledger technology, smart contracts, partially homomorphic encryption, hashing, and digital signatures	Proposes two secure, and fair data trading decentralized schemata built on the Ethereum blockchain. One scheme enables entities to trade raw data, while the other scheme enables them to exchange statistics. The authors chose blockchain in both schemata for its immutability, smart contracts, P2P payment, and disintermediation. Furthermore, for the second scheme, the authors use partially homomorphic encryption to perform confidential statistics. For the data structure to compute the statistics, the authors chose a Merkle Accumulative Tree, where the leaf nodes hold the encrypted data and the non-leaf nodes contain the hash values and a cumulative sum of homomorphic ciphertexts. The data exchanged is verifiable through digital signatures based on asymmetric encryption.
2017	[S33]	Distributed ledger technology, privacy policies, digital signatures, hashing, and k -anonymity	Prototypes a decentralized data market platform for anonymized data. The underlying distributed ledger technology is Hyperledger Fabric, whose peers act as data brokers. The data brokers may only handle datasets based on a set of privacy policies in the interest of the data owner and dictated by a data domain-specific privacy policy manager. The blockchain acts as an auditable ledger for transactions between data brokers and consumers, while the exchange of data is handled off-chain. Furthermore, the anonymization of data is suggested to be performed employing k -anonymity by the broker upon dataset reception from a secure channel, however, the solution remains anonymization-agnostic. For every actor to verify that the correct anonymized dataset has been shared, the broker sends its hash value using SHA-256 to the blockchain before sending it to the consumer. Upon reception, both the policy manager and data receiver may verify the dataset. Lastly, cryptography technologies are employed to encrypt the dataset symmetrically (128-bit AES) before sharing the dataset to the consumer, and the actors use ECDSA to sign confirmations and transactions.

Appendix E. Distribution of selected studies by publication channels

Table E.9: Publication channels for the studies from our SLR.

#	Publication source	Type	No.	%
1	ACM International Conference Proceeding Series	Conference	2	4
2	IEEE International Conference on Internet of Things	Conference	2	4
3	VLDB Endowment	Journal	2	4
4	ACM Conference on Computer and Communications Security	Conference	1	2
5	ACM International Workshop on Mobile Commerce	Workshop	1	2
6	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	Conference	1	2
7	ACM SIGMOD Record	Journal	1	2
8	CEUR Workshop	Workshop	1	2
9	Computer Law and Security Review	Journal	1	2
10	Computer Networks	Journal	1	2
11	Concurrency Computation: Practice and Experience	Journal	1	2
12	Conference on Information and Knowledge Management	Conference	1	2
13	Electronic Markets	Journal	1	2
14	IACR Cryptology ePrint Archive	Journal	1	2
15	IEEE Access	Journal	1	2
16	IEEE Cloud Computing	Journal	1	2
17	IEEE Communications Magazine	Journal	1	2
18	IEEE Computer	Journal	1	2
19	IEEE Eurasia Conference on IOT, Communication and Engineering	Conference	1	2
20	IEEE European Symposium on Security and Privacy	Conference	1	2
21	IEEE International Conference on Big Data	Conference	1	2
22	IEEE International Conference on Blockchain and Cryptocurrency	Conference	1	2
23	IEEE International Conference on Collaboration and Internet Computing	Conference	1	2
24	IEEE International Conference on Pervasive Computing and Communications Workshops	Conference	1	2
25	IEEE International Conference on Software Engineering Research, Management and Applications	Conference	1	2
26	IEEE Internet Computing	Journal	1	2
27	IEEE Internet of Things Journal	Journal	1	2
28	IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking	Conference	1	2
29	IEEE Symposium on Reliable Distributed Systems	Conference	1	2
30	IEEE Transactions on Information Forensics and Security	Journal	1	2
31	IEEE Transactions on Knowledge and Data Engineering	Journal	1	2
32	IEEE Transactions on Network Science and Engineering	Journal	1	2
33	IEEE Transactions on Services Computing	Journal	1	2
34	IEEE Wireless Communications	Journal	1	2
35	Information Sciences	Journal	1	2
36	International Conference on Distributed Computing Systems	Conference	1	2
37	International Conference on Smart Systems and Technologies	Conference	1	2
38	International Conference on Software Engineering	Conference	1	2
39	International Symposium on Mobile Ad Hoc Networking and Computing	Conference	1	2
40	International Workshop on Security and Privacy in Big Data	Workshop	1	2
41	International Workshop on Social Sensing	Workshop	1	2
42	LNCS 7299 – Intelligence and Security Informatics	Journal	1	2
43	Online Information Review	Journal	1	2
44	Security and Communication Networks	Journal	1	2
45	Sensors	Journal	1	2
46	Transportation Research Part C: Emerging Technologies	Journal	1	2
47	Workshop on the Economics of Networks, Systems and Computation	Workshop	1	2
Total			50	100

Appendix F. Electronic data sources and inclusion and exclusion criteria

Table F.10: Electronic data sources (SDS) used in automated search.

ID	Name (Acronym)	Website
EDS1	IEEE Xplore (IEEE)	https://ieeexplore.ieee.org/
EDS2	ACM Digital Library (ACM)	https://dl.acm.org/
EDS3	ISI Web of Science (WoS)	https://www.webofknowledge.com
EDS4	ScienceDirect (SD)	https://www.sciencedirect.com/
EDS5	SpringerLink (SL)	https://link.springer.com/
EDS6	Wiley InterScience (WIS)	https://onlinelibrary.wiley.com/
EDS7	SCOPUS (SCOPUS)	https://www.scopus.com/

Table F.11: Selection criteria used to identify relevant papers. Fulfilling only one exclusion criterion discards the publication from being included.

ID	Facet	Inclusion criterion	Exclusion criterion
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 new to the filtering process	Publication has already been 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

Appendix G. Figures of the metadata analysis



Figure G.12: Map of most active countries in the field of privacy-enhancing data markets for the IoT research.

Table G.12: Classification scheme of research types as described by [150].

Research type	Description
Evaluation research	The authors implement existing techniques, and the solutions are evaluated in practice.
Philosophical papers	These studies present a new perspective on existing research by organizing the domain into a taxonomy or a conceptual framework.
Solution proposal	The authors propose a solution to a problem. The solution can be either novel or a significantly enhanced version of an existent technique. A small example or argumentation demonstrates the benefit and applicability of the solution.

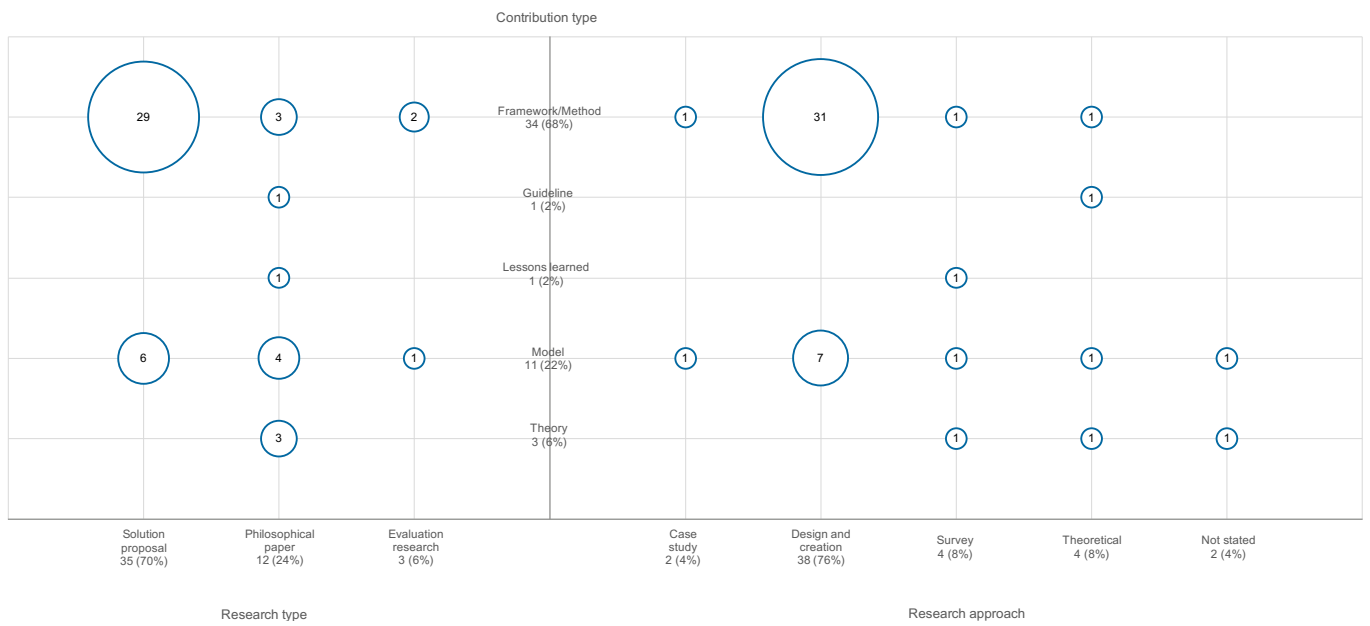


Figure G.13: Mapping of contribution types against research types and approaches.

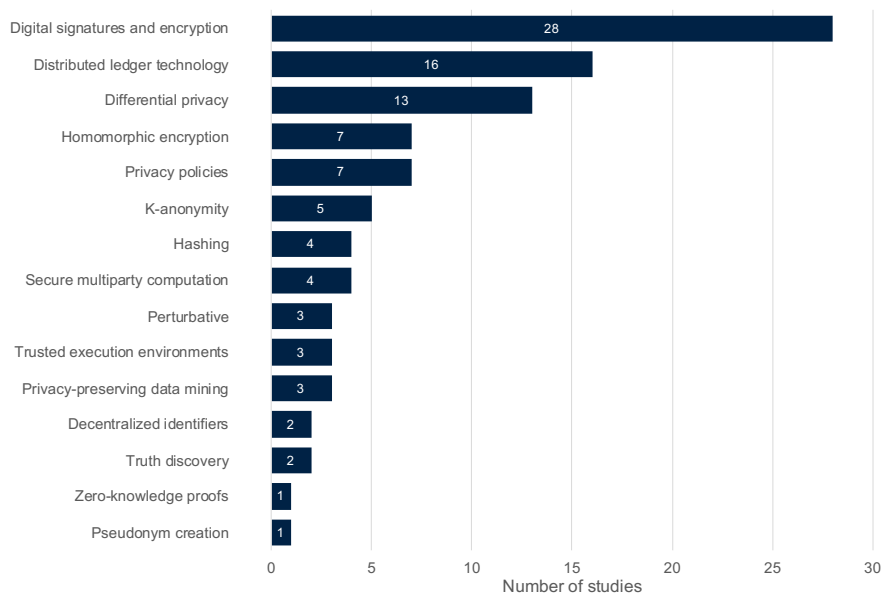


Figure G.14: Distribution of PETs and AETs explicitly employed in the corpus of selected studies.

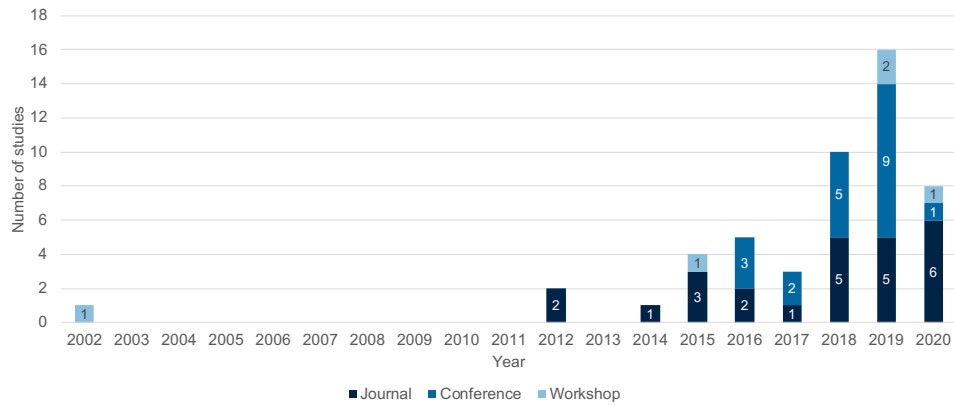


Figure G.15: Distribution of studies over publication domains.

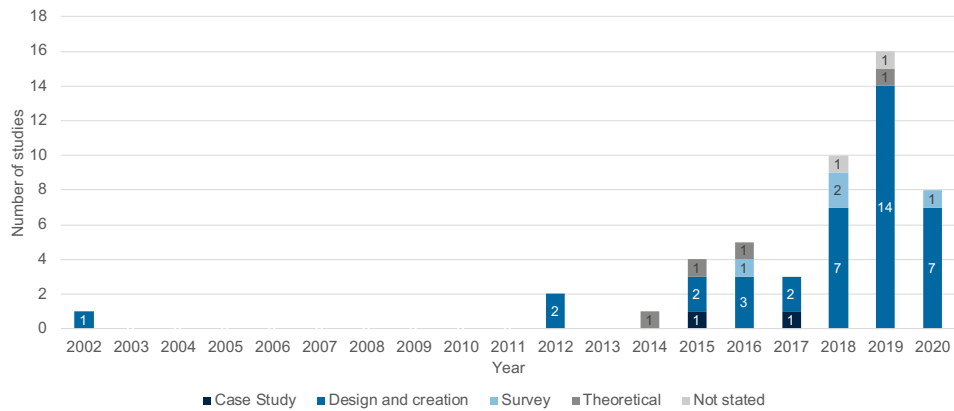


Figure G.16: Number of studies per research approach over time.

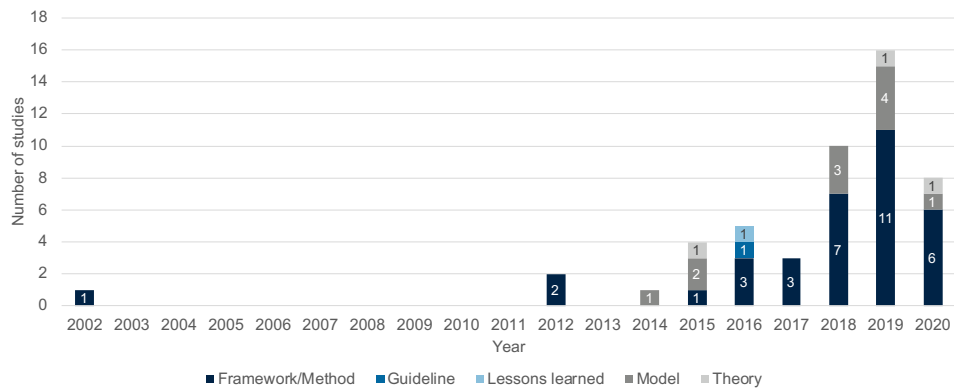


Figure G.17: Distribution of research outcomes over time.