applied sciences

MDPI

Review

Collecting, Processing and Secondary Using Personal and (Pseudo)Anonymized Data in Smart Cities

Silvio Sampaio [1,*], Patricia R. Sousa [2], Cristina Martins [1], Ana Ferreira [3], Luís Antunes [2] and Ricardo Cruz-Correia [3]

[1] Healthy Systems, 4200-135 Porto, Portugal; cristina.martins@hltsys.pt
[2] Competence Centre for Cybersecurity and Privacy (C3P), University of Porto, 4099-002 Porto, Portugal; patricia.sousa@fc.up.pt (P.R.S.); lfa@fc.up.pt (L.A.)
[3] CINTESIS@RISE, MEDCIDS, Faculty of Medicine, University of Porto, 4099-002 Porto, Portugal; amlaf@med.up.pt (A.F.); ricardo.jc.correia@gmail.com (R.C.-C.)
* Correspondence: silvio.sampaio@hltsys.pt

Abstract: Smart cities, leveraging IoT technologies, are revolutionizing the quality of life for citizens. However, the massive data generated in these cities also poses significant privacy risks, particularly in de-anonymization and re-identification. This survey focuses on the privacy concerns and commonly used techniques for data protection in smart cities, specifically addressing geolocation data and video surveillance. We categorize the attacks into linking, predictive and inference, and side-channel attacks. Furthermore, we examine the most widely employed de-identification and anonymization techniques, highlighting privacy-preserving techniques and anonymization tools; while these methods can reduce the privacy risks, they are not enough to address all the challenges. In addition, we argue that de-identification must involve properties such as unlikability, selective disclosure and self-sovereignty. This paper concludes by outlining future research challenges in achieving complete de-identification in smart cities.

Keywords: data privacy; de-identification; anonymization; pseudonymization; re-identification; smart cities

Citation: Sampaio, S.; Sousa, P.R.; Martins, C.; Ferreira, A.; Antunes, L.; Cruz-Correia, R. Collecting, Processing and Secondary Using Personal and (Pseudo)Anonymized Data in Smart Cities. *Appl. Sci.* **2023**, *13*, 3830. https://doi.org/10.3390/app13063830

Academic Editor: Luis Javier García Villalba

Received: 20 January 2023
Revised: 17 February 2023
Accepted: 8 March 2023
Published: 16 March 2023

1. Introduction

Urban data are the backbone of smart cities [1]. However, besides the promises of maximizing control and resources, reducing costs and improving public services, the deployment of smart cities strives in problems related to current technologies, designed for a specific job or context, most needing more privacy as a concern. As urban data include data about the citizens, including personal data, the indiscriminate application of these technologies to the context of a smart city might produce buggy, brittle and hackable urban systems, which create systemic vulnerabilities across critical infrastructure and compromise data security [2].

The intensive collection and processing of personal data, the emergence of privacy regulations, and the ubiquitous Internet of Things (IoT) are among the most challenging aspects of creating privacy-preserving smart city solutions. These challenges are not only caused by the secondary use of data, i.e., data used for another purpose than the one initially collected for, but also by the fact that data should be shared with others to optimize resources, improve the quality of life of inhabitants and create sustainable economic development. One example is provided by [3] using the collected checking-in and checking-out data in the public transportation system as an example. In this case, the primary use for the data is to bill the passenger for the traveled distance. However, the municipalities and transportation companies might post-process these data for better public transportation planning, such as deploying additional or more extended vehicles

for a busy route. The municipalities can also share the data with the police to optimize their police patrols by identifying the hotspots in a city. Commercial companies might also exploit the data to develop new services for citizens, such as (real-time) route planners to guide citizens along the safest and shortest way to their destination. Municipalities might also use these data to improve their environmental policies, e.g., planting trees against air pollution in the identified area. The data types are also diverse and include a wide variety of content, such as geolocation, connections, environment measurements, social networks, text, web, image footage, videos, emails, tweets, audio recordings and others. This diversity is another challenge for the deployment of privacy protection mechanisms. Wide-used solutions, e.g., de-identification and anonymization, become increasingly complex for some data types or even impossible.

The massive amount of data available in a smart city, including personal data, comes from different sources, such as the city's database, IoT sensors and third-party apps. For example, the city's database includes data about geographic information systems; smart building heating, ventilation and air conditioning systems; business supervisor (lighting, parking, waste, finance, etc.) systems; existing video surveillance systems; city police systems; municipal services; etc. In the case of IoT sensors, all things in public spaces must become "smart" to consolidate a truly smart city. So smart cities use sensors and other data collection tools to amass vast information about the town and their citizens that are stored in large data lakes. Later, the city's leadership analyzes all this information to generate an accurate picture of the urban environment, including details on how citizens live in and interact with their city. In most cases, third-party companies or software are employed for the analysis, requiring the data to be shared somewhere (e.g., a private cloud), which brings severe concerns regarding personal data privacy. The data maintained by service providers (bus transport, metro, etc.) and various third-party applications (energy suppliers, weather forecasts, water authorities, Waze, telecom operators, etc) are also a valuable source for the city.

The collection of personal data for the sake of improving services and users' experience is not a new issue of smart cities. Customers have for decades entrusted organizations with their personal information, assuming they will use them to enable better and new services while enhancing the company's decision-making. In recent decades, data collection and intensive processing has accelerated exponentially, enabled by novel communication and processing technologies [4], creating new business areas based on user-related data. Today, data aggregators or brokers range from clear examples such as wide-range private companies (e.g., Google, Amazon, Twitter or Facebook) and government agencies to not-so-clear examples such as our health insurance company or bank. Even if the main activity is not data collection and processing, they all claim ownership and exploitation rights over other people's data. From the data brokers' perspective, consumer data is valuable to businesses and the consumer profile is gold. There are plenty of companies that would love to discover where we shop online, our political preferences, what websites we used to visit, where and how much we spend, our medical history, our lifestyle, in short everything about the consumer. Of course, all this information about users is not always used in abusive transactions but could also be instrumental in helping third-party analysts and investigators answer queries ranging from urban planning [5,6] to treating and curing diseases [7–10]. Even for these legitimate transactions, companies often want to share this information with other parties without compromising the confidentiality of their customers.

Back to the smart cities context, the issue is how to enable the primary and secondary use of the mass of data collected by the different sources and distinct formats in the city without eroding citizens' privacy. The risk of social and economic damage due to the exposure of personal or sensitive information, confirmed by reported abuses in the use of personal data and data breaches, has increased so fast that it has led civil society to demand more protection and guarantees about the privacy of their data, in addition to stricter rules on the collection and processing of this data. As a clear response, several legislations have emerged worldwide to regulate the use of this data and guarantee the individual's

right to privacy. In common, they consider anonymization and pseudonymization as proper technical solutions to obtain the benefits of the secondary use of personal data but without any prejudice to the privacy of individuals. In Europe, the General Data Protection Regulation (GDPR) imposes a uniform data security law on all EU members. With the approach of privacy being a fundamental right, there is a need to review how user data is used and handled by organizations and businesses [11], and cities. Anything that could identify a person, from IP addresses to a digital print [12], is now under protection. GDPR defines personal data as any data of an identified or, at least, identifiable individual—which includes name, email, geolocation, IP address, and also sensitive data related to health or that reveals ethnicity, racial origin or sexual orientation. Even images can be considered personal data, whenever footage or images can be used to directly or indirectly identify an individual.

The risk of re-identification and de-anonymization are critical privacy issues in the development of a smart city. Choosing the right de-identification, anonymization or pseudonymization technique or solution is crucial in ensuring the protection of sensitive information. Once a dataset is made public, it becomes vulnerable to future information releases that could potentially lead to re-identification.

1.1. Contributions

This survey thoroughly examines the fundamental concepts of re-identification risk and de-anonymization and highlights widely used techniques for protecting privacy. Particular emphasis is given to addressing the unique challenges posed by geolocation data and video from surveillance in smart cities, with a focus on presenting proposed mechanisms to mitigate these privacy risks. The methodology for selecting privacy and anonymization techniques and tools in this study was based on the principles of anonymization tools for protecting sensitive data and reducing the risk of re-identification. These tools apply rules and algorithms to data that result in transformations and operations that can be quantified and measured to protect sensitive information. This study also considered ISO/IEC 20889:2018 and used cryptographic techniques to preserve privacy. The aim was to carefully evaluate and choose effective techniques that protect the privacy of individuals and organizations while still allowing for valuable insights to be gained from the data.

This particular study aims to answer the following research questions (RQs) regarding the application of privacy-preserving and cryptographic techniques in Smart Cities:

RQ1. How do de-identification, pseudonymization and anonymization techniques differ in their ability to protect personal information in smart cities?
RQ2. What are the potential risks of de-anonymization and re-identification when using personal data in smart cities, and how can these be minimized?
RQ3. What are the current state-of-the-art methods for preventing de-anonymization and re-identification in smart cities?
RQ4. Are there any limitations to current de-identification and anonymization techniques in the context of smart cities and, if so, how can they be overcome?

1.2. Outline

The remainder of this survey is organized as follows. Section 3 aims to explain the basic principles and terminology used throughout the document. In Section 4, we examine the use cases for de-anonymization and re-identification in smart cities, highlighting potential implications through real-world examples and case studies. Section 5 discuss the application of anonymization tools and privacy-preserving techniques to the use cases discussed in Section 4, demonstrating how they can mitigate the risks associated with de-anonymization and re-identification in smart cities. Section 6 presents the requirements we deem essential for achieving effective de-identification in smart cities. The aim is to provide practitioners with guidelines and recommendations for implementing de-identification in their projects. In Section 7, we examine the current limitations and future research challenges of de-identification in smart cities. It aims to provide insight into ongoing

research and identify areas for further work. Finally, we present the final thoughts and recommendations in Section 9.

2. Related Work

In the recent years, the development of smart cities has brought about numerous advancements in terms of technology and urban management. However, with the increasing amount of data being generated and collected, it is crucial to address security and privacy concerns in the context of smart cities. Previous studies on this topic have attempted to address the functional aspects of anonymization techniques, anonymization operations, privacy models and data anonymity frameworks. Some of them present fuzzy concepts or even misuse of the terms de-identification, anonymization and pseudonymization. Others lack in discussing the implications of using only de-identification, anonymization or pseudonymization. Although most present a comprehensive survey on privacy-preserving and cryptographic techniques [13–15], they need to link the surveyed de-identification techniques to practical privacy problems. Finally, the majority focus on different contexts, such as healthcare [16–21], data mining [22,23], social networks [24,25] and other contexts [26,27], as summarized in Table 1. An exception is [28], which presents a comprehensive review of the privacy-preserving and cryptographic techniques, and briefly elaborates on applying these technologies to some smart city scenarios. Another work worth mentioning is one by Eckhoff and Wagner [29] that, although too general, includes a section to present some privacy-preserving and cryptographic techniques, and also provides some practical privacy problems in the smart city where each technique can be used. Finally, similar to our approach, [30] examines the leading data privacy issues in cyber-physical system deployments in smart cities.

Table 1. Related work case studies.

Literature	Year	Case Study
Our paper	2023	Smart Cities
[29]	2018	Smart Cities
[30]	2019	Smart Cities
[28]	2019	Smart Cities
[16]	2022	Healthcare
[17]	2019	Healthcare
[18]	2021	Healthcare
[19]	2012	Healthcare
[20]	2015	Healthcare
[21]	2015	Healthcare
[22]	2020	Data Mining
[23]	2015	Data Mining
[24]	2020	Social Networks
[25]	2010	Social Networks
[26]	2016	Multimedia Content
[27]	2016	Learning Analytics

Table 2 summarizes the issues addressed by the literature, in order to highlight the main contributions of our work. We consider an issue to be only partially covered if there are only a few examples provided or if the examples are not presented in-depth.

Table 2. Issues addressed by the related work and our work.

Issue	None	Partially Addressed	Fully Addressed
Clarify the terms de-identification, anonymization and pseudonymization	– –	[13–30]	Our Work
Clarify the terms de-anonymization and re-identification	[13–20,22–27,30]	[21,28,29]	Our Work
Link the surveyed privacy-preserving techniques to practical privacy problems faced in smart cities	[13–27]	[28–30]	Our Work
Covers smart city privacy concerns	[13–27]	– –	Our Work and [28–30]

3. Background Concepts

This section explores the main terminologies and concepts that form the basis of de-identification, anonymization and pseudonymization. The different methods and techniques used to protect personal information, as well as the privacy models and anonymization tools currently available, will be examined. For the benefit of discussions in this survey, below, the definitions of a set of key terms are provided. By the end of this section, a comprehensive understanding of the fundamental principles and practices in this field will be provided, serving as a solid foundation for the research presented in the subsequent sections.

3.1. Key Terminology

De-Identification, anonymization and pseudonymization are techniques used to reduce the likelihood of identifying individuals in a dataset with personal data. Its worth to mention there is no standard definition of these terms, and they are highly nuanced and context-dependent. Likewise, the terms De-Anonymization and Re-Identification are interchangeably misused. Therefore, this section provides clear definitions to establish the differences and correct use of these terms that are the basis of the terminology used throughout the paper, answering the RQ1.

De-Identification consists of removing or obfuscating all personal information from a dataset to prevent the identification of individuals. De-identification is not necessarily an irreversible process and it is possible to foresee the existence of a mapping table that allows reversing the process (linking the original records to the de-identified records). In addition to the suppression of all identifying attributes, de-identification usually implies the modification of "quasi-identifier" via the generalization processes (e.g., modifying the scale of an attribute) or by introducing uncertainty factors based on the original values.

Anonymization is usually considered a "strong" case of de-identification, as it aims to make it impractical or even impossible (using all reasonable means) to re-identify (including by the technician who performed the initial operation). In other words, in principle, it should be an irreversible process analogous to destruction. It should be noted that the scope of this definition is adaptable depending on the technological context of the moment: "all means considered reasonable", thus allowing the necessary resources, cost and knowledge necessary to carry out a re-identification to be considered.

Pseudonymization is a process that aims to replace all personal identifiers (e.g., names, addresses and account number) with pseudonyms: artificially generated words or codes, which may function as masked representations of the original data. A "strong" pseudonymization is additionally concerned with focusing on the "quasi-identifier" attributes (e.g., date of birth) and that the attribution of codes is carried out randomly and independently of the original values (although they may eventually continue to be related to each other). Pseudonymization is an approach that provides a form of traceable anonymity

and requires legal, organizational or technical procedures. Consequently, the association can only be carried out under specific and controlled circumstances [31,32]. It should be noted that pseudonymization is not, as a rule, sufficient to guarantee, in light of the GDPR, that the final results of the operation no longer constitute personal data [33]. Additionally, it cannot be neglected that the combination of pseudonymized data with other datasets may allow the total or partial re-identification of individuals.

De-anonymization is a data mining strategy in which anonymous data is cross-referenced with other data sources to re-identify the anonymous data source. Any information that distinguishes one data source from another can be used for de-anonymization. Although this concept goes back several decades, the term made headlines in 2006 when Arvind Narayanan and Vitaly Shmatikov entered a contest hosted by Netflix, a popular movie-rental service. Narayanan and Shmatikov [34] applied their de-anonymization methodology to a dataset that contained the anonymous movie ratings of 500,000 members and were able to identify Netflix data for several specific members successfully. However, the authors emphasized that de-anonymization requires abundant, granular and fairly stable data across time and context.

Re-Identification is the reverse process of de-identification, i.e., data re-identification occurs when personally identifying information is discoverable in de-identified data. The number of re-identification attacks has grown tremendously in recent years. Professionals with experience in the field have realized that removing direct identifiers cannot guarantee correct de-identification [35]. Furthermore, re-identification of personal information is easier and cheaper than ever before, with new databases useful for linking constantly available [36].

According to Lubarsky [37], de-identified (i.e., "pure" de-identified, anonymized or pseudonymized) data can be re-identified through three methods: insufficient de-identification, pseudonym reversal, and combing or linking datasets. Insufficient de-identification occurs when a direct or indirect identifier inadvertently remains in a dataset made available to the public. Pseudonym reversal explores poor de-identification mechanisms that rely on a key that is kept to reverse the process, that use the same pseudonym for a specific individual for a too long period or for which the method used to assign pseudonyms is discovered. Finally, combining or linking one piece of data from the de-identified dataset with other datasets may be enough to reveal the person's real identity. The author also notes these techniques are not mutually exclusive.

El Emam [38] introduced some metrics to quantify the probability of re-identification. These parameters can be applied in datasets or simply in personal information. The probability of re-identification depends on two elements: the number of quasi-identifiers (QIDs) included in the dataset and how disturbed the data is; and the disclosure of information. In general, the more QIDs there are in the available data, the easier the re-identification process [39]. For non-public data, it is essential not to neglect the possibility of a person trying to re-identify an individual in the dataset. The authors in [38,39] developed a methodology based on subjective probability to classify the probability of re-identifying personal data. The method returns a probabilistic value based on risk.

Malin and Sweeney [40] demonstrated in 1997 that 87% of Americans could be re-identified if the date of birth, gender and zip code are provided within a data source. They used a voter register and linked the data sources together to deduce individuals. Several years later, the Facebook myPersonality app published data along with the date of birth, gender and zip code. Therefore, the published data of the myPersonality app was only pseudonymized, but not anonymized.

Summary Notes

In summary, de-identification is a broad term that includes both anonymization and pseudonymization. Figure 1 represents the spectrum of data privacy.

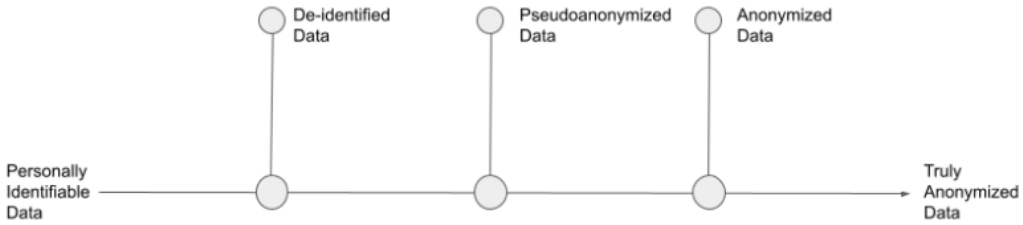

Figure 1. Data privacy spectrum.

The "pure" de-identification only removes the identifiers without any additional analysis of the re-identification risk based on the remaining data—that can include quasi-identifiers—and therefore is a poor technique that must be used carefully for simple and restricted contexts. Since anonymization completely removes any direct and indirect identifiers, it is the most suitable method for the data to be fully open. It is worth noting that anonymized data is complex or, in many cases, impossible to link across multiple sources. Furthermore, the guarantee of non-re-identification is timely as future datasets released might be linked to the anonymized data, allowing re-identification. Due to the poor quality of de-identification methods achieved by the pseudonymization technique compared to anonymization, until now, most of the known re-identification attacks have been successful on pseudonymized data [39,41]. Another drawback of pseudonymization is the difficulty in obtaining metrics that quantify the practical risk of re-identification or even the success rate of the process. Of the various scientific sources consulted, only applications of similarity on numerical data (e.g., [42]) or limited to categories (e.g., [43]) were identified, with an apparent void regarding the application, in this context, of similarity algorithms of texts (e.g., Levenshtein distance, Jaro–Winkler or Jaccard index [44–46]). As such, the reliability of the process will ultimately depend on the responsible analyst and it is impossible to guarantee that all the necessary transformations were carried out correctly to remove the individual character of each record.

As pointed out by Cavoukian and El Emam [47], the growing lack of trust in de-identification, due to a number of noticed cases, and focus on re-identification risks may result in data custodians believing they should not waste their time even attempting to de-identify personal information prior to making it available for secondary purposes.

3.2. Privacy Models

Privacy models are rules and algorithms applied to data, resulting in transformations and operations that can be measured and quantified [48]. The literature shows that some authors also refer to the privacy model as a metric to ensure the re-identification risk threshold has not been surpassed.

The literature includes a variety of competing and complementary privacy models. Besides its well reported weakness [49], *k*-anonymity [50,51] is the most popular [52] privacy model and is still widely accepted as the golden standard [53] for dataset generalization. A dataset is *k*-anonymous if, and only if, for any combination of the associated quasi-identifier attributes, there are at least another $(k-1)$ individuals who share the same values for those same attributes values. In the literature, a large body of work [54–61] contributes with variations of *k*-anonymity. A first weakness relies on the assumption that the set of attributes that an intruder can use to re-identify an individual (the set of quasi-identifiers) is known. This assumption is hazardous when considering the continuous release of new public datasets. Although *k*-anonymity protects against identity disclosure by hiding the individual into a group or equivalence class, it is not enough to provide privacy if sensitive values in an equivalence class lack diversity or if the attacker has background knowledge. The family of *l*-diversity privacy models [62,63] addresses the lack of diversity of sensitive values by ensuring that all equivalence classes contain at least *l*-diverse or "well-represented" values for the sensitive attribute. The *l*-diversity addresses the homo-

geneity and background knowledge attacks. The *l*-diversity is difficult to archive, does not consider the distribution of sensitive attributes and is not even necessary for many datasets. Because of these limitations, [64] presented the *t*-closeness privacy model. The *t*-closeness model achieves privacy by keeping the distribution of each quasi-identifier's sensitive attribute "close" to their distribution in the dataset. This prevents an attacker from learning information about an individual's sensitive attribute value that is not available from the dataset.

Differential privacy [65] is a well-known and mathematical definition-based privacy protection model that ensures that data is collected and analyzed in a way that preserves the privacy of individuals while still allowing for meaningful insights to be drawn from the data. The basic idea behind differential privacy is to add noise to the data so that it is statistically difficult to determine whether or not a particular individual's data was used in the analysis. By definition, the output is not highly affected by the addition or the removal of a single record of the dataset. Differential privacy has been widely adopted in various fields, including healthcare, finance, transportation and smart cities. Among its limitations, it can be computationally expensive to implement and it can be challenging to determine the appropriate noise level for a given dataset.

3.3. Privacy-Preserving and Cryptographic Techniques

This section describes cryptographic techniques as well as privacy preservation mechanisms. The selection of techniques was based on ISO/IEC 20889:2018 [66] and on the techniques referred to throughout this paper, as a solution to the risks of re-identification in smart cities.

Zero-Knowledge Proof (ZKP) [67] is an encryption technique that allows one to verify or prove (mathematically) that a statement about something or someone is real, without having to reveal details about that something or someone. An example of application would be to be able to prove that someone is over eighteen years old to access a site with adult content, without having to reveal identity or even the date of birth, thus guaranteeing anonymity.

Homomorphic Encryption (HE) refers to a class of encryption methods devised by Rivest, Adleman, and Dertouzos as early as 1978 [68] and first constructed by Craig Gentry in 2009 [69]. HE differs from typical encryption methods in that it allows computation to be performed directly in encrypted data without requiring access to a secret key. The result of such a calculation remains in encrypted form and can later be revealed by the owner of the secret key.

Multiparty computation (MPC) is presented as a suitable option to offer the basic building block for building decentralized computers that preserve privacy. The purpose of MPC allows, as parts, the calculation to be a joint function of its private inputs [70]. This protocol must preserve some security properties: the accuracy of the outputs and the privacy of the inputs, even if some of the players are protected by active or passive equipment, without revealing more information about the output of the function itself.

Federated Learning [71] helps in the formation of the machine learning algorithm and keeps the data at the device level. This means that FL allows each device to have its own private and local data. This technology will provide pervasive machine learning solutions as well as flexible, real-time managed data. The technique can be used for numerous tasks and contexts. It includes offline and online learning procedures for the algorithms. Depending on the operational context and data type, the algorithm will choose a suitable technique. Traditional methods, such as centralized machine learning, did not include these benefits and comprise a high risk for data protection and transfer of large files.

3.4. Anonymization Tools

Various off-the-shelf privacy-model-based data de-identification tools have been made available in the past. These tools are commonly adopted for de-identifying tab-

ular data, and a few for unstructured data, in different contexts such as healthcare, financial and governmental.

There are plenty of open-source options. ARX [72] is considered one of the main anonymization tools [73], as it supports a wide variety of (i) privacy and risk models, such *k*-anonymity [50], *l*-diversity, *t*-proximity, δ-presence and differential privacy; (ii) transforming methods for data, such as local and global transformation schemes, value generalization, random sampling, record deletion, attribute and cells, micro-aggregation, top and bottom coding and categorization; risk analyzing of re-identification methods and the usefulness of the resulting data, providing general purpose models, which are cell oriented, with records and attributes [74]. For this reason, in the literature [75] it is used in several studies, either for the anonymization of datasets [76–80] or for the analysis of re-identification risks [81]. Amnesia [82] is another popular anonymization tool [73], following the General Data Protection Regulation (GDPR) guidelines and supporting a few privacy models, such as *k*-anonymity and k^m-anonymity [83,84]. μ-ARGUS [85] is software that aims to help the production of secure microdata. The name ARGUS is an acronym for AntiRe-Identification General Utility System. Initially designed as a private tool, the latest releases have transitioned to open source. To make the microdata safe, the *k*-anonymity privacy model is used in most steps but it is also possible to apply additional transformations, such as local suppression, category grouping, noise addition and synthetic data [86]. Like μ-ARGUS, sdcMicro [87] is an R package that allows anonymization of microdata. SDC is an abbreviation for Statistical Disclosure Control. sdcMicro was developed to assist research on the generation of microdata for public use. In sdcMicro, two privacy models are used, *k*-anonymity and *l*-diversity, as well as methods for data transformation, such as randomization, top and bottom coding, suppression and recoding [18]. Anonimatron [88] performs data anonymization through pseudonymization, and allows fake Roman names, email addresses and universal unique identifiers to be generated, and claims to be GDPR compliant [18]. CHORUS [89] is a framework that provides a Scala library that allows the implementation of differential privacy methods in a cooperative model. The g9 Anonymizer [90] is a tool that comes as an Eclipse plugin, which provides programmable anonymization logic. To achieve data de-identification, the plugin supports data transformations, such as masking, scrambling, data generation synthetics and suppression. The University of Texas at Dallas Anonymization Toolbox [91] is software developed at the UT Dallas Data Security and Privacy Lab and implements six anonymization methods: Datafly [92], Mondrian Multidimensional *k*-Anonymity [93], Incognito [94], Incognito with *l*-diversity [95], Incognito with *t*-closeness [64] and Anatomy [96]. The Cornell Anonymization Toolkit (CAT) [97] is another free tool that allows data anonymization with an intuitive interface. The tool supports the Incognito algorithm, and the *l*-diversity and *t*-proximity privacy models [98].

Some other tools are also free for use but their source code is not available. The Tool for Interactive Analysis of Microdata Anonymization Techniques (TIAMAT) [99] supports different anonymization algorithms, such as Mondrian [93] and k-Member [100] as well as multiple models for analyzing and optimizing the utility of output data, as well as *k*-anonymity, *l*-diversity and *t*-proximity privacy models. The System for Evaluating and Comparing RElational and Transaction Anonymization algorithms (SECRETA) [101] is focused on analyzing the effectiveness and efficiency of anonymization algorithms for tabular as well as set-valued data. SECRETA supports nine algorithms, four to deal with datasets with relational attributes (Incognito [94], Cluster [102], Top-down [103] and Full subtree bottom-up) and five to handle datasets with transaction attributes (COAT [104], PCTA [105], Apriori, LRA and VPA [106]). NLM-Scrubber [107] was developed by the National Library of Medicine to de-identify clinical texts. The goal of NLM-Scrubber is to generate adequate health information with the Health Insurance Portability and Accountability Act (HIPPA). Unlike other tools, NLM-Scrubber performs the de-identification of texts, replacing terms that represent information such as age, address, data and personally identifiable information (PII) by a tag that identifies only the type of information in the text.

For example, in the text: "Dr. Pedro visited the 98-year-old patient…", the de-identified text would contain "Dr. [PERSONALNAME] visited the [AGE90+] patient…".

From the private side, the need of data anonymization tools that protect individuals' and corporations' private activity in compliance with GDPR created a profitable market that explains the emergence of a myriad of tools. Aircloak Insights [108] is a private tool that acts as a proxy between the data analysts and the dataset. Aircloak Insights consists of two components: Insights Air and Insights Cloak. Insights Cloak is responsible for performing the analysis and anonymization of sensitive data, connecting to the databases that contain the sensitive data, without the need for changes. The anonymization is based on a combination of techniques used over time, such as *k*-anonymization, suppression, differential privacy noise, and top and bottom coding. As the limitations of traditional de-identification methods are becoming more evident, modern tools are developed to produce effective results with structured and unstructured data in a vast range of fields and sectors. The new tools mix traditional de-identification methods with new ones, such as synthetic data, natural language processing (NLP) and artificial intelligence. For example, CloverDX [109] is focused on de-identifying production-level datasets for development, visualization, testing, analytics or prototyping. The tool enable a set of data transformations based on a combination of masking and synthetic data generation. Similarly, BizDataX [110] focuses on enabling the anonymization of production data for developing and testing, and offers a data masking toolbox to conceal identities and sensitive data, achieving compliance with GDPR. Created to fit the needs of anonymization of a big pharmaceutical company, Gramener's Data Anonymization Solution [111] uses NLP to redact patients' private information from clinical trial documents, according to HIPAA and GDPR. Another useful tool for redacting sensitive documents is Docbyte's Real-time Automated Anonymization [112]. The tool uses artificial intelligence and machine learning in anonymizing data. The tool can black out or blur images and redact text considered sensitive using image-focused algorithms and object recognition.

Most of the tools listed above have been designed to process only structured data, being unable to extract relevant information from unstructured natural language texts, geolocation data, video footage and images, which represent a significant and essential part of available data in smart cities. Nevertheless, in the following sections, we give some insights into what tools can be helpful in different contexts of the smart city.

4. Smart City De-Anonymization and Re-Identification Use Cases

Peppet [113] argues that IoT objects are more fragile in terms of data protection for three reasons: most companies that intend to integrate products connected to the IoT develop consumer goods, meaning they are not software or hardware developers; much of data security comes from constantly updating software and most IoT objects are not designed for constant updates; and most personal objects are extremely compact (with the exception of cars or refrigerators, for example), which limits processing capacity and sufficient energy to process complex security systems.

The same author also mentions that, although data anonymization is possible, it would actually be an illusion, since the amount of available data would be so large that it would generate a unique digital signature, thus, cross-referencing so much data would make it possible to "re-identify" the user. Re-identification is the process of associating personal data without any type of identifier with the identity of its owner, using auxiliary information [114].

In fact, both problems are threatened by weak anonymization and privacy techniques, but they are two issues that we have to deal with. On the one hand, re-identification and profiling can be done not only through identity management, which can lead to re-identification and profiling, but also through the correlation of data, which we perceive may belong to the same identifier. Even if there is no "literal" identifier, we know that it is a unique person or device that has certain behaviors and patterns, leading to de-anonymization (that is, even with anonymized data, we can correlate or take patterns to

understand behaviors of users and/or devices belonging to specific groups). An example is information that can be shared from users, such as daily professional commitments (which probably have data on the subject of meetings, location and third-party data, such as names and contact information of those involved, etc.), information about daily routines, geolocation and consumption habits, just to mention a few. Therefore, these are the main risks of de-anonymization and re-identification particularly related to smart cities, answering RQ2. We will also refer to de-identification that could lead to re-identification, along with the de-anonymization resulting from lack of or non-effective anonymization and pseudonymization techniques. It is important to note that these are complex issues that require a multi-faceted approach, including the development of advanced privacy-preserving techniques, as well as the implementation of strong security controls and data minimization practices.

This section presents some smart city use cases and targets for de-anonymization and re-identification of personal data/identifiers, answering RQ2 (with the main risks).

4.1. WiFi Probes

WiFi is the most used communication protocol in the IoT. The article written by [115] aims to study behavioral aspects of cell phones with regard to the transmission of probe requests and the exposure of users' privacy. Researchers in a previous work [116] have identified privacy risks associated with WiFi probe requests, such as leaking service set identifiers (SSIDs) already connected by users. Despite several efforts to develop privacy-preserving alternatives, modern mobile devices continue to expose the SSIDs already used by users during WiFi probe requests. In the work in question, the threats of WiFi probe requests to privacy are quantified, carrying out an experimental study of the most popular smartphones in different configurations. The authors' objective is to identify how different factors influence the frequency of probes and the average number of probes transmitted. The findings are worrying: on average, some mobile devices send probe requests at a frequency of 55 times per hour, thus revealing their unique MAC address at a high frequency. When a mobile device is not charging the battery and is also not in sleep mode, it can transmit up to 2000 probes per hour.

The work of [117] presents the idea of capturing and reading WiFi probe request frames to de-anonymize the origin of participants in large events. The authors collected around 11 million records of probe request boards captured in events of different levels of relevance held in Italy. When transmitting probe requests, in many cases, cell phones end up citing in their request the SSID of a network already connected to by the user of the cell device. These networks are maintained by the cell phone in a structure known as the Preferred Network List, or NLP. Some cellular devices even send several requests in a row exposing several networks already connected to by the cellular device user. Capturing such information, it is possible to know which networks users present at a given event usually connect to. By itself, this information would be enough to classify the crowd of people into groups of individuals who frequent the same places. However, the authors also proposed to cross-reference this data with a public database, which maps the geographic location of known wireless networks. An example of a public database with the function of storing information about wireless networks on a global scale is Wigle.net. The mapping of these networks is done cooperatively by voluntary users, who install the application on their cell phones and activate the scanning of wireless networks with the respective upload of the data to the Wigle.net database. Therefore, knowing the MAC addresses of the cellular devices that generated the probe requests; knowing the wireless networks that such users frequently connect to; and knowing the geographic location of such networks; it is then possible to know the origin of the people who are present at a particular public event. The experiments conducted by the authors showed that it is possible to explore the semantic information brought by probe requests, to discover with high precision the provenance of the crowds in each event. An example is the de-anonymization result of two political meetings held a few days before the elections in Italy, which surprisingly

coincide with the reported official voting results. In [118], the authors propose the use of passive monitors to monitor WiFi evidence in a museum to extract information about the behavior of its visitors. More than 1.7 million probes were collected during the six months of capture. The authors obtain promising results regarding the visitors' trajectory, knowing some behavior patterns of their visitors. In [119], a mechanism is presented to detect the social interaction of people, through evidence left in the wireless environment by WiFi networks. For this, they use probe requests to detect devices not associated with a network, and null data frames to detect those that are associated. The article presents techniques for identifying co-location, when two or more cell phones are very close to each other. Using these techniques, the researchers obtained good results for estimating the size of groups of people in a university cafeteria. Through the results, it was possible to observe patterns of behavior on the formation of groups in the different meals of the day: breakfast, lunch, afternoon coffee and dinner.

4.2. Geolocation Data

Leaking users' location information can allow a range of attacks by malicious individuals, ranging from physical surveillance and stalking to identity theft. Another risk is that of inferences from sensitive information. Location data is a strong target for attackers, as it allows direct access to citizens' personal lives, allowing them to know where they have been and for how long, giving the ability to perceive daily habits and use that information for targeted attacks like stalking and spear-phishing. For example, if an individual's location information indicates a hospital, in this case, the data already suggests a series of information related to the place and healthcare, for example, diseases, opening hours, profession and visits to acquaintances, among others.

Location-based services such as parking spot finders disclose not only spatio-temporal user data, but also user queries. The notion that individuals implicitly consent to being monitored when moving in public space is worrying because the lack of alternatives means that consent cannot be meaningfully withheld.

There are other things that may seem harmless but that can result in the identification of a person. Let us imagine an open-data system that shows the information about the real-time location of the electric scooters publicly available, even when they are being driven: this imposes a risk to the privacy, and even security, of the citizens who use this type of transportation. This is an example of information that, if correlated, can lead to the re-identification of a person.

Geolocation is very present in our lives, especially with the adoption of numerous IoT devices that accompany us 24/7. At the moment, we have wearables that, many of them, already come with GPS that can always be on or, at least, during physical exercise such as a running workout. For example, in November 2017, sensitive information about the location and staffing of military bases and spy outposts around the world was revealed by a fitness tracking company. In this case, population density or user density was so low that those facilities could be identified for lack of other data points nearby [7]. The details were released by Strava in a data visualization map that shows all the activity tracked by users of its app, which allows people to record their exercise and share it with others [120]. In this case, it is a matter of security and privacy breach rather than direct re-identification. However, if we think about all the other data that the bracelet will record (heartbeat, daily habits and other personal information), it is possible to arrive at the re-identification of the specific individual, since the spectrum of people who, for example, would be within the base, would already be restricted.

More specifically, there are many studies of re-identification of individuals through the analysis of location data. In 2013, researchers in Europe studied the location data of 1.5 million people and found that the data was so specific to individual habits that they could identify 95% of people with just four location data points [121].

Many of these data are stored in databases. We know that the IoT collects a lot of data, mainly geolocation, and this can lead to the re-identification of people, even without

the individual names or identifiers. The authors of [122] described an investigation of the sensitivity of a dataset with taxi trips without identifiers. The dataset included pick-up location, drop-off location and time, without the names of individuals. The authors claim that, in the investigation, the researchers found photos of celebrities getting into taxi cabs and used metadata from those photos to match up with starting times and locations in the taxi dataset. With this, researchers found the drop-off location of those individuals, getting their home addresses. This means that any large database can expose PII, when combined with other data and other datasets. Anonymization and de-identification techniques attempt to remove PII from datasets. For large datasets, especially datasets with many fields, re-identification is possible [123].

4.3. Medical Data

The correlation of data and re-identification challenges are not limited to publicly released datasets. Google recently acquired an "anonymous" medical record dataset from the University of Chicago Medical Center [124] and a lawsuit alleges that Google has the ability to correlate these datasets with other information at its disposal provision, such as location tracking data from mobile devices, thereby re-identifying individuals and circumventing healthcare-related privacy regulations, such as the Health Insurance Portability and Accountability Act HIPAA) [125]. HIPAA is the Privacy Rule, which raises questions about the US personally identifiable health information privacy standards. This rule defines appropriate security measures to protect the privacy of personal health information and sets limits on the use and disclosure of this information without prior authorization from patients. The Privacy Rule also gives patients rights to their information, such as reviewing it, obtaining a copy of their health information and even requesting corrections to be made.

As mentioned in Section 3.4, NLM-Scrubber can be used for medical data, allowing the de-identification of clinical texts.

4.4. Smart Security

Nowadays, video surveillance systems are present anywhere, whether indoor or outdoor spaces. The ease of acquisition and their availability, alongside their versatility, makes them a primary choice for security purposes. However, they have inherent privacy and security requirements [126], especially about who is watching and has access to both live and video recordings. Although legal and regulatory frameworks try to regulate the use and application of these systems, technological advances tend to be faster and make them obsolete. A concern is the lack of security awareness of the people who manage these systems, where an example is the use of default passwords. Given the ease of obtaining standard vendor passwords and the cracking tool options available with a quick Internet search, it is essential to invest in training. Otherwise, we have a perfect target for hackers.

There are two standards that specify the minimum requirements and give recommendations for video surveillance systems (VSSs), namely the EN 62676-1-1 [127] and EN 62676-4 [128] standards. Furthermore, the introduction of the EU General Data Protection Regulation [129] tries to address the privacy concerns of citizens. Anything that could identify a person, from IP addresses to a digital print [12], is now under protection. GDPR pays special attention to video surveillance (or CCTV).

However, CCTV systems, mostly coupled with facial recognition or automatic number plate recognition, enable the provider to track individuals throughout the city.

4.5. Smart Grids

A key component of smart grids are smart meters, which are devices that measure the energy consumption of households with high precision and frequency. These data allow for the inference of certain information about the daily routines of households. In 2013, a study conducted by [121] demonstrated the potential privacy risks associated with smart meters. The study analyzed features of energy consumption data and was

able to re-identify 68 households out of a sample of 180 households, by studying 60,480 energy-consumption records metered at a frequency of one hour over a period of two weeks. Specifically, the authors were able to re-identify households through features such as the consumption on weekdays and the first increase of energy consumption in the morning. This study highlights the importance of considering the privacy implications of smart grids, as sensitive information about household daily routines can be inferred from energy consumption data.

Reference [130] proposes a privacy-preserving data aggregation scheme, named PPDB, for a fog-computing-based smart grid (FCSG) that supports dynamic billing and arbitration. The scheme uses fog nodes and the ElGamal cryptosystem to encrypt and aggregate electricity consumption data, and employs a trusted third party to arbitrate disputed bills. The scheme is designed with a four-layer data aggregation framework that guarantees confidentiality, authentication and integrity of data. However, as the trusted authority knows the key of each participant, it is easy to pose a threat to privacy.

4.6. Wearable Devices

Wearable devices have grown in popularity in recent years, with a wide range of applications from fitness tracking to medical monitoring. However, these devices also pose significant privacy risks, such as wireless eavesdropping, flawed protocol design, software vulnerabilities and side-channel attacks [131]. Wireless eavesdropping refers to unauthorized individuals or entities intercepting and listening in on wireless communications between the wearable device and other devices, such as smartphones or servers, which can lead to the exposure of sensitive personal information, such as location data or health information.

Flaws in the design and implementation of the device's software and protocols can also pose a significant privacy risk. These flaws can create vulnerabilities that can be exploited by attackers to gain access to personal information or take control of the device. For example, a flaw in the Bluetooth protocol used by a wearable device could allow an attacker to connect to the device without the user's knowledge or consent, potentially giving the attacker access to sensitive data.

Side-channel attacks are another type of attack that can be used against wearable devices. These attacks exploit weaknesses in the device's hardware or software to extract sensitive information, such as encryption keys, without directly accessing the device. For example, a side-channel attack could analyze the power consumption of a wearable device to extract encryption keys used to protect the device's data.

Moreover, sensor data can also be used to re-identify individuals and a wide range of their behaviors and psychological states, as demonstrated in the usage of electrocardiograms (ECGs) and respiration sensors in medical services [132]. Additionally, the geolocation data on these devices can also be subject to re-identification, as stated in Section 4.2.

4.7. Smart Homes

The amount of information that an IoT device can collect is substantial. Webcams can see everything, smart TVs and personal assistants can hear everything, and smart cars can give clues as to whether a person is at home or not. The amount of data that an IoT device can send back to its manufacturers and how they are stored depends solely on them. Most of the time, users are unaware that this information is being sent and shared with external sources. These data can still be intercepted or forwarded to a malicious server if not properly protected. In addition to sound and images, depending on the device, data sent to external sources may include sensitive information, such as IP addresses, other devices connected to the network and location. Some manufacturers may collect confidential users' information and gather patterns about their lives (whether they are at home, the content of their conversations and other information).

Several articles explore the use of voice assistants and feature emerging privacy issues from them [133]. In most cases, users cannot control their data, nor are they aware of data

sharing to external entities. The authors of [134] claim that, despite this and even with the possibility of devices' privacy settings, users do not have enough knowledge to access and edit these privacy settings and often prefer to turn off the voice when they are talking about more private things in the same room, because they do not trust that the assistant is not even listening.

There are numerous devices from brands, such as the Amazon Echo [135], Google Home [136], Apple HomePod [137] or Amazon Echo Dot [138] smart speakers. Millions of smart speakers (from Google in this case) have been sold [139] and worldwide spending on these wireless smart speaker devices reached the $2 billion in 2020 [140]. A vast portion of the population is using these types of devices in their homes, but what are the privacy implications to them? Of course, we are not saying that all these devices disregard privacy policies but the truth is that, with a microphone on, there is always an intrinsic fear.

Reference [134] shows that many users who have voice assistants say they do not care about privacy, but when faced with the possibility of hearing everything they are saying, they prefer to hang up and not use it. Many users who do not have a voice assistant guarantee that they do not trust these devices regarding privacy.

In fact, any smart home devices have always-on sensors that capture users' offline activities in their living spaces and transmit information about these activities on the Internet. However, even when this information is encrypted, an ISP or other network observer can infer privacy-sensitive in-home activities by analyzing Internet traffic from smart homes [141].

4.8. Intelligent Transportation Systems

The evolution of intelligent transport systems occurred in an accelerated, multifaceted way and often based on technological advances considered revolutionary for the urban mobility sector. Recently, the widespread use of ITS in the operation and management of urban mobility has become part of everyday life. Numerous tools are available today for different contexts and scales, with applications that directly impact both locally and across global society [142]. In case of re-identification, a victim/target privacy is heavily compromised in cases where access to vehicle systems provides "attackers" with near-complete information about where, when and for how long the victim/target visited a specific location. It can provide additional information about who the victim/target called. This privacy breach is a major security risk.

Vehicle re-identification methods require sets of detectors mounted along the road. In this technique, a unique serial number for a device in the vehicle is detected at one location and then re-detected (re-identified) further down the road. Travel times and speed are calculated by comparing the time a specific device is detected by sensor pairs. This can be done using the MAC addresses of Bluetooth or other devices, or using the RFID serial numbers of electronic toll collection (ETC) transponders (also called "toll tags") [143].

Furthermore, an increasing number of vehicles are equipped with satellite/GPS navigation systems (satellite navigation) that have two-way communication with a traffic data provider. The position readings from these vehicles are used to calculate vehicle speeds. Modern methods may not use dedicated hardware, but smartphone-based solutions using so-called Telematics 2.0 approaches [142].

Finally, smartphones with multiple sensors can be used to track traffic speed and density. Accelerometer data from smartphones used by car drivers is monitored to find out traffic speed and road quality. Audio data and GPS tagging from smartphones allow the identification of traffic density and possible traffic jams. This was implemented in Bangalore, India, as part of an experimental Nericell research system [144].

4.9. Social Networks

Social network data can be integrated into smart cities to enhance various urban services and operations, such as traffic management, public safety and emergency response. For example, social media can be used to monitor real-time traffic conditions and adjust

traffic lights, to collect information about road accidents and provide real-time alerts, or to facilitate communication between citizens and government agencies.

However, this integration raises privacy concerns as personal information and location data from social media can be easily accessed, collected and analyzed. This can lead to the violation of citizens' privacy rights, as their personal information and location data can be used for unauthorized purposes, such as targeted advertising or political manipulation. Additionally, there is a risk of sensitive information being disclosed or hacked, leading to further privacy breaches. To address these privacy concerns, it is crucial to establish clear and stringent privacy policies and guidelines, as well as secure data management and protection systems.

Francesca et al. [145] focused on analyzing privacy requirements offered by social networks through collecting data from 5000 users with different social network profiles, using image recognition techniques to retrieve personal data accessible through these networks. The aim is to raise awareness about the spread and management of social network data and highlight privacy issues by showing how easily users' data can be retrieved.

4.10. Summary Notes

Smart cities rely heavily on technology and connectivity, making them vulnerable to various forms of cyber attacks. Throughout this section, we evaluated different types of attacks in the context of smart city use cases. From the aforementioned use cases, we can extract the following categorization:

Linking attacks—A linkage attack is a type of attack in which an attacker uses indirect identifiers, also known as quasi-identifiers, to re-identify individuals in an anonymized dataset by combining it with another dataset.

Linking attacks can be executed by combining an anonymous medical database with another, because by overlaying the common attributes of these two databases, it becomes possible to re-identify the individual. An example can be seen in Figure 2.

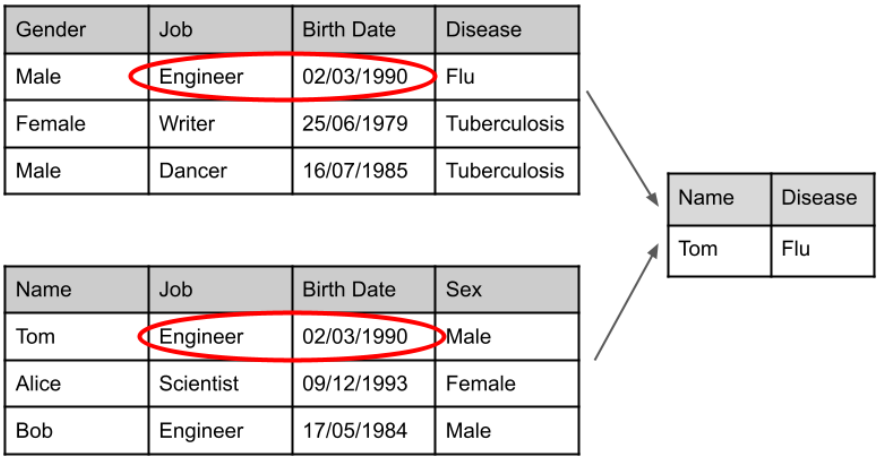

Figure 2. Database linkage attack. The red circles indicate the common attributes of both databases.

Another example is that an attacker can use two databases with spatiotemporal points. One has the linkage with the identifier, and the other has only attributes such as gender, roles or salary. By combining the two equal spatiotemporal points of both databases, the attacker can combine the identifier with the attributes, leading to re-identification of an individual.

Predictive, membership, reconstruction and inference attacks—These types of attack involve using information from an anonymous dataset to make predictions or reconstructions about an individual's identity.

An attacker may try to infer the identity of a specific individual by analyzing their mobility traces, potentially leading to re-identification. An attacker can also analyze

WiFi traces to understand users' daily patterns and habits, potentially leading to de-anonymization, as represented in Figure 3. Such an attack can be executed by combining an anonymous dataset with other easily obtainable information about the individuals.

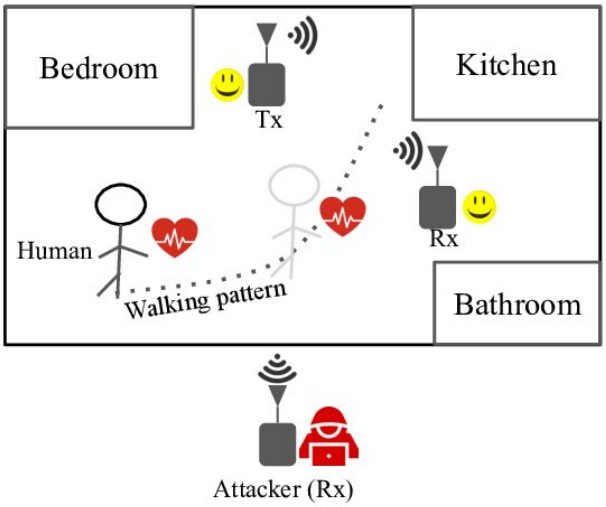

Figure 3. WiFi inference attacks (reference: [146]).

Side-channel attack—This type of attack involves using information from external sources, such as timing data or network traffic, to infer sensitive information about individuals in an anonymous dataset.

For example, a side-channel attack could analyze the power consumption of a wearable device to extract encryption keys used to protect the device's data. Different types of side-channel attacks on a crypto device are represented in Figure 4.

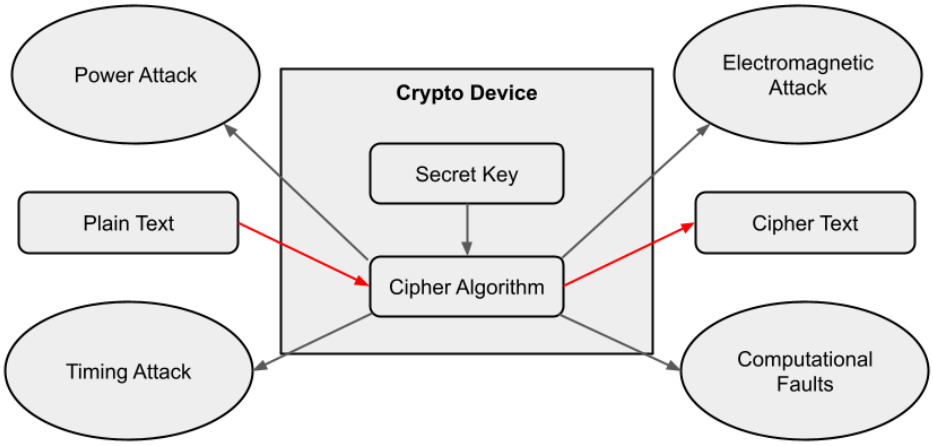

Figure 4. Side-channel attacks on a crypto device.

5. Application of Anonymization Tools and Privacy-Preserving Techniques

Privacy-preserving techniques have been developed to protect personal information while still allowing the data to be used for beneficial purposes. These techniques can include data anonymization, cryptography and access control, and secure computation protocols. Additionally, other privacy-preserving techniques include differential privacy, which adds noise to data to protect individual privacy, and homomorphic encryption, which allows computations to be performed on encrypted data without the need to decrypt it first. These techniques have been developed to enable data to be used for research, analysis and other beneficial purposes while still protecting the personal information of individuals. Overall, these privacy-preserving techniques provide a way to balance the

use of data for beneficial purposes with the need to protect personal privacy. Thus, they answer RQ3 which is about how to protect personal information while still allowing the data to be used for beneficial purposes.

We conducted a comprehensive review of previous studies on techniques and tools to evaluate them theoretically. We analyzed and combined various opinions and identified the limitations of the technologies in question.

5.1. Cryptographic Techniques

Cryptographic techniques are mathematical algorithms and protocols used to secure communication and protect data from unauthorized access. They are an essential component of modern information security systems. Cryptographic techniques can be used for various purposes such as confidentiality (hiding the content of a message), integrity (detecting changes to the message) and authentication (verifying the identity of the sender). Some common cryptographic techniques include symmetric key encryption, asymmetric key encryption, hash functions, digital signatures and zero-knowledge proofs. These techniques are designed to be computationally infeasible to break, making it difficult for attackers to access or modify sensitive information. Cryptographic techniques are widely used in various applications, such as online banking, e-commerce and secure communication over the internet.

Stromire and Potoczny-Jones [147] argue that strong encryption, when implemented correctly, can provide protection against de-anonymization techniques that use statistical analysis, which can uniquely identify a person from surprisingly small pieces of information. They argue that users must be in control of their data. Combining the traditional cryptographic applications with end-to-end cryptography, it is possible to ensure that breaches reveal nothing about the data protected, while maintaining data integrity and authenticity.

There are many use cases where it is necessary to encrypt private data. Mutual authentication is useful for many scenarios, and one of the most known for IoT devices is the leakage of sensitive data during service discovery, such as owner name and service type. By using an identity-based encryption, clients can reveal their identity only to authorized clients. In this way, only authorized customers can decrypt the information [148].

Cryptography can support device-local operations even if the provider has to be assured of their correctness. By supporting the processing of data locally on devices, it is possible to discard raw data [149]. Furthermore, with ZKP for example, it is possible to perform time-of-use billing on smart meters [150] and enforce honesty of vehicles for local processing, e.g., for electronic tolling [151].

CHORUS is an anonymization technique that uses cryptographic techniques to protect the sensitive information in a database. This method is particularly useful in smart cities when data is collected from multiple sources and it is important to ensure the integrity and authenticity of the data.

5.2. Privacy-Preserving Techniques

In this section, we will be discussing various privacy-preserving techniques that are used to protect personal information and prevent re-identification or deanonymization. It is important to note that these are just a few examples of the many techniques available to safeguard sensitive information.

5.2.1. Homomorphic Encryption

In a context of smart city vehicular data sharing, there are electricity providers that have to receive location data and weather data from cars that do the measurement. With HE, it is possible for the electricity supplier to compute on this data without having access to the actual location of the vehicles and still having the supposed results [152].

Another application of HE is given by [153], where fully homomorphic encryption (FHE) is applied to encrypt patient average data for a health-based smart city initiative.

The authors also shows that is possible to compute the heart rates and detect long QT syndrome privately.

5.2.2. Zero-Knowledge Proof

Most location-based services (LBSs) require proof of location (PoL) to prove that the user satisfies the service requirement, which exposes the user's privacy. Wei Wu et al. [154] propose a zero-knowledge proof of location protocol to better protect the user's privacy. With this protocol, the user can choose necessary information to expose to the server, so that hierarchical privacy protection can be achieved.

Another application case is authentication, where an entity proves that it knows a password, without revealing it to the verifier. Furthermore, it can be applied to show the steps in a protocol or process have been done correctly (honest behavior). This approach has been used in smart meter and electronic toll pricing [149].

5.2.3. Multiparty Computation

One approach for vehicular MPC communication is described by [152] that makes a cooperative control strategy incorporating efficient MPC, reducing latency and integrating a secret function sharing scheme. The MPC can be performed using a separate map on different clusters, where each cluster has different vehicles, which together calculate the average energy demand of a given cluster with secure multiparty computing. One vehicle from each cluster is then chosen as the cluster leader which sends the computed result to the destination. To avoid identification of certain vehicles by the address, "shuffling between the cluster leaders" is also possible.

There are other applications of multiparty computation, namely using two-party computation for a privacy-preserving location recommendation scheme for LBS [155]. The scheme supports multi-attribute queries and returns accurate results while ensuring privacy protection for both the service provider and the users. The proposed scheme is based on the Paillier cryptosystem [156] and uses secure equal test protocols to check the equality of encrypted values. The security of the scheme is analyzed in the semi-honest model and experimental results demonstrate its practicality in real-world applications, making it a potential solution for privacy-sensitive smart city applications.

5.2.4. Federated Learning

Federated learning is a state-of-the-art method that enables the creation of machine learning models using datasets distributed across multiple devices, while maintaining the privacy and security of the data. In the context of smart cities, this approach is particularly relevant as centralized data integration can often result in privacy and security concerns [157].

For example, in the field of healthcare, centralized data integration for training machine learning models can result in improved performance compared to separate training using data from just one institution [158]. However, this approach is not feasible due to privacy and security issues. In such cases, federated learning provides a solution as it enables training of models without the direct transfer of data, protecting the privacy and confidentiality of the data. This has been demonstrated through comparative analysis of medical data from multiple institutions, which showed that the training effect of the model obtained through federated learning is nearly identical to that of the centralized approach, while preserving the privacy of the data [158].

Despite its benefits, federated learning is not immune to security concerns. One such vulnerability is the susceptibility to man-in-the-middle attacks, as well as inference attacks aimed at re-identifying data subjects. A recent study [159] has demonstrated this vulnerability through the use of a mobility model called mobility Markov chain, built from the mobility traces observed during the training phase and used to perform the attack during the testing phase. The study used a combination of the closeness between two mobility Markov chain distances to build de-anonymizers that can re-identify users.

To enhance the privacy and security of federated learning, several approaches have been proposed in the literature. One such approach is the use of secure multiparty computation techniques, which enable secure collaboration between multiple parties without revealing their private data [160]. Another approach is the use of homomorphic encryption, which allows for computations to be performed on encrypted data without the need to decrypt it [161]. These techniques can help to improve the privacy and security of federated learning in smart cities and other applications.

In recent years, there has also been growing interest in the application of federated learning in various domains, including IoT [162], mobile networks [163] and autonomous driving [164]. These applications highlight the potential of federated learning to address the privacy and security challenges in smart cities and other domains.

In conclusion, federated learning holds great promise for smart cities, providing a solution to the privacy and security concerns associated with centralized data integration. However, further research is necessary to address the vulnerabilities of this approach, and to enhance its privacy and security features. With its potential to address these challenges, federated learning is poised to play a significant role in the future of data-driven applications in smart cities and beyond.

5.2.5. *K*-Anonymity

Reference [165] presents a privacy notion of client-based personalized *k*-anonymity for autonomous vehicles querying services in cyber-physical systems, allowing users to specify different report sizes representing the anonymity level of each query content.

For example, in a smart city that uses cameras to monitor traffic flow, the cameras may collect images of cars and license plate numbers. To protect the privacy of drivers, the license plate numbers can be replaced with a code that represents a cluster of at least *k* vehicles. This way, it would be impossible to identify a specific vehicle or driver from the data.

The authors of [166] present an approach called heatmap confusion (HMC), which is a location privacy protection mechanism (LPPM) that acts to protect against re-identification attacks. It uses a heat map alteration process to confuse the attacker and to make the re-identification fall to the wrong user. Another approach also focuses on local mobility features: micro-mobility (e.g., individual geographical coordinates). LPPMs are often classified depending on the privacy guarantees they offer to users, mainly *k*-anonymity and differential privacy.

L-diversity and *t*-proximity are enhancements for *k*-anonymity, which is a technique used to protect the privacy of individuals in a dataset by ensuring that each group of records (or quasi-identifiers) with the same values for certain attributes (e.g., age, gender) contains at least *k* records. *L*-diversity is a technique used to ensure that, within each group of records, there is a sufficient number of distinct sensitive attribute values (e.g., disease diagnosis). *T*-proximity is a technique used to ensure that, within each group of records, the sensitive attribute values are similar to one another.

Reference [167] proposes a novel privacy-preserving data collection scheme for IoT-based healthcare service systems that utilizes clustering-based anonymity models to ensure privacy and prevent privacy attacks. The scheme aims to tackle various privacy threats, such as attribute and identity disclosure, and is efficient in reducing communication costs while improving privacy protection. The authors argue that this scheme proved to be more efficient in terms of information loss and data utility compared to *k*-anonymity. Another approach of the same authors, in another publication [168], proposes an attribute-focused privacy-preserving data publishing scheme for sharing healthcare data while protecting patient privacy. The scheme consists of a fixed-interval approach for numerical attributes and an improved *l*-diverse slicing approach for categorical and sensitive attributes. Experiments show improved accuracy of 13% in classification models and reduced information loss by 12% compared to similar approaches.

In the context of smart cities, these techniques could be used to protect the privacy of individuals when sharing data about urban services, such as transportation or energy

usage. For example, when sharing data about the usage of public transportation, *l*-diversity and *t*-proximity could be used to ensure that groups of individuals with similar travel patterns also have a sufficient number of distinct destinations and that the destinations are similar to one another.

The TIAMAT and Cornell Anonymization Toolkit allow for creating *k*-anonymous, *l*-diverse and *t*-closeness-anonymous datasets. They are based on the concept of generalization and suppression to protect the privacy of individuals.

Harmanjeet Kaur et al. [169] propose an improved version of *k*-degree-anonymization for preserving privacy in social network graphs, using the NeuroSVM hybrid technique which reduces distortion in average path length and information loss. The proposed technique is shown to have improved accuracy (over 75%) compared to existing methods in preserving privacy in social networks.

5.2.6. Differential Privacy

Differential privacy is a mathematical framework that allows organizations to share aggregate information about a dataset while ensuring that any individual's data cannot be inferred from the shared information. This is particularly important in the context of smart cities, where large amounts of data are collected from a wide range of sources, including sensors, cameras and mobile devices.

Reference [170] proposes a method for privacy-preserving medical data collection that considers many missing values for more accurate data analysis. The patient data is anonymized and processed by a data collection server using a generative model and a contingency table based on expectation maximization and Gaussian copula methods. The method is evaluated using differential privacy and results show improved accuracy compared to existing methods that do not consider missing values.

Smart cities rely on this data to provide efficient and effective services to citizens, such as transportation, energy and public health. However, this data often contains sensitive information about individuals, such as their location, movement patterns and personal health information. If these data are not protected, they could be used to infer sensitive information about individuals, such as their whereabouts or medical conditions.

In the context of smart cities, Aircloak Insights—that was described in Section 3.4—could be used to protect the privacy of individuals when sharing data about various urban services, such as transportation, energy usage or public health. For example, an organization could use Aircloak Insights to share aggregate information about transportation usage patterns in the city, such as the most popular routes and times of day, without revealing any information about individual travelers.

Aircloak Insights could also be used to share data about other smart city services, such as parking, waste management and air quality, while still protecting the privacy of individuals. It allows multiple parties to access data without revealing the identity of the individual. Overall, Aircloak Insights can be seen as a powerful tool for smart city applications, as it allows organizations to share valuable data while also complying with data privacy regulations and protecting the privacy of citizens.

5.2.7. Obfuscation Techniques

In smart cities, CCTV cameras are often used for traffic management, public safety and crime prevention. However, the footage captured by these cameras often contains sensitive information about individuals, such as their faces and license plate numbers.

To anonymize this data, it is possible to use an anonymization technique called "obfuscation", which aims to conceal sensitive information by making it difficult to read or understand. Other obfuscation techniques include masking, which replaces sensitive information with a fixed value, and generalization, which replaces sensitive information with a less specific value.

Docbyte's Real-time Automated Anonymization uses a combination of image processing and computer vision algorithms to automatically detect and anonymize sensitive

information in video streams, such as faces and license plates. The anonymization technique used in this technology is called "blurring" or "pixelation", which is a method of obscuring sensitive information by replacing it with a blurred or pixelated version. This technique makes it difficult to identify specific individuals or objects while still allowing the overall scene to be visible.

This technology could be applied in the context of smart cities to enhance protection of the privacy of individuals captured on CCTV cameras.

5.2.8. Data Anonymization

Data anonymization through privacy-preserving techniques is crucial in the context of smart cities. With the increasing amount of data being collected from IoT devices, it is important to ensure that sensitive information is protected while still allowing for data analysis and decision making.

One category of tools that can be used for this purpose includes ARX, sdcMICRO and μ-ARGUS. These tools are specifically designed for data anonymization and privacy compliance, allowing cities to automatically anonymize datasets according to specific privacy requirements, ensuring compliance with regulations such as HIPAA and GDPR.

Another category of tools that can be used in smart city contexts include data integration and data masking tools such as CloverDX, BizDataX and Gramener's Data Anonymization Solution. These tools can be used to integrate and anonymize location data from various sources, such as GPS data from public transportation, traffic cameras and mobile devices. The use of these tools can provide valuable insights, such as traffic prediction, people's movement patterns and smart transportation management. However, it is important to note that these techniques and tools can also be applied to other use cases in smart cities.

5.2.9. Other Techniques Related to Location

In this section, we will investigate alternative techniques that are pertinent to location privacy preservation and can be adapted to such use cases.

One widely studied approach is the implementation of dummy locations to safeguard the real mobility trace of individuals. Methods such as SybilQuery [171] aim to preserve privacy while enabling the analysis of mobility patterns.

In the context of smart cities, a plethora of sensor data is generated and stored. In the case of location data, which is considered private information, a set of mobility traces can be collected by a trusted third party or sensors. However, this data must be sanitized before it is shared with other parties, such as a data analytics application running on an untrusted cloud platform.

In the evaluation of patterns, there are lists of users' points of interests (POIs), which are specific places where a user has stopped for a given duration. Promesse [172] eliminates clusters of points that correspond to user stops by utilizing a speed-smoothing algorithm, thus erasing sensitive information about the users. Bindschaedler and Shokri [173] propose the generation of fake traces that share statistical properties with the real traces as a replacement strategy.

6. Requirements for De-Identification

We define as main requirements to be resistant to re-identification:

- Unlinkability—the adversary should not be able to determine whether two blinded credentials are produced from the same self-blindable credential [174]. Therefore, this property ensures that different presentations of the same credential cannot be linked [175].
- Self-sovereignty—individuals no longer depend on a third entity to issue an identifier to them. The individuals will create their own identifiers, maintaining their control and ownership, as well as the information they wish to share, with whom and under what conditions.

- Selective disclosure for minimal information disclosure—in terms of credentials, only a few attributes are required to complete authentication. To protect confidential information, only bits of information are presented to the verifiers [176]. This can be called a "partial identity".

These properties are based on the surveys [177,178], where the authors address the requirements for privacy-preserving identity management.

6.1. Partial Identities

One of the initiatives to preserve user information is through the use of partial identities with minimal information sharing. IBM developed Identity Mixer (IDEMIX) [179], which protects the personal data of online shoppers. This concept allows internet users to use a "partial identity", which can be a "persona" issued by a reliable source, such as a bank. A "persona" is a type of character that differs from the real character of a person or device, acting as a representation. As such, it would be possible to have different personas depending on the context of the user and/or device, and the environment in which they are inserted, considering several characteristics.

With the software, users can make purchases without having to reveal their real name or credit card number, for example. The system gives credibility to the buyer with the store, attesting that it is someone backed by an institution and with funds to make the acquisition. This is important in the context of human identity as, when downloading music from the internet or buying a book from a retailer, the internet user leaves data that can be tracked. This information can be tracked for other users (in the case of public information) or from a data leak (through a database disclosure or de-anonymization) or even re-identification through context data (related users' data).

IDEMIX is a solution from PRIME [180] and its effectiveness in protecting against re-identification has been evaluated by [181].

The solution can be adapted for IoT as it is necessary to implement new techniques that preserve privacy, along with credentials and independent of the usage scenario. Therefore, this could be an interesting future research challenge.

Another interesting technique is the creation of fake profiles, which could leverage synthetic traces of fake users to be the target for confusion. With such a technique, it would be important for the fake profiles to be representative of real users, in order to ensure that the confusion falls on other users rather than the correct one. In this way, even if the data is de-anonymized or re-identified, it would be difficult to determine the real identity of the user, thus protecting their privacy.

In conclusion, IDEMIX and synthetic trace techniques, as well as other privacy-preserving techniques such as encryption, are important methods to protect personal data in smart cities and IoT. However, further research is needed to develop and improve these techniques to effectively protect individual privacy in smart city data. It is also important to consider the trade-offs between the level of privacy protection and the usability of the data, in order to find the right balance and ensure the best possible solution.

6.2. Pseudonymous Credentials

A pseudonym is an identifier that is different from a person's real name and can be used to protect their identity. In smart cities, pseudonymous credentials are used in industry standards for intelligent vehicles [182,183] to ensure drivers' location privacy. These credentials can be used to identify a vehicle or driver without revealing their true identity, thus protecting their privacy and personal information from being tracked or compromised. This is particularly important in smart cities where the use of transportation data is prevalent, and the collection and analysis of this data can pose significant privacy risks.

7. Current Limitations

Anonymization is the process of eliminating any information that relates to an identified or identifiable person to prevent re-identification by any reasonable means. However,

in practice, many anonymization techniques that are currently in use are ineffective, leaving personal information vulnerable to re-identification; while pseudonymization allows for the reconstruction of the original dataset by design, anonymization should make re-identification impossible.

The ISO/IEC 20889:2018 [66] standard provides guidelines for privacy-enhancing data de-identification techniques, including the use of HE as a cryptographic tool for de-identification. HE has several benefits, such as protecting sensitive details while allowing data to be analyzed and processed. However, traditional encryption still contains personal data and may not be reliable or secure. Additionally, HE requires significant computational resources and is currently slow [184], making it less practical for some use cases, and requires accelerators, such as Amazon AWS FPGA. There is also a lot of hard work required to get multiparty computation up and running quickly and with all the security properties correctly implemented, in order to guarantee the privacy of inputs and the correct outputs. Furthermore, multiparty computation still needs to be based on an authenticated protocol so that there are no impersonation attacks, which can also limit its use.

A recent study by [185] evaluated the potential privacy risks of using federated learning in smart cities and found that, even with *k*-anonymity, the technique could still leak private information and allow for re-identification. Similarly, a study by [121] found that it was possible to re-identify up to 95% of individuals in a dataset of anonymous call detail records (CDRs) used for traffic analysis in a smart city, despite *k*-anonymity being applied.

The current methods of anonymization and pseudonymization are inadequate to safeguard personal data in smart city systems. Advanced privacy-preserving techniques are needed for proper protection of individual privacy. Pseudonymization, although cheap and efficient, may not always be enough, and even de-characterized data can still lead to re-identification and be subject to GDPR regulations. Anonymization is the most effective approach for privacy compliance but complete anonymity is not possible, especially with visual data. The use of AI-based tools can speed up the process but 100% protection cannot be guaranteed.

Another key limitation is scalability. Anonymization and pseudonymization techniques may not be able to handle large and complex datasets effectively, particularly when multiple data sources are involved. This can result in data loss and reduced quality of the anonymized data, making it less useful for analysis and decision-making. Another limitation is the potential loss of information during the anonymization process. Therefore, balancing privacy and utility is another major challenge in the field of anonymization and pseudonymization; while it is important to protect personal information, the anonymized data must also retain sufficient information to be useful for analysis and decision-making. This balance is difficult to achieve, particularly in the context of smart cities, where a large amount of data is collected from various sources.

8. Future Research Challenges

De-identification and anonymization of data is a crucial aspect for protecting individual privacy in the era of smart cities. However, current methods are far from perfect and there are several research challenges that must be addressed to achieve greater privacy protection.

One of the key challenges is to develop de-identification methods that ensure unlinkability, which is the ability to prevent the re-identification of an individual from multiple data sources. This can be done by using advanced techniques such as differential privacy or *k*-anonymity.

Another challenge is to improve selective disclosure, which involves allowing access to only the minimum amount of information necessary to perform a specific task, while maintaining privacy. This requires the development of new methods for data masking, data sharing and data access control.

The concept of self-sovereign identity is also gaining momentum, especially using Blockchain technologies. This concept gives individuals control over their own data, allowing them to choose what information they want to share and with whom.

There is also a need to improve the scalability and interoperability of de-identification methods across multiple data sources. This will enable the integration of data from different sources to support data analysis and decision-making without compromising privacy.

The security and reliability of de-identification methods is also a major challenge, as the risk of human error and security breaches must be reduced. This requires the development of secure protocols for data sharing and the implementation of robust security measures, such as encryption and secure key management.

Finally, harmonizing privacy regulations and data protection standards across countries is another important challenge, as privacy laws vary widely from one country to another. This requires international cooperation and a shared understanding of the importance of privacy protection in smart city data.

9. Final Thoughts

In conclusion, our survey provides insight into the current state of privacy concerns in the context of smart cities and the IoT. Our findings highlight the challenges posed by increased connectivity and data collection, and the limitations of privacy-preservation methods such as anonymization and pseudonymization. The complexity of smart cities and IoT, the incentives for malicious actors to de-anonymize personal data, and the rapid pace of technological change and innovation all present additional difficulties in ensuring the privacy of individuals.

It is evident that, while privacy-preservation techniques can reduce privacy risks, they are not enough to fully solve the problem. Therefore, it is imperative for organizations to implement additional measures such as data minimization and strong security controls to further protect personal data and ensure individual privacy.

This survey highlights the importance of continued research and development in privacy-preservation methods and the need for organizations to be proactive in addressing privacy concerns in the context of smart cities and the IoT. Our findings provide a foundation for future research and discussions on how to effectively balance the benefits of increased connectivity with the need to protect personal privacy.

Author Contributions: Conceptualization, S.S. and P.R.S.; methodology, S.S., P.R.S. and C.M.; investigation, S.S. and P.R.S.; writing—original draft preparation, S.S. and P.R.S.; writing—review and editing, C.M., A.F., L.A. and R.C.-C.; supervision, L.A. and R.C.-C.; funding acquisition, A.F. and R.C.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Project *SMART-HEALTH-4-ALL: Smart medical technologies for better health and care*, financed by Programa Operacional Competitividade e Internacionalização da Agência Nacional de Inovação—POCI-01-0247-FEDER-046115; and the Project *City Catalyst—Catalisador para cidades sustentáveis*, ref. POCI-01-0247-FEDER-046112, financed by Fundo Europeu de Desenvolvimento Regional (FEDER), through COMPETE 2020 and Portugal 2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaginalkar, A.; Kumar, S.; Gargava, P.; Niyogi, D. Review of urban computing in air quality management as smart city service: An integrated IoT, AI, and cloud technology perspective. *Urban Clim.* **2021**, *39*, 100972. [CrossRef]
2. Coletta, C.; Evans, L.; Heaphy, L.; Kitchin, R. *Creating Smart Cities*; Routledge: London, UK, 2019; ISBN 9780815396253. [CrossRef]
3. Choenni, S.; Bargh, M.S.; Busker, T.; Netten, N. Data governance in smart cities: Challenges and solution directions. *J. Smart Cities Soc.* **2022**, *1*, 31–51. [CrossRef]

4. Gates, C.; Matthews, P. Data are the new currency. In Proceedings of the 2014 New Security Paradigms Workshop, Victoria, BC, Canada, 15–18 September 2014; pp. 105–116. [CrossRef]
5. Ma, R.; Lam, P.T.I.; Leung, C.K. Big Data in Urban Planning Practices: Shaping Our Cities with Data. In Proceedings of the 21st International Symposium on Advancement of Construction Management and Real Estate, Honk Kong, China, 14–17 December 2016; Chau, K.W., Chan, I.Y., Lu, W., Webster, C., Eds.; Springer: Singapore, 2018; pp. 365–373. [CrossRef]
6. Babar, M.; Arif, F. Smart urban planning using Big Data analytics to contend with the interoperability in Internet of Things. *Future Gener. Comput. Syst.* **2017**, *77*, 65–76. [CrossRef]
7. Venkatesh, R.; Balasubramanian, C.; Kaliappan, M. Development of big data predictive analytics model for disease prediction using machine learning technique. *J. Med. Syst.* **2019**, *43*, 272. [CrossRef]
8. Bansal, S.; Chowell, G.; Simonsen, L.; Vespignani, A.; Viboud, C. Big data for infectious disease surveillance and modeling. *J. Infect. Dis.* **2016**, *214*, S375–S379. [CrossRef]
9. Khan, Z.F.; Alotaibi, S.R. Applications of artificial intelligence and big data analytics in m-health: A healthcare system perspective. *J. Healthc. Eng.* **2020**, *2020*, 8894694. [CrossRef] [PubMed]
10. Zhu, H. Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **2020**, *60*, 573. [CrossRef]
11. Cate, F.H. The EU data protection directive, information privacy, and the public interest. *Iowa L. Rev.* **1994**, *80*, 431.
12. Goddard, M. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *Int. J. Mark. Res.* **2017**, *59*, 703–705. [CrossRef]
13. Pawar, A.; Ahirrao, S.; Churi, P.P. Anonymization techniques for protecting privacy: A survey. In Proceedings of the 2018 IEEE Punecon, Pune, India, 30 November–2 December 2018; IEEE: New York, NY, USA, 2018; pp. 1–6. [CrossRef]
14. Vovk, O.; Piho, G.; Ross, P. Anonymization methods of structured health care data: A literature review. In *Lecture Notes in Computer Science, Proceedings of the International Conference on Model and Data Engineering, Tallinn, Estonia, 21–23 June 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 175–189. [CrossRef]
15. Mogre, N.V.; Agarwal, G.; Patil, P. A review on data anonymization technique for Data publishing. *Int. J. Eng. Res. Technol.* **2012**, *1*, IJERTV1IS10210.
16. Olatunji, I.E.; Rauch, J.; Katzensteiner, M.; Khosla, M. A review of anonymization for healthcare data. *Big Data* **2022**. [CrossRef] [PubMed]
17. Puri, V.; Sachdeva, S.; Kaur, P. Privacy preserving publication of relational and transaction data: Survey on the anonymization of patient data. *Comput. Sci. Rev.* **2019**, *32*, 45–61. [CrossRef]
18. Zuo, Z.; Watson, M.; Budgen, D.; Hall, R.; Kennelly, C.; Al Moubayed, N. Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. *JMIR Med. Inform.* **2021**, *9*, e29871. [CrossRef]
19. Gkoulalas-Divanis, A.; Loukides, G. *Anonymization of Electronic Medical Records to Support Clinical Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012. [CrossRef]
20. Gkoulalas-Divanis, A.; Loukides, G. A survey of anonymization algorithms for electronic health records. In *Medical Data Privacy Handbook*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 17–34. [CrossRef]
21. Nelson, G.S. Practical implications of sharing data: A primer on data privacy, anonymization, and de-identification. In Proceedings of the SAS Global Forum Proceedings, Dallas, TX, USA, 26–29 April 2015; pp. 1–23. Available online: https://www.pharmasug.org/proceedings/2016/IB/PharmaSUG-2016-IB06.pdf (accessed on 9 December 2022).
22. Yang, Y.; Zhou, Y. A Survey on Privacy-Preserving Data Mining Methods. *IOP Conf. Ser. Mater. Sci. Eng.* **2020**, *782*, 022011. [CrossRef]
23. Eze, B.; Peyton, L. Systematic literature review on the anonymization of high dimensional streaming datasets for health data sharing. *Procedia Comput. Sci.* **2015**, *63*, 348–355. [CrossRef]
24. Majeed, A.; Lee, S. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. *IEEE Access* **2020**, *9*, 8512–8545. [CrossRef]
25. Wu, X.; Ying, X.; Liu, K.; Chen, L. A survey of privacy-preservation of graphs and social networks. In *Managing and Mining Graph Data*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 421–453. [CrossRef]
26. Ribaric, S.; Ariyaeeinia, A.; Pavesic, N. De-identification for privacy protection in multimedia content: A survey. *Signal Process. Image Commun.* **2016**, *47*, 131–151. [CrossRef]
27. Khalil, M.; Ebner, M. De-identification in learning analytics. *J. Learn. Anal.* **2016**, *3*, 129–138. [CrossRef]
28. Curzon, J.; Almehmadi, A.; El-Khatib, K. A survey of privacy enhancing technologies for smart cities. *Pervasive Mob. Comput.* **2019**, *55*, 76–95. [CrossRef]
29. Eckhoff, D.; Wagner, I. Privacy in the Smart City—Applications, Technologies, Challenges, and Solutions. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 489–516. [CrossRef]
30. Habibzadeh, H.; Nussbaum, B.H.; Anjomshoa, F.; Kantarci, B.; Soyata, T. A survey on cybersecurity, data privacy, and policy issues in cyber-physical system deployments in smart cities. *Sustain. Cities Soc.* **2019**, *50*, 101660. [CrossRef]
31. Neubauer, T.; Heurix, J. A methodology for the pseudonymization of medical data. *Int. J. Med. Inform.* **2011**, *80*, 190–204. [CrossRef] [PubMed]
32. Riedl, B.; Neubauer, T.; Goluch, G.; Boehm, O.; Reinauer, G.; Krumboeck, A. A secure architecture for the pseudonymization of medical data. In Proceedings of the Second International Conference on Availability, Reliability and Security (ARES'07), Vienna, Austria, 10–13 April 2007; IEEE: New York, NY, USA, 2007; pp. 318–324. [CrossRef]

33. Esayas, S. The role of anonymisation and pseudonymisation under the EU data privacy rules: Beyond the 'all or nothing'approach. *Eur. J. Law Technol.* **2015**, *6*.
34. Narayanan, A.; Shmatikov, V. Robust de-anonymization of large sparse datasets. In Proceedings of the 2008 IEEE Symposium on Security and Privacy (sp 2008), Oakland, CA, USA, 18–22 May 2008; IEEE: New York, NY, USA, 2008; pp. 111–125. [CrossRef]
35. Article 29 Data Protection Working Party. Opinion 05/2014 on Anonymisation Techniques. European Commission. 2014. Available online: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed on 10 December 2022).
36. Cavoukian, A. Dispelling the Myths Surrounding De-Identification. Technical Report. Information and Privacy Commissioner, Ontario. 2011. Available online: https://www.ipc.on.ca/wp-content/uploads/2016/11/anonymization.pdf (accessed on 17 November 2022).
37. Lubarsky, B. Re-Identification of "Anonymized" Data. Georgetown Law Technology Review. 2010. Available online: https://georgetownlawtechreview.org/re-identification-of-anonymized-data/GLTR-04-2017/ (accessed on 10 July 2022).
38. El Emam, K. *Guide to the De-Identification of Personal Health Information*; CRC Press: Boca Raton, FL, USA, 2013. [CrossRef]
39. El Emam, K.; Rodgers, S.; Malin, B. Anonymising and sharing individual patient data. *Br. Med. J.* **2015**, *350*, h1139. [CrossRef]
40. Malin, B.; Sweeney, L. Re-identification of DNA through an automated linkage process. In *Proceedings of the AMIA Symposium*; American Medical Informatics Association: Washington, DC, USA, 2001; p. 423. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243547/ (accessed on 16 November 2022).
41. El Emam, K.; Jonker, E.; Arbuckle, L.; Malin, B. A systematic review of re-identification attacks on health data. *PLoS ONE* **2011**, *6*, e28071. [CrossRef]
42. Sarathy, R.; Muralidhar, K. A Common Index of Similarity for Numerical Data Masking Techniques [Invited Paper]. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. 2009. Available online: https://www.iiisci.org/journal/pdv/sci/pdfs/GS315JG.pdf (accessed on 5 December 2022).
43. Singh, A.; Yu, F.; Dunteman, G. MASSC: A New Data Mask for Limiting Statistical Information Loss and Disclosure. In Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality, Luxembourg, 7–9 April 2003; pp. 373–394. Available online: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2003/04/confidentiality/wp.23.e.pdf (accessed on 2 December 2022).
44. Myers, E.W. AnO (ND) difference algorithm and its variations. *Algorithmica* **1986**, *1*, 251–266. [CrossRef]
45. Ukkonen, E. Algorithms for approximate string matching. *Inf. Control* **1985**, *64*, 100–118. [CrossRef]
46. Choi, S.S.; Cha, S.H.; Tappert, C.C. A survey of binary similarity and distance measures. *J. Syst. Cybern. Inform.* **2010**, *8*, 43–48.
47. Cavoukian, A.; El Emam, K. De-Identification Protocols: Essential for Protecting Privacy. Technical Report. Information and Privacy Commissioner, Ontario. 2014. Available online: https://www.ipc.on.ca/resource/de-identification-protocols-essential-for-protecting-privacy/ (accessed on 17 October 2022).
48. Silva, P.; Monteiro, E.; Simões, P. Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges. *IEEE Access* **2021**, *9*, 10473–10497. [CrossRef]
49. Torra, V.; Domingo-Ferrer, J. A Critique of k-Anonymity and Some of Its Enhancements. In Proceedings of the 2008 3rd International Conference on Availability, Reliability and Security (ARES 08), Barcelona, Spain, 4–7 March 2008; IEEE Computer Society: Los Alamitos, CA, USA, 2008; pp. 990–993. [CrossRef]
50. Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* **2002**, *10*, 557–570. [CrossRef]
51. Samarati, P. Protecting respondents identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 1010–1027. [CrossRef]
52. Slijepčević, D.; Henzl, M.; Klausner, L.D.; Dam, T.; Kieseberg, P.; Zeppelzauer, M. k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Comput. Secur.* **2021**, *111*, 102488. [CrossRef]
53. Jha, N.; Favale, T.; Vassio, L.; Trevisan, M.; Mellia, M. z-anonymity: Zero-Delay Anonymization for Data Streams. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 3996–4005. [CrossRef]
54. Wong, R.C.W.; Li, J.; Fu, A.W.C.; Wang, K. (α, k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing. In *KDD '06: Proceedings of the 12th ACM SIGKDD*; Association for Computing Machinery: New York, NY, USA, 2006; pp. 754–759. [CrossRef]
55. Zhang, Q.; Koudas, N.; Srivastava, D.; Yu, T. Aggregate Query Answering on Anonymized Tables. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 116–125. [CrossRef]
56. Truta, T.; Vinay, B. Privacy Protection: P-Sensitive k-Anonymity Property. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 3–7 April 2006; p. 94. [CrossRef]
57. Nergiz, M.E.; Clifton, C.; Nergiz, A.E. Multirelational k-Anonymity. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1104–1117. [CrossRef]
58. Gionis, A.; Mazza, A.; Tassa, T. k-Anonymization revisited. In Proceedings of the 2008 IEEE 24th International Conference on Data Engineering, Cancun, Mexico, 7–12 April 2008; IEEE: New York, NY, USA, 2008; pp. 744–753. [CrossRef]
59. Terrovitis, M.; Mamoulis, N.; Kalnis, P. Privacy-Preserving Anonymization of Set-Valued Data. In Proceedings of the VLDB Endowment, VLDB Endowment, Seattle, WA, USA, 29 August–3 September 2008; Volume 1, pp. 115–125. [CrossRef]

60. Zhang, Q.; Lin, Z.; Zheng, Q.; Liu, H. (K, G)-anonymity model based on grey relational analysis. In Proceedings of the 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS), Macao, China, 15–17 November 2013; pp. 16–19. [CrossRef]
61. El Emam, K.; Dankar, F.K. Protecting Privacy Using k-Anonymity. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 627–637. [CrossRef] [PubMed]
62. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. L-Diversity: Privacy beyond k-Anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3-es. [CrossRef]
63. Liu, J.; Wang, K. On optimal anonymization for l+-diversity. In Proceedings of the 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), Long Beach, CA, USA, 1–6 March 2010; pp. 213–224. [CrossRef]
64. Li, N.; Li, T.; Venkatasubramanian, S. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 15–20 April 2007; pp. 106–115. [CrossRef]
65. Dwork, C. Differential Privacy. In *Automata, Languages and Programming*; Bugliesi, M., Preneel, B., Sassone, V., Wegener, I., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–12. [CrossRef]
66. ISO/IEC 20889:2018: Privacy Enhancing Data De-Identification Terminology and Classification of Techniques. 2018. Available online: https://www.iso.org/standard/69373.html (accessed on 6 December 2022).
67. Goldreich, O.; Oren, Y. Definitions and properties of zero-knowledge proof systems. *J. Cryptol.* **1994**, *7*, 1–32. [CrossRef]
68. Rivest, R.L.; Adleman, L.; Dertouzos, M.L. On data banks and privacy homomorphisms. *Found. Secur. Comput.* **1978**, *4*, 169–180.
69. Gentry, C.; Sahai, A.; Waters, B. Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based. In Proceedings of the Annual Cryptology Conference, Santa Barbara, CA, USA, 18–22 August 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 75–92. [CrossRef]
70. Yao, A.C. Protocols for secure computations. In Proceedings of the 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), Chicago, IL, USA, 3–5 November 1982; IEEE: New York, NY, USA, 1982; pp. 160–164. [CrossRef]
71. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. Federated learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **2019**, *13*, 1–207. [CrossRef]
72. ARX. ARX—Data Anonymization Tool: A Comprehensive Software for Privacy-Preserving Microdata Publishing. 2022. Available online: https://arx.deidentifier.org (accessed on 6 December 2022).
73. Tomás, J.; Rasteiro, D.; Bernardino, J. Data Anonymization: An Experimental Evaluation Using Open-Source Tools. *Future Internet* **2022**, *14*, 167. [CrossRef]
74. Prasser, F.; Eicher, J.; Spengler, H.; Bild, R.; Kuhn, K.A. Flexible data anonymization using ARX—Current status and challenges ahead. *Softw. Pract. Exp.* **2020**, *50*, 1277–1304. [CrossRef]
75. Vovk, O.; Piho, G.; Ross, P. Evaluation of Anonymization Tools for Health Data. In Proceedings of the International Conference on Model and Data Engineering, Tallinn, Estonia, 21–23 November 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 302–313. [CrossRef]
76. de Oliveira Silva, H.; Basso, T.; de Oliveira Moraes, R.L. Privacy and data mining: Evaluating the impact of data anonymization on classification algorithms. In Proceedings of the 2017 13th European Dependable Computing Conference (EDCC), Geneva, Switzerland, 4–8 September 2017; IEEE: New York, NY, USA, 2017; pp. 111–116. [CrossRef]
77. Jakob, C.E.; Kohlmayer, F.; Meurers, T.; Vehreschild, J.J.; Prasser, F. Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19. *Sci. Data* **2020**, *7*, 435. [CrossRef]
78. Gentili, M.; Hajian, S.; Castillo, C. A case study of anonymization of medical surveys. In Proceedings of the 2017 International Conference on Digital Health, New York, NY, USA, 2–5 July 2017; pp. 77–81. [CrossRef]
79. De Boeck, K.; Verdonck, J.; Willocx, M.; Lapon, J.; Naessens, V. Dataset anonymization with purpose: A resource allocation use case. In Proceedings of the 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), Rome, Italy, 12–14 November 2021; IEEE: New York, NY, USA, 2021; pp. 202–210. [CrossRef]
80. Prasser, F.; Bild, R.; Eicher, J.; Spengler, H.; Kohlmayer, F.; Kuhn, K.A. Lightning: Utility-Driven Anonymization of High-Dimensional Data. *Trans. Data Priv.* **2016**, *9*, 161–185.
81. Jyothi, M.; Rao, M. Preserving the Privacy of Sensitive Data using Data Anonymization. *Int. J. Appl. Eng. Res.* **2017**, *12*, 1639–1663.
82. Amnesia. Amnesia Anonymization Tool. 2022. Available online: https://amnesia.openaire.eu/ (accessed in 6 December 2022).
83. Kulkarni, S.; Bedekar, M. Perception of privacy in a data driven world. *Int. J. Mod. Trends Sci. Technol.* **2022**, *8*, 380–388. [CrossRef]
84. Crutzen, R.; Ygram Peters, G.J.; Mondschein, C. Why and how we should care about the General Data Protection Regulation. *Psychol. Health* **2019**, *34*, 1347–1357. [CrossRef]
85. μ ARGUS. μ-ARGUS—Research. 2022. Available online: https://research.cbs.nl/casc/mu.htm (accessed on 6 December 2022).
86. Stenersen, H.W. Anonymization of Health Data. Master's Thesis, University of Oslo, Oslo, Norway, 2020. Available online: http://hdl.handle.net/10852/79902 (accessed on 11 November 2022).
87. Templ, M.; Kowarik, A.; Meindl, B. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro. *J. Stat. Softw.* **2015**, *67*, 1–36. [CrossRef]
88. Anonimatron. Providing GDPR Compliance Since 2010. 2022. Available online: https://realrolfje.github.io/anonimatron/ (accessed on 6 Deceber 2022).

89. Johnson, N.; Near, J.P.; Hellerstein, J.M.; Song, D. Chorus: A programming framework for building scalable differential privacy mechanisms. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, 7–11 September 2020; IEEE: New York, NY, USA, 2020; pp. 535–551. [CrossRef]
90. esito. g9 Anonymizer-Database Anonymization Tool. 2022. Available online: https://www.esito.no/en/products/anonymizer/ (accessed on 6 December 2022).
91. UTD Anonymization Toolbox. 2022. Available online: http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php (accessed on 6 December 2022).
92. Sweeney, L. Guaranteeing anonymity when sharing medical data, the Datafly System. In *Proceedings of the AMIA Annual Fall Symposium*; American Medical Informatics Association: Washington, DC, USA, 1997; p. 51. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233452/ (accessed on 9 November 2022).
93. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Mondrian multidimensional k-anonymity. In Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 3–7 April 2006; IEEE: New York, NY, USA, 2006; p. 25. [CrossRef]
94. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Incognito: Efficient full-domain k-anonymity. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 49–60. [CrossRef]
95. Han, J.; Yu, H.; Yu, J. An improved l-diversity model for numerical sensitive attributes. In Proceedings of the 2008 Third International Conference on Communications and Networking in China, Hangzhou, China, 25–27 August 2008; IEEE: New York, NY, USA, 2008; pp. 938–943. [CrossRef]
96. Xiao, X.; Tao, Y. Anatomy: Simple and effective privacy preservation. In Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB Endowment, Seoul, Republic of Korea, 12–15 September 2006; pp. 139–150.
97. Xiao, X.; Wang, G.; Gehrke, J. Interactive anonymization of sensitive data. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, Providence, RI, USA, 29 June–2 July 2009; pp. 1051–1054. [CrossRef]
98. Maier, J. Anonymity: Formalisation of Privacy–k-anonymity. In Proceedings of the Seminars Future Internet (FI), Innovative Internet Technologies and Mobile Communications (IITM), and Autonomous Communication Networks (ACN), Seminar Paper, Technische Universität, Munich, Germany, 30 April–31 July 2013; pp. 41–48. Available online: https://www.net.in.tum.de/fileadmin/TUM/NET/NET-2013-08-1.pdf (accessed on 11 November 2022).
99. Dai, C.; Ghinita, G.; Bertino, E.; Byun, J.W.; Li, N. TIAMAT: A tool for interactive analysis of microdata anonymization techniques. *Proc. VLDB Endow.* **2009**, *2*, 1618–1621. [CrossRef]
100. Byun, J.W.; Kamra, A.; Bertino, E.; Li, N. Efficient k-anonymization using clustering techniques. In Proceedings of the International Conference on Database Systems for Advanced Applications, Bangkok, Thailand, 11–14 April 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 188–200. [CrossRef]
101. Poulis, G.; Gkoulalas-Divanis, A.; Loukides, G.; Skiadopoulos, S.; Tryfonopoulos, C. SECRETA: A system for evaluating and comparing relational and transaction anonymization algorithms. In Proceedings of the Advances in Database Technology—EDBT 2014, 17th International Conference on Extending Database Technology, Athens, Greece, 24–28 March 2014; pp. 620–623 . [CrossRef]
102. Poulis, G.; Loukides, G.; Gkoulalas-Divanis, A.; Skiadopoulos, S. Anonymizing data with relational and transaction attributes. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Prague, Czech Republic, 23–27 September 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 353–369. [CrossRef]
103. Fung, B.C.; Wang, K.; Yu, P.S. Top-down specialization for information and privacy preservation. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; IEEE: New York, NY, USA, 2005; pp. 205–216. [CrossRef]
104. Loukides, G.; Gkoulalas-Divanis, A.; Malin, B. COAT: Constraint-based anonymization of transactions. *Knowl. Inf. Syst.* **2011**, *28*, 251–282. [CrossRef]
105. Gkoulalas-Divanis, A.; Loukides, G. Utility-Guided Clustering-Based Transaction Data Anonymization. *Trans. Data Priv.* **2012**, *5*, 223–251. [CrossRef]
106. Terrovitis, M.; Mamoulis, N.; Kalnis, P. Local and global recoding methods for anonymizing set-valued data. *VLDB J.* **2011**, *20*, 83–106. [CrossRef]
107. NLM-Scrubber . 2022. Available online: https://lhncbc.nlm.nih.gov/scrubber/ (accessed on 12 June 2022).
108. Aircloak. Aircloak: Peace of Mind—Immediate Insights. 2022. Available online: https://aircloak.com/ (accessed on 6 December 2022).
109. CloverDX. CloverDX | Solve Demanding, Real-World Data Challenges. 2022. Available online: https://www.cloverdx.com/ (accessed on 6 December 2022).
110. BizDataX. BizDataX: Data Masking Done Right. 2022. Available online: https://bizdatax.com/ (accessed on 6 December 2022).
111. Gramener. Gramener: Data Science and AI Company. 2022. Available online: https://gramener.com/ (accessed on 6 December 2022).
112. Docbyte. Intelligent Document Processing Solution Anonymization. 2022. Available online: https://www.docbyte.com/solutions/anonymization/ (accessed on 6 December 2022).
113. Peppet, S.R. Regulating the internet of things: First steps toward managing discrimination, privacy, security and consent. *Tex. L. Rev.* **2014**, *93*, 85.
114. Buchmann, E.; Böhm, K.; Burghardt, T.; Kessler, S. Re-identification of smart meter data. *Pers. Ubiquitous Comput.* **2013**, *17*, 653–662. [CrossRef]

115. Freudiger, J. How talkative is your mobile device? An experimental study of Wi-Fi probe requests. In Proceedings of the 8th ACM Conference on Security & Privacy in Wireless and Mobile Networks, New York, NY, USA, 22–26 June 2015; pp. 1–6. [CrossRef]
116. Cunche, M.; Kaafar, M.A.; Boreli, R. I know who you will meet this evening! Linking wireless devices using wi-fi probe requests. In Proceedings of the 2012 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), San Francisco, CA, USA, 25–28 June 2012; IEEE: New York, NY, USA, 2012; pp. 1–9. [CrossRef]
117. Di Luzio, A.; Mei, A.; Stefa, J. Mind your probes: De-anonymization of large crowds through smartphone WiFi probe requests. In Proceedings of the IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, USA, 10–14 April 2016; IEEE: New York, NY, USA, 2016; pp. 1–9. [CrossRef]
118. Hong, H.; De Silva, G.D.; Chan, M.C. Crowdprobe: Non-invasive crowd monitoring with wi-fi probe. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–23. [CrossRef]
119. Hong, H.; Luo, C.; Chan, M.C. Socialprobe: Understanding social interaction through passive wifi monitoring. In Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, New York, NY, USA, 28 November–1 December 2016; pp. 94–103. [CrossRef]
120. Hern, A. Fitness tracking app Strava gives away location of secret US army bases. *Support Guard.* **2018**, *28*, 2018.
121. De Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1376. [CrossRef]
122. Potoczny-Jones, I.; Kenneally, E.; Ruffing, J. Encrypted Dataset Collaboration: Intelligent Privacy for Smart Cities. In Proceedings of the 2nd ACM/EIGSCC Symposium on Smart Cities and Communities, Portland, OR, USA, 101–12 September 2019; pp. 1–8. [CrossRef]
123. Rocher, L.; Hendrickx, J.M.; De Montjoye, Y.A. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **2019**, *10*, 3069. [CrossRef]
124. Schencker, L. How Much is Too Much to Tell Google? Privacy Lawsuit Allenges U. of C. Medical Center Went Too Far When Sharing Patient Data. Chicago Tribune. 2019. Available online: https://www.chicagotribune.com/business/ct-biz-lawsuit-university-of-chicago-google-patient-records-20190627-4vnmvfdnv5gcdl5fakgp5zwtna-story.html (accessed on 2 November 2022).
125. Annas, G.J. HIPAA regulations: A new era of medical-record privacy? *N. Engl. J. Med.* **2003**, *348*, 1486. [CrossRef]
126. Kalbo, N.; Mirsky, Y.; Shabtai, A.; Elovici, Y. The security of ip-based video surveillance systems. *Sensors* **2020**, *20*, 4806. [CrossRef]
127. *BS EN 62676-1-1:2014*; Video Surveillance Systems for Use in Security Applications. System Requirements. General. British Standard Institution: London, UK, 2014. [CrossRef]
128. *BS EN 62676-4:2015*; Video Surveillance Systems for Use in Security Applications. British Standard Institution: London, UK, 2015; pp. 1–82. [CrossRef]
129. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, *119*, 1–88.
130. Wang, H.; Gong, Y.; Ding, Y.; Tang, S.; Wang, Y. Privacy-Preserving Data Aggregation with Dynamic Billing in Fog-Based Smart Grid. *Appl. Sci.* **2023**, *13*, 748. [CrossRef]
131. Rushanan, M.; Rubin, A.D.; Kune, D.F.; Swanson, C.M. Sok: Security and privacy in implantable medical devices and body area networks. In Proceedings of the 2014 IEEE Symposium on Security and Privacy, Berkeley, CA, USA, 18–21 May 2014; IEEE: New York, NY, USA, 2014; pp. 524–539. [CrossRef]
132. Raij, A.; Ghosh, A.; Kumar, S.; Srivastava, M. Privacy risks emerging from the adoption of innocuous wearable sensors in the mobile environment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 7–12 May 2011; pp. 11–20. [CrossRef]
133. Ammari, T.; Kaye, J.; Tsai, J.Y.; Bentley, F. Music, search, and IoT: How people (really) use voice assistants. *ACM Trans. Comput.-Hum. Interact.* **2019**, *26*, 1–28. [CrossRef]
134. Lau, J.; Zimmerman, B.; Schaub, F. Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.* **2018**, *2*, 1–31. [CrossRef]
135. Purington, A.; Taft, J.G.; Sannon, S.; Bazarova, N.N.; Taylor, S.H. "Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, CO, USA, 6–11 May 2017; pp. 2853–2859. Available online: https://cpb-us-e1.wpmucdn.com/blogs.cornell.edu/dist/1/8892/files/2013/12/Alexa_CHI_Revise_Submit-22ay4kx.pdf (accessed on 12 December 2022).
136. Noda, K. Google Home: Smart speaker as environmental control unit. *Disabil. Rehabil. Assist. Technol.* **2018**, *13*, 674–675. [CrossRef]
137. Foxx, C. Apple Reveals HomePod Smart Speaker. *BBC News*, 5 June 2017. Available online: https://www.bbc.com/news/technology-40158158 (accessed on 12 December 2022).
138. Berger, A.A. The Amazon Echo. In *Perspectives on Everyday Life*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 79–82. [CrossRef]
139. Kastrenakes, J. Google Sold over 6 Million Home Speakers since Mid-October. 2018. Available online: https://www.theverge.com/2018/1/5/16855982/google-home-sales-figures-holidays-2017 (accessed on 11 December 2022).

140. Gartner Newsroom Gartner Says Worldwide Spending on VPA-Enabled Wireless Speakers Will Top $2 Billion by 2020. Gartner Newsroom. 2016. Available online: https://www.gartner.com/en/newsroom/press-releases/2016-10-03-gartner-says-worldwide-spending-on-vpa-enabled-wireless-speakers-will-top-2-billion-by-2020 (accessed on 16 November 2022).
141. Apthorpe, N.; Huang, D.Y.; Reisman, D.; Narayanan, A.; Feamster, N. Keeping the smart home private with smart (er) iot traffic shaping. *arXiv* **2018**, arXiv:1812.00955.
142. Ravi, S.; Mamdikar, M.R. A Review on ITS (Intelligent Transportation Systems) Technology. In Proceedings of the 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 9–11 May 2022; IEEE: New York, NY, USA, 2022; pp. 155–159. [CrossRef]
143. Tarnoff, P.J.; Bullock, D.M.; Young, S.E.; Wasson, J.; Ganig, N.; Sturdevant, J.R. Continuing Evolution of Travel Time Data Information Collection and Processing. Technical Report. Transportation Research Board. 2009. Available online: https://trid.trb.org/view/881513 (accessed on 16 December 2022).
144. Mohan, P.; Padmanabhan, V.N.; Ramjee, R. Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In Proceedings of the 6th ACM Conference on Embedded Network Sensor Systems, New York, NY, USA, 5–7 November 2008; pp. 323–336. [CrossRef]
145. Francesca, C.; Stefano, C.; Domenico, D.; Gambardella, S.M.; Giuseppe, P. Social network data analysis to highlight privacy threats in sharing data. *J. Big Data* **2022**, *9*, 19. [CrossRef]
146. Sun, W.; Chen, T.; Gong, N. SoK: Inference Attacks and Defenses in Human-Centered Wireless Sensing. *arXiv* **2022**, arXiv:2211.12087.
147. Stromire, G.; Potoczny-Jones, I. Empowering smart cities with strong cryptography for data privacy. In Proceedings of the 1st ACM/EIGSCC Symposium on Smart Cities and Communities, Portland, OR, USA, 20–22 June 2018; pp. 1–7. [CrossRef]
148. Wu, D.J.; Taly, A.; Shankar, A.; Boneh, D. Privacy, discovery, and authentication for the internet of things. In Proceedings of the European Symposium on Research in Computer Security, Heraklion, Greece, 26–30 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 301–319. [CrossRef]
149. Jawurek, M.; Johns, M.; Kerschbaum, F. Plug-in privacy for smart metering billing. In Proceedings of the International Symposium on Privacy Enhancing Technologies Symposium, Leuven, Belgium, 12–14 September 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 192–210. [CrossRef]
150. Rial, A.; Danezis, G. Privacy-preserving smart metering. In Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society, Chicago, IL, USA, 17 October 2011; pp. 49–60. [CrossRef]
151. Balasch, J.; Rial, A.; Troncoso, C.; Geuens, C.; Preneel, B.; Verbauwhede, I. PrETP: Privacy-Preserving Electronic Toll Pricing (extended version). In Proceedings of the 19th USENIX Security Symposium, Washington, DC, USA, 11–13 August 2010. Available online: https://www.usenix.org/legacy/event/sec10/tech/full_papers/Balasch.pdf (accessed on 8 November 2022).
152. Löbner, S.; Tronnier, F.; Pape, S.; Rannenberg, K. Comparison of de-identification techniques for privacy preserving data analysis in vehicular data sharing. In Proceedings of the Computer Science in Cars Symposium, New York, NY, USA, 30 November 2021; pp. 1–11. [CrossRef]
153. Sun, X.; Zhang, P.; Sookhak, M.; Yu, J.; Xie, W. Utilizing fully homomorphic encryption to implement secure medical computation in smart cities. *Pers. Ubiquitous Comput.* **2017**, *21*, 831–839. [CrossRef]
154. Wu, W.; Liu, E.; Gong, X.; Wang, R. Blockchain Based Zero-Knowledge Proof of Location in IoT. In Proceedings of the ICC 2020-2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–7. [CrossRef]
155. Han, L.; Luo, W.; Yang, A.; Zheng, Y.; Lu, R.; Lai, J.; Cheng, Y. Fully privacy-preserving location recommendation in outsourced environments. *Ad Hoc Netw.* **2023**, *141*, 103077. [CrossRef]
156. O'Keeffe, M. The Paillier Cryptosystem. Mathematics Department, 18 April 2008; pp. 1–16. Available online: https://www.cae.tntech.edu/~mmahmoud/teaching_files/grad/ECE7970/S16/slides/Homomorphic_basics.pdf (accessed on 7 November 2022).
157. Kapoor, S.R.; Jain, V.; Jain, R. A Privacy Preserving Repository For Data Integration Across Data Sharing Services. *Int. J. Eng. Res. Technol.* **2013**, *1*, 130–140. [CrossRef]
158. Federated learning approach to protect healthcare data over big data scenario. *Sustainability* **2022**, *14*, 2500. [CrossRef]
159. Gambs, S.; Killijian, M.O.; Núñez del Prado Cortez, M. De-anonymization attack on geolocated data. *J. Comput. Syst. Sci.* **2014**, *80*, 1597–1614. [CrossRef]
160. Truex, S.; Baracaldo, N.; Anwar, A.; Steinke, T.; Ludwig, H.; Zhang, R.; Zhou, Y. A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security, London, UK, 15 November 2019; pp. 1–11. [CrossRef]
161. Fang, H.; Qian, Q. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet* **2021**, *13*, 94. [CrossRef]
162. Nguyen, D.C.; Ding, M.; Pathirana, P.N.; Seneviratne, A.; Li, J.; Poor, H.V. Federated learning for internet of things: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 1622–1658. [CrossRef]
163. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 2031–2063. [CrossRef]
164. Li, Y.; Tao, X.; Zhang, X.; Liu, J.; Xu, J. Privacy-preserved federated learning for autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8423–8434. [CrossRef]

165. Wang, J.; Cai, Z.; Yu, J. Achieving personalized *k*-anonymity-based content privacy for autonomous vehicles in CPS. *IEEE Trans. Ind. Inform.* **2019**, *16*, 4242–4251. [CrossRef]
166. Maouche, M.; Ben Mokhtar, S.; Bouchenak, S. Hmc: Robust privacy protection of mobility data against multiple re-identification attacks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2018**, *2*, 1–25. [CrossRef]
167. Onesimu, J.A.; Karthikeyan, J.; Sei, Y. An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services. *Peer-Netw. Appl.* **2021**, *14*, 1629–1649. [CrossRef]
168. Onesimu, J.A.; Karthikeyan, J.; Eunice, J.; Pomplun, M.; Dang, H. Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access* **2022**, *10*, 86979–86997. [CrossRef]
169. Kaur, H.; Hooda, N.; Singh, H. k-anonymization of social network data using Neural Network and SVM: K-NeuroSVM. *J. Inf. Secur. Appl.* **2023**, *72*, 103382. [CrossRef]
170. Sei, Y.; Andrew, J.; Okumura, H.; Ohsuga, A. Privacy-preserving collaborative data collection and analysis with many missing values. *IEEE Trans. Dependable Secur. Comput.* **2022**, 1. [CrossRef]
171. Shankar, P.; Ganapathy, V.; Iftode, L. Privately querying location-based services with sybilquery. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 Spetember–3 October 2009; pp. 31–40. [CrossRef]
172. Primault, V.; Mokhtar, S.B.; Lauradoux, C.; Brunie, L. Time distortion anonymization for the publication of mobility data with high utility. In Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 20–22 August 2015; IEEE: New York, NY, USA, 2015; Volume 1, pp. 539–546. [CrossRef]
173. Bindschaedler, V.; Shokri, R. Synthesizing plausible privacy-preserving location traces. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2016; IEEE: New York, NY, USA, 2016; pp. 546–563. [CrossRef]
174. Yang, Y.; Ding, X.; Lu, H.; Weng, J.; Zhou, J. Self-blindable credential: Towards anonymous entity authentication upon resource constrained devices. In *Information Security*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 238–247. [CrossRef]
175. Khovratovich, D.; Law, J. Sovrin: Digital Identities in the Blockchain Era. Github Commit Jasonalaw. 17 October 2017. Available online: https://sovrin.org/wp-content/uploads/AnonCred-RWC.pdf (accessed on 6 December 2022).
176. Garcia-Alfaro, J.; Navarro-Arribas, G.; Hartenstein, H.; Herrera-Joancomartí, J. Data Privacy Management, Cryptocurrencies and Blockchain Technology. In Proceedings of the ESORICS 2021 International Workshops, DPM 2021 and CBT 2021, Darmstadt, Germany, 8 October 2021. [CrossRef]
177. Sousa, P.R.; Resende, J.S.; Martins, R.; Antunes, L. The case for blockchain in IoT identity management. *J. Enterp. Inf. Manag.* **2020**, *35*, 1477–1505. [CrossRef]
178. Bernabe, J.B.; Hernandez-Ramos, J.L.; Gomez, A.F.S. Holistic Privacy-Preserving Identity Management System for the Internet of Things. *Mob. Inf. Syst.* **2017**, *2017*, 6384186. [CrossRef]
179. Neven, G. IBM Identity Mixer (idemix). Presented at the NIST Meeting on Privacy Enhancing Technology, Zurich, Switzerland, 8–9 December 2011; pp. 8–9. Available online: https://csrc.nist.gov/csrc/media/events/meeting-on-privacy-enhancing-cryptography/documents/neven.pdf (accessed on 11 November 2022).
180. Camenisch, J.; Leenes, R.; Sommer, D. *Digital Privacy: PRIME-Privacy and Identity Management for Europe*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6545. [CrossRef]
181. Clauβ, S.; Kesdogan, D.; Kölsch, T. Privacy enhancing identity management: Protection against re-identification and profiling. In Proceedings of the 2005 Workshop on Digital Identity Management, New York, NY, USA, 11 November 2005; pp. 84–93. [CrossRef]
182. Eckhoff, D.; Sommer, C. Driving for big data? Privacy concerns in vehicular networking. *IEEE Secur. Priv.* **2014**, *12*, 77–79. [CrossRef]
183. Petit, J.; Schaub, F.; Feiri, M.; Kargl, F. Pseudonym schemes in vehicular networks: A survey. *IEEE Commun. Surv. Tutor.* **2014**, *17*, 228–255. [CrossRef]
184. Turan, F.; Roy, S.S.; Verbauwhede, I. HEAWS: An accelerator for homomorphic encryption on the Amazon AWS FPGA. *IEEE Trans. Comput.* **2020**, *69*, 1185–1196. [CrossRef]
185. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; IEEE: New York, NY, USA, 2017; pp. 3–18. [CrossRef]