IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 16, 2020, accepted January 3, 2021, date of publication January 6, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049599

Privacy in the Cloud: A Survey of Existing Solutions and Research Challenges

PAULO SILVA, EDMUNDO MONTEIRO, (Senior Member, IEEE), AND PAULO SIMÕES, (Senior Member, IEEE)

Centre for Informatics and Systems of the University of Coimbra, Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal

Corresponding author: Paulo Silva (pmgsilva@dei.uc.pt)

This work was supported by the POSEIDON H2020 Project under Grant 786713 H2020-DS-2016-2017/DS-08-2017.

ABSTRACT Private data is transmitted and stored online every second. Therefore, security and privacy assurances should be provided at all times. However, that is not always the case. Private information is often unwillingly collected, sold, or exposed, depriving data owners of their rightful privacy. In this article, various privacy threats, concepts, regulations, and personal data types are analyzed. An overview of *Privacy Enhancing Technologies* (PETs) and a survey of anonymization mechanisms, privacy tools, models, and metrics are presented together with an analysis of respective characteristics and capabilities. Moreover, this article analyses the applicability of the reviewed privacy mechanisms on today's Cloud Services and identifies the current research challenges to achieve higher privacy levels in the Cloud.

INDEX TERMS Anonymization, cloud computing, privacy metrics, privacy enhancing technologies.

I. INTRODUCTION

Over time different technologies and solutions have been proposed to secure users' information online or offline. These solutions range from privacy policies to security mechanisms, including encryption, authentication methods, anonymization techniques, laws, and regulations. All these solutions play an essential role in providing proper data privacy protection and security to users' information. Traditional authentication systems (e.g., password-based authentication) are among the most common and widely used methods of securing access to data, systems, databases, and services. The problem is that authentication systems can be subject to attacks or can fail. An example is the JPMorgan attack [1], which resulted in the exposure of personal information (e.g., names, addresses, and email) that compromised 87 million customers. Uber was also a target of an attack [2], and information about 57 million customers, as well as drivers, was compromised. These events could be minimized or eventually avoided if suitable privacy measures are adopted (e.g., encryption or reversible anonymization).

Encrypted data and communications should improve safety from attacks and eavesdropping. However, this is not always the case. There were situations where faults in the implementation of the encryption mechanisms (e.g., bugs) or man-in-the-middle attacks caused problems [3]. Other examples are brute-force attacks with online services or tools, such as FPGA or ASIC [4], plain text disclosure [5], and backdoors [6], which allowed illegitimate access to private data. There are options for increased security, such as Mayers' proposal of quantum cryptography [7], but this would compromise data utility (i.e. the usefulness of data after the application of PETs). Homomorphic encryption can be used to provide data utility and privacy. Using this method, it becomes possible to encrypt and still perform calculations and computations on data, therefore providing data utility (although at a limited scope). Moreover, it allows one to perform secure database search queries, which in many cases are translated into increased privacy. There are approaches, like the one proposed by Smart and Vercauteren [8], that use smaller keys and ciphertexts or, like the one proposed by Gentry *et al.*, that have simpler and faster implementation [9], but the performance is still an issue unless computational power is outsourced, as suggested by Mittal *et al.* [10]. There is a clear motivation to pursue improvements in this field. Nevertheless, in our article, the focus is on anonymization and privacy metrics, which should also provide a higher level of data utility against standard encryption.

The associate editor coordinating the review of this manuscript and approving it for publication was Jing Bi.

Since a decade ago, that data has been doubling every two years [11]. Such data is likely to contain sensitive *Personally Identifiable Information* (PII) in the most variate forms (e.g., biographical, technical, biological, and behavioral). Therefore, proper privacy mechanisms for handling sensitive information in such massive amounts of data are required. Data anonymization, pseudonymization, data minimization, and encryption are examples of such mechanisms. They are often denominated as *Privacy Enhancing Technologies* (PETs), but they are not the only resource available to enhance privacy. All the algorithms, tools, policies, and other mechanisms that provide privacy protection can be classified as PETs. This article stresses one of the ways of directly achieving information privacy: by performing data anonymization. These mechanisms may not only help to avoid (or to minimize) the problems mentioned above, but they also increase safety while publishing data (e.g., preserving privacy on publicly released data). Therefore, while this survey presents an overview of the different types of PETs, it particularly emphasizes data anonymization mechanisms, the related privacy metrics, and their applicability in Cloud contexts.

PETs and Privacy Metrics are many times associated with offline data and with the respective process of transforming and publishing data. Moreover, even anonymized data is prone to linkage attacks if not duly treated. A known example is when Sweeney showed that 97% of 54,805 voters were identifiable with their birth date and full postal code [12]. Moreover, Cloud Computing and the associated services and applications are every day more involved in our digital lives. The implications are significant, as massive amounts of data are being generated and held online every day. Therefore, data privacy should be a requirement and fundamental characteristic of offline processing and online services in the Cloud.

A survey on privacy-preserving data publishing by Fung *et al.* [13] presents a consistent review of anonymization algorithms, metrics, and different publishing scenarios. The authors provide examples of several cases and consider different data types. However, Cloud implications and related regulations are not covered. Wagner and Eckhoff's survey on privacy metrics [14] also details several metrics that can be used in the context of data privacy. However, considerations about the Cloud applicability are not sufficiently covered. A significant aspect of achieving privacy online is by assuring secure and anonymous communications. Shirazi *et al.* [15] provide insights on several anonymous communication protocols and systems. Nevertheless, the authors emphasize that security and anonymity are conflicting aspects, and there are trade-offs that are still an open research issue - especially between anonymity and performance.

Further literature [16]–[19] covers security and privacy aspects in Cloud environments. However, it tends to be focused on technical aspects and less inclusive. Due to the previously described reasons, we felt compelled to perform a comprehensive literature review for the privacy expert and, at the same time, accessible to the reader outside the specialty. The contributions of this article are the following:

1) a presentation and discussion of the different concepts related to privacy, including a review and discussion of privacy regulations, data privacy, and types of privacy threats;
2) a literature review of the most representative PETs with respect to anonymization options available to obtain data privacy, with a presentation of the different algorithms and models available, their operation, constraints, file types, and other relevant features;
3) since, to assess the work done on privacy algorithms and models, it is necessary to rely on privacy metrics, a review of the available privacy metrics, their operation domains, and their characteristics and a review of the privacy tools available to perform data anonymization;
4) an analysis of the Cloud applicability of the PETs and privacy metrics presented;
5) the identification and discussion of open issues and research challenges that need to be addressed in order to enhance privacy assurances in the Cloud.

The remainder of this article is organized as follows: Section II provides a background on privacy, privacy definitions, regulations, and threats, as well as privacy integration in the Cloud. PETs such as anonymization mechanisms and privacy models are presented and analyzed in Section III. Privacy metrics used to assess aspects such as privacy risk levels or data utility are presented and discussed in Section IV. Privacy tools supporting the application of PETs as well as privacy metrics are presented in Section V. Sections III, IV, and V also compare and discuss the Cloud applicability of PETs, privacy metrics, and privacy tools. Section VI highlights the current and future research challenges on the field. Final considerations are presented in Section VII.

II. BACKGROUND ON PRIVACY

Before the increase of Cloud Services, the Internet already had an abundance of services that required privacy mechanisms. As services related to Cloud Computing emerged and spread, privacy concerns were raised. That was due to the Cloud's intrinsic characteristics, such as distributed online storage, data replication, data integration, data regulation in different countries, privacy policies, and different types of threats. This section considers different privacy concepts, threats, and regulations and discusses specific privacy challenges in the Cloud.

A. PRIVACY CONCEPTS AND APPLICABILITY DOMAINS

Nowadays, the word *privacy* can be ambiguous and, therefore, more difficult to define accurately. There are several forms and definitions of privacy, none of them less relevant. In simple terms, to have privacy is to have the ability to control which personal information is known and used. Personal information is every piece of information that is related to an identifiable person. Nissenbaum, for instance, links privacy with contextual integrity such as a medical urgency episode

and social norms [20]. The following concepts represent common expressions and keywords used in the field:

- *Anonymization* – Daintith defines anonymization as ''a process that removes or replaces identity information from a communication or record'' [21]. For instance, a subject in communications or records can be made pseudonymous. The same subject will then always have the same replacement identity but cannot be identified.
- *Concealing* – Petitcolas *et al.* [22] state that concealing is the act of keeping from sight, to hide. By doing so, it means to keep something secret or prevent something from being known or noticed.
- *Data Confidentiality* – According to the Oxford dictionary [23], something confidential is ''intended to be kept secret,'' meaning that confidential information is the information intended to be kept secret. It can be seen as a set of rules that limit access or impose restrictions on certain types of information. Thereby, providing data confidentiality means keeping data secret.
- *Data Curator* – A data curator is an individual in charge of managing data. As Cragin *et al.* state: ''Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time'' [24].
- *Data Privacy* – Data privacy is the ability of an individual or group to stop information about themselves from becoming known to people, other than those whom they choose to give the information to. Privacy is sometimes related to anonymity, and Solove [25] considers that it is often most highly valued by people who have private data publicly known.
- *Data Utility (or Data Usability)* – After the anonymization process, there is the matter of the utility of the information, which is of high importance. Sweeney [26] considers utility or usability as the representational value of the amount of information preserved in the anonymized data.
- *De-identification* – De-identification is the process of identification, selection, and removal of sensitive information in a document or data set.
- *Observable Data* – The information that is available for a limited amount of time. In this case, an attacker might need to be present to observe or collect the data. Examples are communication systems where contents or intervening parties are actively or passively compromised.
- *Personally Identifiable Information* – Krishnamurthy and Wills [27] define PII as the information which can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linkable to a specific individual.
- *Published Data* – Published data is all the information willingly released and available to the public, considering all formats: databases, logs, traces, social network profiles, posts, and others.
- *Quasi-identifier* – A quasi-identifier is an attribute of the private information that can be linked with external data. Some identifiers, such as a person's name or address, are explicit. A quasi-identifier is an attribute that, combined with others, can identify individuals [26].
- *Re-Identification* – Re-identification is the name of the process that matches anonymized data with other datasets (publicly available or not). The matching process returns an estimate of the re-identification of records.
- *Risk of Disclosure* – There are a few variations (e.g., log-linear models or weight sampling) of the method to calculate the risk of disclosure. However, a common ground stated in Polettini [28] is based on the probability of a sampled record being re-identified. In other words, a record among the entire sample being identified.

It is also important to fully understand the kind of data to which these concepts can be applied. Many areas hold PII by default. Those areas include, but are not limited to, *health care*, *criminal*, *financial*, and *social* information. *Health care* information is one of the most sensitive types as it relates to an individual's health record. Blood samples, urine, *Deoxyribonucleic Acid* (DNA), and saliva test results are examples of health information as they relate with biological and genetic profiles, regardless of the origin. *Criminal*-related information can range from criminal records to court rulings, charges, convictions, speed tickets, *Driving Under the Influence* (DUI) of alcohol or drugs, and many other associated records. *Financial* information regards all information related to an individual's finances, such as salary, debt, mortgage, and other records such as bank accounts, credit and debit cards, bank extracts, loans, leases, and taxes. *Social* information includes, for instance, address, marital status, family, gender, sexual orientation, education, voter information or political preferences, location, and shopping habits.

The Cloud comprises an enormous amount of information stored online, somewhere, with no expiration date and often with no permanent deleting options. Along with all sorts of personal information or media like image and video stored in social networks or applications and web services, there are online communication services such as email. An example is a company processing email content to provide targeted advertising or personal assistant-related features. There are other aspects, such as shopping habits, product preferences, interaction and communication with others, and many others. What usually applies in most cases is that most online users leave a track, thus forming a digital fingerprint that can lead to complete or partial identification. Location, browser, search queries, visited websites, cookies, canvas, and window size are examples of data used to identify users.

Based on the previous concepts and respective applicability, information (i.e. data) is our focus. Nevertheless, PETs are also applicable in other domains that are beyond the scope of this article. We consider that *Architecture and Design*, *Communications and Networking*, *Data Information*, and *Identify*

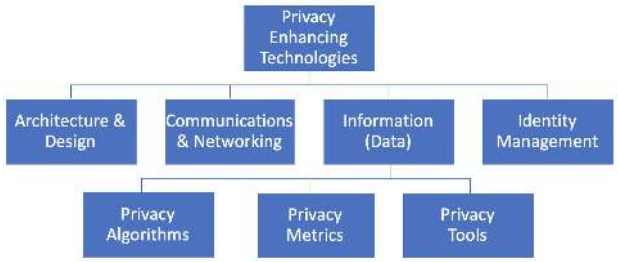

FIGURE 1. Proposed PETs applicability domains (emphasis on Information Privacy).

Management (depicted in Figure 1) is what best define other applicable areas for PETs. This is justified by the coverage that these four aspects provide: (1) the architecture and design of applications with privacy embedded by design; (2) providing secure networks and private communications; (3) keeping private the information available in the most variate data types; (4) keeping users' identity private. Therefore, following a development approach that applies state-of-the-art mechanisms and methodologies in these four application domains should result in a privacy-assuring product or service.

B. PRIVACY THREATS

An invasion of privacy occurs when personal information is used without consent or knowledge of the owner. It can happen through a data breach, attack, eavesdropping, or other forms of appropriation. According to Drake [29], Robison [30], and Thomson Reuters' Find Law [31], privacy threats (as shown in Figure 2) can be classified as follows:

- *Intrusion* – An intrusion of privacy includes all the actions that directly or indirectly invade an individual or organization's private affairs. Phone calls or conversations recorded without authorization and knowledge, taking pictures or trespassing on private property, repeatedly making non-requested phone calls, and spying on someone are examples of privacy intrusion.
- *Public Disclosure* – Releasing previously unknown or private information to the public is a public disclosure. This information can be offensive or embarrassing when publicly released. Therefore, if the data does not provide any public concern, the one(s) responsible for the release can be liable for privacy invasion. Typical examples that have their private information publicly disclosed are individuals in public offices, celebrities, and politicians.
- *False Light* – Similar to the previous point (public disclosure) is false light. It is a form of public disclosure of false or malicious statements. It is usually done by distorting the truth or using fictional facts.
- *Appropriation* – This case refers to the appropriation of an individual or organization's name or identity. It usually happens by using an individual's name, image, or any other personal characteristic without authorization or knowledge. It is common to see such media cases, references in books, stories, and marketing. Although it is possible to happen with any person, the issue is more recurrent with celebrities or famous personalities. In the digital era, this happens with online profiles or accounts as well.

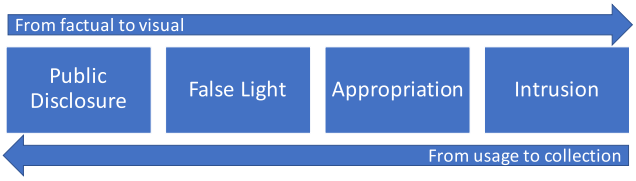

FIGURE 2. Privacy threats categories.

There are additional types of privacy threats. For instance, Solove [25] proposes a similar, yet more fine-grained taxonomy: information collection (e.g., surveillance), information processing (e.g., identification or re-identification), information dissemination (e.g., disclosure), and invasion (e.g., decisional interference). Other types of privacy invasion are attacks directed to data records. As defined by the *International Organization for Standardization* (ISO), an attack is an ''attempt to destroy, expose, alter, disable, steal, or gain unauthorized access to or make unauthorized use of anything that has value'' [39] to an individual or organization.

Within the data privacy scope, the overall consensus is that there are three different ways an attacker gathers information (i.e. attacker estimates) [40]–[42]. These attacker estimates are based on the type of information available to an attacker and the resemblance with other gathering information methods. The three main attacker estimates are as follows:

- *Prosecutor* – The attacker knows that data about the targeted individual is contained in the data set.
- *Journalist* – The attacker has no background knowledge.
- *Marketer* – The attacker is not interested in re-identifying just a specific individual.

It is also possible to enforce particular attack models that operate on specific data conditions. The attack models identified by Fung *et al.* [13] are the following:

- *Record Linkage* – This occurs when an attacker successfully matches a record owner to a sensitive attribute from datasets published or obtained elsewhere.
- *Attribute Linkage* – This occurs when there is no specific record identification, but the attacker can still infer sensitive values supported by the information of the group where the record owner belongs.
- *Table Linkage* – This occurs when attacks successfully derive the presence or the absence of the targeted record owner in a table.
- *Probabilistic Attack* – This is based on the uninformative principle from Machanavajjhala *et al.* [43]. Instead of focusing on actual records, it assures that the beliefs before and after accessing published data do not change significantly.

TABLE 1. Examples of privacy breaches and exploited threats.

Who?	What?	How?	When?	Threat Exploited	Source
AOL	Published search data led to identification of users	Cross referencing	2006	Public Disclosure	[32]
Netflix	Released data sets led to identification of users	Cross referencing	2006	Public Disclosure	[33]
Yahoo	500 million user accounts stolen	Hacking	2014	Intrusion and Appropriation	[34]
JPMorgan	87 million customer details exposed	Hacking	2014	Intrusion and Appropriation	[1]
Uber	57 million customer and driver details exposed	Hacking	2016	Intrusion and Appropriation	[2]
Equifax	Sensitive information of 140 million people	Hacking	2017	Intrusion and Appropriation	[35]
Cambridge Analytica	Unauthorized profiling	Personal data scraped from Facebook accounts	2018	Intrusion	[36]
Facebook	Over 540 million records exposed	Third-party security issues (Cultura Colectiva)	2019	Intrusion and Appropriation	[37]
Microsoft	Over 250 million records exposed	Security issues	2020	Intrusion and Appropriation	[38]

Table 1 provides examples of privacy invasions in which private data was exposed, and citizens' or organization's privacy was compromised. Political interests, credit card details, and addresses were publicly disclosed. In some cases, such as Yahoo and Uber, the data breaches happened due to security reasons. However, in other cases (such as Netflix or AOL), it was due to the incorrect usage of anonymization mechanisms. Cross-referencing or linkage attacks pose a significant risk for anonymized data. Nevertheless, the risk can be minimized or possibly avoided if proper anonymization mechanisms—ideally, a combination of mechanisms—are used, and attacker models are considered.

C. PRIVACY REGULATIONS

Many countries have laws and regulations regarding privacy, data access, data sharing, and handling. In Europe, some directives should be enforced and/or followed by the countries that are part of the *European Union* (EU). Moreover, with the *General Data Protection Regulation* (GDPR) enforcement in May 2018, any services or businesses handling data from European citizens are forced to comply with this regulation. In the *United States of America* (USA), the *Gramm-Leach-Bliley Act* (GLBA) [44], [45] is being enforced, while in Canada there is the *Personal Information Protection and Electronic Documents Act* (PIPEDA) [46]. To the East, there is the Russian Federation with its *Personal Data Protection Act* (PDPA) [47]. Regarding China, Greenleaf and Chen [48] show that, although there is no national privacy law enforced, the *Computer Processed Personal Data Protection Act* (CPPDPA) and *Personal Information Protection Act* (PIPA) are examples of regulations created for that effect.

Although regulations vary from country to country, they have a common objective: to provide legal protection and regulation over its citizens' personal and private information. The particularities of the regulations in the USA and Europe are analyzed next.

In the USA, different activity sectors (e.g., insurance, financial, and health care) have their own regulations.

- *Health Insurance Portability and Accountability Act* (HIPAA) is a health care regulation that assures that individuals' health information is properly protected while still (1) simplifying administrative processes by standardizing health care transactions and (2) reforming insurance conditions so that a job change does not affect coverage. Failure to comply with this regulation can result in fines up to $ 250K [49] and up to 10 years of jail time.
- *Gramm-Leach-Bliley Act (GLBA)* regulates how financial institutions manage financial information. Banks, insurance companies, securities firms, and even retailers must provide confidentiality about customers' credit information. Furthermore, according to the Federal Deposit Insurance Corporation [44] and the U.S. Code [45], these institutions must inform their customers of how their information is kept confidential and secure.
- The *Clarifying Lawful Overseas Use of Data* (CLOUD) Act regulates authorities' access to data held by American companies across the border of the USA. The act allows the *Department of Justice* (DOJ) data access without authorization from the courts or the Senate [50], [51].

While in the USA there is a sectoral approach for privacy regulation, in the EU the GDPR [52] regulates citizens' data privacy transversally with regard to all types of personal information. Some of the key points of the GDPR are as follows:

- *Territorial Applicability* is directed to all companies that process the personal data of European Union residents, regardless of the company's location.
- *Penalties* are up to 4% of annual sales volume or a maximum of € 20M. This penalty is applied in severe cases (for instance, lacking customer consent to process data).

- *Consent* requests must also be given in an easily accessible form. The purpose of data processing should also be present in the consent request.
- *Right to Access* intends to provide citizens with access to copies of all personal data held by a company. Furthermore, it is the right to know whether their data is being processed, the purpose, and the location.
- *Breach Notification* is a mandatory action (with a 72-hour limit) in cases where the data breach can pose a risk for the rights and freedom of citizens.
- *Right to be Forgotten* gives the right of having a citizen's data erased. It also has the potential to prevent data processing from third parties.
- *Data Portability* is the option that grants a citizen the right to receive and transmit his/her data.
- *Privacy by Design* is the inclusion of privacy and data protection mechanisms at each stage of development of a system or service, rather than addition. Companies such as Microsoft already adopt this principle when developing new products or services [53].
- *Data Protection Officers (DPOs)* are mandatory for those whose core activities consist of processing operations that require regular and systematic monitoring of data subjects on a large scale, particular categories of data, or data relating to criminal convictions and offenses.

Since 2000, there has been an agreement concerning privacy between the *European Commission* (EC) and the USA Government: the Safe Harbor agreement [54]. The primary purpose was to prevent and avoid accidental disclosures of personal information.

Despite the enforcement of such an agreement, after an EU citizen complained about Facebook's handling of his data, the agreement was declared invalid [55] by the European Court of Justice. After a modification of data collection terms between the USA and the EU, a new agreement was drafted: the EU-USA Privacy Shield. It is described as a framework for transatlantic exchanges of personal data for commercial purposes between the EU and the USA, and it is designed to accommodate the European regulations.

D. PRIVACY IN THE CLOUD

Cloud Services differ from more traditional Internet Services. The distributed data processing or the servers' location are aspects to consider in regard to privacy. All the aspects discussed in the previous sections should be suitable and adapted to the Cloud's context. The main requirements for privacy in the Cloud are the following:

- *Data Location* – Privacy laws and regulations differ from country and region. Therefore, compliance in different locations is a challenge. Companies processing data from international customers (e.g., European or American citizens) face some difficulties since the servers with databases and computing power might be distributed across different countries. There are at least two aspects to consider: local laws regarding the storage and management of customers data (e.g., GLBA) and laws regarding the country of origin of the customer's data (e.g., GDPR). Failure to comply might incur in significant losses (i.e. fines) for the companies in question.
- *Data Status* – Another aspect that Cloud Service Providers should consider is the disclosure of the methods used to protect data (e.g., the disclosed privacy policy). The status of the data during the stage(s) of processing or handling should be indicated (e.g., plain text, encrypted, anonymized, or pseudonymized).
- *Data Usage and Access* – It is necessary to assure proper handling and access to data at all times. A system or service might be compromised even if the best security measures and policies are in place. Ensuring proper data usage policies and (both physical and logical) access is sometimes not given due diligence. Suppose a more specific data processing is intended. In that case, it is recommended to disclose usage policies in two directions: user/customer to service provider, and service provider to user/customer (nevertheless, the latter predominates in most cases). Regarding data access, it is crucial to accurately define access rules. As such, a series of questions can be addressed: who can access, why, how, where, and for how long?
- *Security* – In addition to data status, there is a security point of view. In this case, infrastructure (e.g., an Intrusion Detection System), communications (e.g., a Secure Sockets Layer), and other security features play a crucial role in keeping data secure, regardless of the data state (e.g., plain text or encrypted). Common aspects such as establishing strong passwords, antivirus, and regular software updates can effectively increase security on both ends: users/customers and service providers.

FIGURE 3. Considerations for privacy in the Cloud.

As depicted in Figure 3, other requirements such as auditability, portability, and availability should also be considered. Nevertheless, there might still be vulnerabilities despite the Cloud providers' active measures to meet high privacy standards. Typically, data owners or users of such services have no physical control over the system. Therefore, instead of full trust, there is a semi-trust relationship. Nevertheless, in cases where Cloud Services are used for the single purpose of outsourcing data (i.e. data storage), users may take more proactive approaches such as anonymizing their data. For that purpose, several privacy algorithms and tools can limit the exposure of sensitive information.

In the next section, we review such algorithms and tools, as well as related privacy models.

III. PRIVACY ENHANCING TECHNOLOGIES

Privacy Enhancing Technologies perform data transformations or operations that result in increased data privacy levels (e.g., data anonymization or encryption). Data anonymization is the group of PETs that we emphasize in this article. This section presents an overview of PETs, anonymization mechanisms, privacy models, and a discussion of the applicability to the Cloud.

A. OVERVIEW OF PRIVACY TECHNOLOGIES

Several technologies are available in each applicability domain of PETs (cf. Figure 1). Nevertheless, a methodological approach concerning classification should be applied. The following models proposed by Hansen *et al.* [56] contain indicators for an evaluation based on the quality and readiness of each technology:

- *Quality assessment* – based on nine indicators with different weights (from the highest to lowest weight): Protection, Trust assumptions, Side effects, Reliability, Performance efficiency, Operability, Maintainability, Transferability, and Scope.
- *Readiness assessment* – six stages of the technology or mechanism: Idea, Research, Proof-of-concept, Pilot, Product, and Outdated.

PETs can be chosen according to the scores provided by the indicators mentioned above. These scores are useful to everyone interested in following a systematic approach to choose the right combination of PETs, for instance, software developers, online users, researchers, and data protection authorities.

Other types of PETs can be placed between end-users and services. For instance, the *The Onion Router* (TOR) browser and its network offer increased online privacy by anonymizing web traffic. Clark *et al.* compared several tools designed to be used in conjunction with TOR and concluded that none was satisfactory from a usability point of view [57]. TOR later provided a bundle that includes a modified version of the Firefox browser, which positively enhanced usability by offering a more straightforward way of keeping users' identity private. Nevertheless, Abbott *et al.* [58] showed that it is possible to identify TOR clients by performing browser attacks. Although it is difficult to attack TOR, Schneier [59] reported that the *National Security Agency* (NSA) was able to identify TOR users and attack computers by using private servers and a privileged position on the Internet's backbone.

In the Communications and Networking domain, encryption and even *Domain Name Server* (DNS) encryption (although with some drawbacks, such as transitive trust and traffic overhead [60]) are used to enhance privacy. Nevertheless, for increased data privacy, these technologies should complement each other whenever possible. For applicability in the Information domain, Zhou *et al.* [61] identified data anonymization and pseudonymization, and Pfitzmann and Hansen [62] identified data minimization.

Data publishing is a significant challenge for researchers or data protection authorities due to the trade-off between data utility and disclosure risk and the inherent risk of re-identification by cross-referencing or linkage attacks. These types of attacks rely on the information published elsewhere that might be matched with anonymized data sets, leading to individuals' identification. A systematic literature review of 14 studies about re-identification attacks on anonymized datasets [63] concluded the following: (1) 26% of all records were re-identified (with a *Confidence Interval* (CI) of 95%). (2) From the data anonymized (de-identified) *without* HIPAA standards, 33% of the health records were re-identified (CI of 95%). (3) From the only study with health records that used anonymization (de-identification) *following* HIPAA standards, only 1% of the records were re-identified (with a CI of 95%). These results indicate that enforcing anonymization mechanisms by itself does not suffice to provide low re-identification thresholds. A data curator must be experienced not only with anonymization mechanisms but also with the regulations available, which is also why we review privacy regulations (cf. Section II-C). Moreover, these results show that enforcing anonymization procedures following standards lead to a strong reduction of record re-identification. In a study about a hybrid anonymization heuristic [64], Mivule and Anderson aimed towards higher data utility. In the study, the *timestamp* and Internet Protocol *(IP) address* were anonymized using distinct methods, and the results demonstrated effective protection against re-identification. For the *timestamp*, enumeration and multiplicative noise were used to preserve the flow structure. For the IP addresses, a partial prefix-preserving heuristic was chosen. In the first octet, generalization is applied to preserve the prefix. A transformation is performed by multiplying or adding noise. k-anonymity is also applied to assure that no unique values appear and that the values appear $k > 1$ times. This way, the data retains usability by getting a synthetic flow of the IP address. Differential privacy was applied to the remaining octets.

Some technologies are more suitable than others to particular data types or applications (e.g., structured data, unstructured data, offline application, real-time, reversible, and non-reversible). According to the type of application,

data curators should consider characteristics such as suppressiveness (i.e. permanently suppressing or replacing original data), order preservation, the ability to operate on structured (e.g., tabular information) or unstructured data (e.g., text), and parameter tuning. Next, we discuss suppressive mechanisms.

B. SUPPRESSIVE ANONYMIZATION MECHANISMS

This section analyses mechanisms (mainly algorithms) that perform data transformations to enhance privacy. Since these transformations reduce or eliminate PII, the process results in increased data privacy. The transformations can provide specific data anonymization thresholds and assurances, depending on which methods are used. Datasets with specific characteristics can be generated by modifications that range from the slightest generalization or swapping of attributes to more elaborate operations and combinations of techniques. The algorithms and mechanisms described next permanently suppress or replace the original data.

Enumeration is defined by Goulden and Jackson [65] as the mapping of original values to new values, in a way that order is preserved. It is applicable in any well-ordered set. This algorithm is very straightforward. Consider for instance the following sequence: 3.6, 16.8, 21, 0.9, 27.9, 14.4. When the enumeration algorithm is applied, a sequence like 3, 12, 15, 3, 18, 9 is generated. The order is preserved, but any specific information is removed.

Slagell *et al.* [66] describe the *Black Marker* technique as going with a black marker on a piece of paper to conceal some field or part of the information. Though simple, this is a powerful algorithm in practical terms because replacing fields by *NULL* or 0 is an effective way of hiding sensitive information. However, this technique does not provide high usability levels (i.e. data utility) due to the information's deletion.

According to Slagell *et al.* [66], the difference between *truncation* [67] and the *black marker* is that the first shortens the values, while the latter maintains the length or structure of the privatized data. The transformation is performed by taking a data field and selecting a point after which all bits are removed. In the case of an email address (a string), it is possible to truncate the domain information (e.g., *john@domain.com* becoming *john*). When dealing with the binary values with fixed lengths, it is possible to choose a point after which the truncation is applied. With an IP address like 192.168.0.0, it is possible to apply right shifts until all bits to the right of the selected point are shifted to the end. This shift results in a truncated IP address such as 0.0.192.168.

In a usability awareness study [64], *precision degradation* is defined as a generalization technique that removes the most precise components of a specific field (e.g., replacing it by 0). With a timestamp, it is possible to have different precision levels (e.g., days, hours, minutes, seconds, and milliseconds). There is a higher precision with milliseconds and a lower precision with days or hours. A high precision degradation (e.g., milliseconds) applied to the time stamp 1000001001 would result in a less precise timestamp 1000001000.

Suppression [68] is similar to the black marker technique. This technique works by suppressing (i.e. deleting) sensitive fields from a data set, at a cellular level. There are different ways of suppressing a record. For instance, it is possible to delete the entire record (e.g., remove all the information of an attribute, either by removing it or replacing the field by zero or ''$*$'') or to suppress part of it (e.g., a zip code like 35684 can be suppressed as $35 * **$). Similar to the *black marker* technique, as suppression range increases, the data utility reduces accordingly. This technique is often used for generalization purposes such as precision degradation, k-anonymity adherence, and other similar mechanisms.

Time Unit Annihilation is mentioned by Slagell *et al.* [66] as a combination of the black marker technique with partitioning, for time and date fields. First, the values are dismantled into year, month, day, hour, minute, second, and millisecond. After this step, it is possible to annihilate any of the fields by replacing them with 0. It is possible to remove the time information (i.e. hour, minute, and second) and still have a valid date. The opposite can also occur: removing the date information (i.e. year, month, and day) and holding the time information.

Rastogi *et al.* [69] proposed the α β-*algorithm*, an anonymization algorithm on random insertions and deletions of tuples from a database. The algorithm considers the attacker's prior knowledge for its estimates. Adhering to a set of restrictions, the algorithm performs an enhanced type of noise addition and suppression. With a somewhat limited empirical study, the authors claim balanced privacy and utility tradeoffs. To the best of our knowledge, a practical implementation of the algorithm has not yet been released. The algorithm is also hard to evaluate given the limited study provided by this article, and the lack of publicly available implementation.

Anonimytext, developed by Pérez-Laínez *et al.* [70], is designed to ''de-identify sensible data from unstructured documents'' and preserve its structure. It is not an algorithm. Nevertheless, it is a system that follows a systematic approach to provide data anonymization. By using *Natural Language Processing* (NLP) and *Information Extraction* (IE) techniques, this method performs a semantic analysis of the documents (i.e. unstructured text) and creates tokens. Subsequently, with those tokens, it can detect sensitive information. This operation is supported by pre-loaded induction dictionaries and legal information (e.g., laws and regulations such as HIPAA). Finally, an expert reviews it and approves or rejects the suggested de-identification. This mechanism was validated on clinical notes (i.e. limited scope), and the results showed that, although it performed relatively well to de-identify patients' names, it showed little effect de-identifying medical facilities due to the ambiguity of the terms.

In the *partitioning technique* [71], the fields chosen to be anonymized are partitioned into meaningful sets. Afterward,

TABLE 2. Properties of suppressive anonymization and concealing mechanisms.

Mechanisms	Preserves Order	Structured Data	Unstructured Data	Parameter Tuning	Source
Enumeration	✓	✓	✗	✗	[65]
Black Marker	✓	✓	✓	✗	[66]
Truncation	✗	✓	✗	✗	[67]
P. Degradation	✗	✓	✗	✗	[64]
Suppression	✗	✓	✓	✓	[68]
Time Unit Ann.	✗	✓	✗	✓	[66]
α β-algorithm	✗	✓	✗	✓	[69]
Anonimytext	✓	✗	✓	✗	[70]
Partitioning	✗	✓	✗	✗	[71]
Permutation	✓	✓	✗	✗	[64]
Prefix-preserving	✗	✓	✗	✗	[72]
Hashing	✗	✓	✓	✗	[73]

the actual values are replaced with a fixed value from the same set. For instance, with TCP ports ($0-65535$), a possible solution could be to have the port numbers within 0 and 4095 replaced with a 0, and have the ports within the other set, 4096 and 65535, replaced with 65535. The black marker and truncation techniques are special cases of partitioning.

A study by Mivule and Anderson's [64] defines *permutation* as a one-to-one mapping of values. This method is useful when it is necessary to preserve the count or the order of the data sets without maintaining its value. There are several variations of permutation functions with different characteristics, such as performance or guarantees of non-collision or security. Nevertheless, a general feature in all permutation functions should be the difficulty in reversing it without knowing the parameters. For instance, using a hash function as a selection function while anonymizing an IP address can be dangerous if the hash function is known. The limitation, in this case, is that given the small space of IPv4 addresses, if additional parameters are not applied, the hash function could be retrieved through brute force.

Boschi and Trammell [72] proposed *prefix-preserving* as a particular type of permutation, due to the direct substitution technique that it enforces, with the restriction of having to preserve the structure of the value. For instance, considering two private IP addresses that match on a prefix of n bits, the two anonymized IP addresses that will be generated will match on a prefix of n bits as well. Therefore, the structure of subnets is preserved at each level while anonymizing IP addresses.

Hashing functions [73] can be useful for anonymization of both text and binary data. A hash function maps each value to a new value, not necessarily unique (as the permutation). Nevertheless, it has a limitation with binary data: truncating the result of a hash function to the shorter length of the value is often required. Consequently, the hash function is weaker and suitable for more collisions.

Table 2 summarizes the discussed suppressive mechanisms, considering properties such as order preservation, parameterization, and the ability to handle structured or unstructured data. Most methods are applied to structured files such as *Comma Separated Values* (CSV), *Extensible Markup Language* (XML), or other file types with organized information (e.g., tables or logs). Four mechanisms can preserve the order of data, and another three allow the tuning of its anonymization parameters.

C. NON-SUPPRESSIVE ANONYMIZATION MECHANISMS

This section analyzes mechanisms that perform data transformations to enhance privacy without suppressing or replacing the original data.

Mivule and Anderson [64] claim that *random time shifts* can be considered a particular case of permutation. When applied to time stamps, it adds a random offset (e.g., seconds or milliseconds) to every record within the timestamp attribute. As such, since all timestamps are shifted at once, an entire data set can be anonymized at once. The duration and the chronological order of the events are preserved in this technique. However, an attacker with external knowledge about the network traffic can easily revert the anonymization. Although a pseudorandom shift could avoid its reversal, it would also mean a loss of order preservation.

Differential Privacy, proposed by Dwork [74], tries to maximize query accuracy from a given dataset while minimizing the chances of identifying its records. There are similarities with the noise addition method. However, differential privacy performs anonymization by adding the Laplace transform to the data set queries. With this, it is not possible to distinguish if a particular value was modified or not. It is a method widely used across the industry (e.g., Microsoft or Apple) and with many application types. Due to this method's mathematical rigor and characteristics, the US Census Bureau adopted [81] differential privacy as the disclosure avoidance methodology of the 2020 census. Another application was demonstrated by Li *et al.* [82], which applied differential privacy to collaborative filtering (used in recommender systems) in such a way so that a two-party collaboration scheme can still be privacy-preserving. Despite being a state-of-the-art mechanism, it also has limitations. Since the Laplace transform accounts for outliers and influential observations, when the privacy is higher (i.e. by setting the epsilon parameter to a low value such as 0.0001), according to Fienberg *et al.* [83], data

TABLE 3. Properties of non-suppressive anonymization and concealing mechanisms.

Algorithms	Preserves Order	Structured Data	Unstructured Data	Parameter Tuning	Source
Rand. Time Shift	✓	✓	✗	✗	[64]
Diff. Privacy	✗	✓	✓	✓	[74]
Noise Addition	✗	✓	✓	✗	[75]
Mult. Noise	✗	✓	✓	✗	[76]
Log. Mult. Noise	✗	✓	✓	✗	[77]
Data Swapping	✗	✓	✗	✓	[78]
FRAPP	✗	✓	✗	✓	[79]
DNA-inspired	✗	✓	✓	✓	[80]

utility remains a challenge because the statistical properties change.

When dealing with noise perturbation, it is common to add or multiply values. *Noise Addition* [75] works by adding random values chosen between the mean and standard deviation of the original data. The random values are then added to the original data set's sensitive attribute values, providing confidentiality. *Multiplicative Noise* [76] is similar to noise addition; however, instead of adding the values to the original data, the random values are typically picked between the mean and the variance and afterward multiplied with the original data. *Logarithmic Multiplicative Noise* [77] is described as a variation of multiplicative noise. In this case, a logarithmic change is performed on the original data.

Exchanging sensitive cell values with other cell values (within the same attribute) is the basis of *data swapping* [78]. This data transformation technique preserves the underlying statistics and frequency of the data. In turn, this makes it difficult for an attacker to map the original values with the anonymized values. In this case, what needs to be considered is the selection of the swapping rate, the attributes to be swapped, and the respective candidate data records for the data swapping.

By using perturbation mechanisms characterized as random variables elements, Agrawal *et al.* [79] claim that FRAPP facilitates "a systematic approach to the design of perturbation mechanisms for privacy-preserving mining." FRAPP is designed to provide acceptable tradeoffs between privacy and utility. This high-accuracy privacy-preserving method intends to reduce mining classification errors for association rule mining and achieve a classification accuracy comparable to direct mining. The privacy improvements obtained by the usage of random variable elements have a marginal impact on accuracy. The mechanism registered 2.46% lower classification accuracy when compared to the direct classification on the original database.

Kencl and Loebl [80] developed a *DNA-inspired information concealing algorithm* able to conceal information based on the introduction and maintenance of families of repeats. There are two types of applications for this algorithm: weak concealing and strong concealing. The weak concealing method is intended for non-specific inputs where there is no outside knowledge about the likelihood or presence of some segments in the input. The strong concealing method is intended for all other applications. It comprises five procedures that operate on unstructured data sequences (e.g., characters and audio samples). The transformations add *dust* (a random part of the sequence itself) in the process. There are four procedures with a common pattern: (1) partition the input sequence into consecutive disjoint blocks; (2) in front of each block, add the terminal part of its predecessor (overlap); (3) add dust at the end of each block; (4) rearrange the blocks into an output sequence. The final results are documents, such as text files or audio tracks, with mathematically scrambled information. However, the generated files are from three times (using the weak concealing method) up to twenty-four times (using the strong concealing method) larger than the original files [84].

Table 3 summarizes the discussed non-suppressive mechanisms, considering order preservation, parameter tuning, and the ability to handle structured or unstructured data. Only Random Time Shift can preserve to order of the original data. All the mechanisms work over structured data, while Differential Privacy, Noise Addition, Multiplicative Noise, and the DNA-inspired information concealing algorithm also support unstructured files (e.g., files with email contents or notes) with text or even audio files.

D. PRIVACY MODELS

Privacy models are composed of rules and algorithms applied to data, resulting in transformations and operations that can be measured and quantified. Such measurements can lead to ambiguous interpretations of privacy metrics and privacy models. Literature also shows that some authors refer to privacy models as privacy metrics and vice versa. The following section describes what we consider to be privacy models.

Generalization [85] is a way of transforming a sensitive attribute into less specific information. There are several ways of performing these transformations. For instance, by attributing a single value to a group of sensitive fields such as a *ZIP code*, it is possible to group specific zip codes as part of a region or state. Instead of specifying whether an individual is a male or female, *gender* can be generalized, stating that it is a person. The same applies to *age* or any other numeric field with ranges or intervals that are suitable to use (e.g., "$<=$ 25" or 18-25).

The *k-anonymity* model was first proposed after Sweeney was able to identify a USA senator using ZIP codes and

voter information [26]. The method intends to provide a solution to the following challenge: ''given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful'' [26]. As such, *k-anonymity* provides data privacy by ensuring that the sensitive attributes are repeated k times, with k being always greater than one in order to provide confidentiality and make more difficult the identification of individual values. The algorithm relies on the combination of generalization and suppression methods to achieve *k-anonymity*. However, it does not guarantee data privacy against attackers with background knowledge.

The *MinGen* algorithm [86] was developed with a default condition of adhering to *k-anonymity* with a minimal generalization. The algorithm starts by creating a frequency list with the distinct sequences of values of the private data. Until there is a minimum of k or fewer tuples with distinct sequences of values in the frequency list, the generalization proceeds, making it a greedy algorithm. If there is some sequence of values occurring less than k times, it is suppressed. Nevertheless, since this method does not guarantee privacy against attackers with background knowledge (similarly to *k-anonymity*), other mechanisms like L-diversity or T-closeness (described next) were proposed so that this limitation could be reduced.

L-diversity was proposed by Machanavajjhala *et al.* [43] as an effort to overcome the drawbacks of attackers with background knowledge experienced in *k-anonymity* or *MinGen*. A certain block of information (e.g., a set of attributes of a table) is *l-diverse* if it provides at least l values of the sensitive attribute(s). An entire table is *l-diverse* if all blocks are *l-diverse*. However, despite overcoming the background knowledge attack problem, according to [87], if there are two or more sensitive attributes, then it is more difficult to apply *l-diversity* due to the additional dimensions.

Achieving l-diversity might be difficult. Moreover, even if it is reached, it may be insufficient to prevent disclosure. As such, *t-closeness* was proposed by Li *et al.* [88] to address the concerns related with *l-diversity* and *k-anonymity*. The authors state it ''requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e. the distance between the two distributions should be no more than a threshold t)''. Based on Aggarwal and Yu's findings [87], applying t-closeness on numerical attributes is far more effective than several other mechanisms.

The *LKC-privacy* model proposed by Mohammed *et al.* [89] is intended for anonymizing high-dimensional data where the LKC-privacy is formulated to ensure that all combinations of Quasi Identifiers with maximum length L in the data table T are shared by at least K records. Moreover, the confidence of inferring any sensitive values in S is, as the authors describe, ''not greater than C, where L, K, and C are thresholds, and S is a set of sensitive values specified by the data holder.'' Respectively, L is the maximum adversary's knowledge, K is the minimum anonymity, and C is the maximum confidence threshold.

Cao and Karras proposed the *β-likeness* model [90] for microdata anonymization, using a β threshold satisfaction rule with low utility loss. The authors claim that an attacker's confidence in a tuple of sensitive attributes is not higher after seeing published data. For that effect, it uses generalization methods and data perturbation. The combination of such methods claims to result in better privacy assurances than state-of-the-art t-closeness models and to be more efficient than similar approaches. The results that support the authors' claims only apply for categorical data as the applicability on numerical data is limited and vulnerable to proximity attacks. Nevertheless, according to Li *et al.* [91], an extension of this algorithm where the neighboring values are also perturbed would make these mechanisms immune to proximity attacks.

Seen as an extension of the k-anonymity model, the *p-sensitive* model proposed by Truta *et al.* [92] prevents disclosure of sensitive information by considering more than one sensitive attribute. Therefore, an anonymized dataset satisfies p-sensitivity if it satisfies k-anonymity and guarantees that each sensitive attribute's cardinality is at least P. Holding these two properties (k-anonymity and p-sensitive) provides further protection against homogeneity and background attacks. The results show that, as K increases, the number of disclosed attributed decreases. Therefore, p-sensitiveness and k-anonymity are only guaranteed with large P values.

Brickell and Shmatikov [93] define *δ-disclosure* considering a table as ''δ-disclosure private if the distribution of sensitive attribute values within each quasi-identifier class is roughly the same as their distribution in the entire table.'' With this model, it is also possible to calculate a gain in adversarial knowledge by relating δ with the information gain used by decision tree classifiers (e.g., ID3 and C4.5 by Quinlan [94]). In such a relation, there is a difference between the entropy of the set of sensitive attributes and the conditional entropy of quasi-identifiers and sensitive attributes. On the other hand, Cao and Karras [90] find that δ-disclosure's properties can become unnecessarily rigid and exceedingly lax at the same time. Moreover, they state that Brickell and Shmatikov [93] do not propose a mechanism (or tool) capable of applying δ-disclosure in real use cases.

Xiao and Tao [95] present *M-invariance* as a generalization method that aims at providing a strong level of privacy protection on re-published data. The authors mention that most methods rely on on-time publishing and cannot guarantee privacy assurances after deletions or insertions of records. This method's differentiating factor is the consideration of data re-publication, which is not respected by the well-known *k-anonymity* or *l-diversity* models.

Unlike most methods, which operate on data, *PrivAPP* [99] is an integrated approach for the design of privacy-aware applications that operates at the design level. It uses *Unified Modified Language* (UML) to introduce privacy in the design of applications. With the aim of systematizing privacy concepts for web applications and Cloud Services,

TABLE 4. Properties of the privacy models.

Models	Type	Range	Source
Generalization	Similarity / Diversity	-	[85]
K-anonymity	Similarity / Diversity	1 - x	[26]
MinGen	Similarity / Diversity	1 - x	[86]
L-diversity	Similarity / Diversity	0 - ∞	[43]
T-closeness	Similarity / Diversity	0 - ∞	[88]
LKC-privacy	Information Gain or Loss	1 - x	[89]
β-likeness	Similarity / Diversity	1 - x	[90]
P-sensitive	Similarity / Diversity	1 - x	[92]
δ-disclosure	Information Gain / Loss	0 - ∞	[93]
M-invariance	Similarity / Diversity	0 - ∞	[95]
κ-map	Similarity / Diversity	1 - x	[96]
δ-presence	Adversary Success Prob.	0 - 1	[97]
ϵ-indistinguishability	Indistinguishability	0 - 1	[98]

this approach uses a conceptual model, a referential architecture, and an extension of UML with a privacy profile. After validating the approach with an example bookstore, the authors claim this method is suitable for a model-based approach implementation.

Other models include, for instance, *κ-map* [96], *δ-presence* [97], and *ϵ-indistinguishability* [98]. All of them share an ultimate goal: modify data in such a way that private identities and information is protected but still able to provide useful information.

Table 4 shows that these privacy models rely on measures of distinguishability (e.g., similarity or diversity) and information gain (or loss). Overall, most models present a low or medium complexity and support structured data. Moreover, Kelly *et al.* applied such models on general datasets and concluded that they provide highly generic applicability characteristics [100].

E. PRIVACY TECHNOLOGIES FOR THE CLOUD

It is possible but not easy to modify and adapt the previously described mechanisms to apply them to the Cloud. Moreover, most of the previously described mechanisms rely on full access to the datasets and need to load all the data into memory. Nevertheless, other state-of-the-art PETs can be found in the literature, with some even taking advantage of *Artificial Intelligence* (AI) to perform their tasks. Such mechanisms are analyzed next.

Kohlmayer *et al.* [101] proposed a flexible approach to distribute data anonymization (more specifically, sensitive biomedical data). The authors rely on an encrypted global view of the dataset and then apply K-anonymity, L-diversity, T-closeness, and δ-presence. The global view is built using an *Secure Multiparty Computation* (SMC) protocol. In SMC protocols [102], [103], functions are computed and evaluated anonymously by different members of a group, with each member only getting to know its input and output. This allows members of a group to perform common computations using private data from each other while still ensuring privacy.

There are cases in which SMC can be adapted for different purposes. For instance, on-the-fly SMC on the Cloud via Multikey Fully Homomorphic Encryption [104] allows arbitrary computations on data in a non-interactive fashion. Another case is the usage of proactive SMC with a dishonest majority [105], addressing the possibility of members having been corrupted over a lifetime of secrets that should remain confidential for a long time (e.g., cryptographic keys). This mechanism uses a tradeoff between an adversary penetration rate and the resetting speed of other members to achieve such a feat.

In a *Zero Knowledge Proof* (ZKP) system [106], a party (a prover) can prove to another party (a verifier) that a given statement (or any form of data) is true, without disclosing any additional information. There are two kinds of ZKP systems: interactive [107] and non-interactive [108]. The interactive method requires the prover to perform a series of actions to the verifier. On the other hand, non-interactive systems offer a way for verifiers to perform the verification by themselves, thus not requiring actions to be repeated by the entity that owns or claims the truthfulness of a statement. This method can also be used to perform private queries to third parties. Despite the advantages, there are limitations such as the required computational power and the fact that the proof cannot be given with 100% certainty.

Shinde and Vishwa show how to preserve privacy by using a data partitioning technique for secure Cloud storage [109]. The proposed scheme divides data files into small blocks. In this way, the authors claim it provides security, integrity, and privacy.

Fu *et al.* [110] proposed a framework where, among others, tensorization (i.e. the mapping of lower-order data to higher-order data), Fourier transform, and homomorphic encryption are used to provide a privacy-preserving analysis of multimedia data in cloud environments. The authors show that their approach has a lower error ratio than similar approaches, such as those not applying secure tensorization.

Another approach is to perform anonymization techniques (e.g., permutation or truncation) before the data is outsourced to the Cloud. The difference here is the need to securely store the real values' mapping to the anonymized ones, which creates an overhead in storage. Companies such as Intel are following this approach and claiming success [111].

Other privacy-preserving methods provide data privacy and have minimal impact on the usability of a cloud service. For instance, a similarity-based method proposed by Pang and Shen [112] provides privacy assurances in text retrieval by anonymizing search results from authorized servers and preventing the reconstruction of queries and documents.

Homomorphic encryption, discussed by Mittal *et al.* [113], can be used to preserve privacy and still provide data utility because it allows computations to be performed on encrypted data. It also allows private database queries by taking advantage of its encrypted computation capabilities. Nevertheless, with homomorphic or any other type of encryption, a few aspects should always be considered. For instance, the *Cloud Security Alliance* (CSA) recommends the following guidelines:

- sensitive data should be encrypted with approved algorithms and long, random keys;
- data should be encrypted at all times: in transit, at rest, and in use;
- data should be encrypted before it passes from the enterprise to the cloud provider;
- the Cloud provider or service, as well as its staff, should never have access to the decryption keys used.

The development of AI services usually demands large amounts of (at times sensitive) data to produce capable *Machine Learning* (ML) models. Moreover, the training of such models often requires data to be sent to a centralized server, potentially jeopardizing data privacy. To optimize the training process, Konecný *et al.* [114], [115] proposed Federated Learning. This technology allows each device to contribute to the improvement of a shared model by training it with local data. The updates made are summarized and sent to the Cloud server, which aggregates the updates sent by different devices. In this way, as no private data leaves the device, higher privacy guarantees are offered.

Since its inception, Federated Learning is evolving, accommodating different application scenarios and having even stronger privacy focus. From Blockchain integration [116] to deployment in wireless networks [117], a large focus is on performance and privacy improvements [118]–[122]. For instance, Niu *et al.* [122] proposed a framework where clients download only portions of the model, train it locally, and then upload the renewed version. Thus, avoiding inefficient large-scale learning tasks for resource-constrained mobile devices. Moreover, they increased privacy assurances by coupling the framework with features like differential privacy, secure aggregation, randomized response, and a bloom filter.

In addition to the application of privacy algorithms and technologies, it is crucial to assess the privacy levels that each one of them can provide. Therefore, in the following section, we analyze privacy metrics.

IV. PRIVACY METRICS

When dealing with data privacy, it is evident that privacy metrics are required as they facilitate the assessment, quantification, and evaluation of privacy-enhancing processes and anonymized data sets. This section provides an overview of privacy metrics, a discussion and analysis of privacy metrics that can be applied for anonymization purposes, and the analysis of its applicability to the Cloud.

A. OVERVIEW OF PRIVACY METRICS

Metrics are quantitative assessment measures typically adopted across industries to assess and compare operations, performance, or any other measurable indicator. Several metrics can be used as privacy metrics due to its applicability in the data privacy domain. Information theory (e.g., Shannon's Entropy [123] or *Mutual Information* (MI)), descriptive statistics (e.g., percentage or average), and even advanced clustering algorithms (e.g., K-means or Davies Bouldin Index) provide many of those metrics. However, they do not represent all the options available.

There are population models such as Pitman's [124] or McNulty's [125] that estimate the characteristics of the overall population (i.e. dataset). They do that with probability distributions fine-tuned with sample characteristics. These models are particularly useful for determining disclosure and re-identification risks. Such models are part of statistical methods analyzed by Dankar *et al.* [126], which compared different models to estimate the number of population uniqueness accurately.

In a survey of technical privacy metrics by Wagner and Eckhoff [14], aspects such as data inputs, outputs, and types of data were addressed. For instance, uncertainty, similarity, diversity, indistinguishability or information gain, and loss are types of outputs that are attainable from privacy metrics. They also propose an informal method for choosing metrics based on data types, input sources, target audiences, and others. Nevertheless, it was also concluded that the combination and aggregation of privacy metrics are necessary.

Aside from data publishing and datasets, some metrics focus on information available on the Internet. By attempting to quantify how much of a user's information is online, Blauw and von Solms [134] show that it is possible to calculate users' visibility or invisibility scores incrementally. Nevertheless, the scores are subject to the classification and weight given to each layer of visibility, which can lead to subjective interpretations of the scores in different scenarios.

A similar work by Becker and Chen [135] targeted social media and the quantification of a user's privacy. A slightly different approach by Braunstein *et al.* [136] tried to infer privacy scores with privacy surveys. However, the authors discovered that the formulation of the surveys has a high impact on responses. Therefore, it was proposed to continue research on mapping indirect answers instead.

B. ANONYMIZATION METRICS

Information theory and *descriptive statistics* provide several metrics that can be used in contexts such as anonymization and privacy. Anonymization metrics are measures applied under specific conditions and datasets that provide the output required to estimate privacy scores.

The usage of *descriptive statistics* [137] is a general but effective way of estimating the amount of privacy granted to data or its usability. With this method, several measures can be taken to analyze the anonymized data. Mean, standard deviation, average, variance, covariance, and dispersion are some of the measures that can quantify the distortion between anonymized and original data.

Classification Error Metrics [127] are similar to descriptive statistics as they measure the classification error of an anonymized dataset and compare it to the original data. Mivule and Anderson [64] show that the difference between the two indicators presents a trade-off between data privacy and usability. In this case, both original and anonymized data

TABLE 5. Properties of the privacy metrics.

Metrics	Type	Structured Data	Unstructured Data	Range	Source
Class. Error Metrics	Error	✓	✓	-	[127]
Entropy	Uncertainty	✓	✓	0 - X	[123]
Mutual Information	Inf. Gain or Loss	✓	✓	0 - ∞	[128]
Person's Corr. Coeff.	Inf. Gain or Loss	✓	✓	(-1) - 1	[129]
Euclidean Distance	Error	✓	✓	0 - X	[130]
Davies Bouldin Index	Error	✓	✓	0 - X	[131]
Cosine Similarity	Similarity	✓	✓	0 - 1	[132]
Re-identification risk	Uncertainty	✓	✗	0 - 1	[133]

are passed through a machine learning algorithm that returns a classification error for original and anonymized data.

Shannon's Entropy is a metric proposed by Shannon [123] that is widely used in *Information Theory* (IT). It measures the amount of information in a particular block of information based on the data's uncertainty or randomness.

Mutual Information [128] is a statistical method that calculates the amount of shared information in two data sets. This metric is useful with anonymization processes (i.e. original vs. anonymized dataset). Using MI, there are several ways to improve the assumptions taken of the privatized data and, with the same principle, provide better anonymizations when the MI is used in the anonymization algorithm.

The correlation metric, also known as *Pearson's Correlation Coefficient* [129], measures the linear correlation between two data sets (the original data set and the anonymized one). Mivule and Anderson [64] show how it measures the correlation's direction, being positive or negative. This method returns values between -1 and 1. The signal indicates the direction of the correlation—positive if the data from the two data sets moves in the same direction and negative if it moves in the opposite direction.

When working with anonymization techniques, it is possible to implement clustering for anonymization performance purposes, for instance, k-means. In this situation, the *Euclidean Distance* [130] becomes handy in measuring the distances within the original and privatized cluster. Furthermore, with these results, it is possible to assess how well the anonymization went.

The *Davies Bouldin Index* [131] evaluates the quality of data clustering. It is similar to the Euclidean distance. It also quantifies how functional the clustering is. Furthermore, the resultant distance between clusters (and distances within the cluster) can be useful for further analysis (e.g., using Euclidean Distance).

The *Cosine Similarity* [132] is a function of the inner product of vectors (i.e. the representation of files or documents), divided by the product of their lengths. Usually applied in information retrieval, this function generates a normalized value between zero and one. The files being compared have the same information if the Cosine Similarity is one. On the other hand, the files are completely different if the Cosine Similarity is zero.

It is possible to assess the *risk of re-identification* [133] by relating the uniqueness of the records of a dataset. If uniqueness can be measured accurately, then this kind of risk (i.e. re-identification of disclosure) can be managed. Nevertheless, in practice, it is often not possible to measure uniqueness directly. Therefore, it must be estimated, for instance, as proposed by Dankar *et al.* [126].

Table 5 summarizes identified metrics, considering their types, the nature of the input data with which they operate (structured or unstructured), and the metrics' range. Apart from re-identification risk, all metrics can operate on structured and unstructured data.

C. PRIVACY METRICS FOR THE CLOUD

Only a few privacy metrics, in their original form, have characteristics suitable for Cloud concepts and services. The reason is that it is necessary to consider different data sources, instances, locations, life cycles, automation, and regulation. Another challenging step is to develop software capable of adapting and implementing such metrics in Cloud environments (cf. Section V-C). The following improvements may enhance their applicability in the Cloud:

- Setting sensitive attributes – As mentioned before, some metrics operate according to the sensitive attributes indicated by the data curator. Since a faulty identification of sensitive attributes or quasi-identifiers may increase the risk of re-identification, this is an issue that must be further researched and improved. Although still a challenge and object of research, such mechanisms could benefit from AI approaches to identify sensitive attributes and automate the anonymization process. For instance, cloud services that have run-time flows could benefit from these adaptations.
- Linkage and cross-referencing attacks – It is not possible to know with certainty what the background knowledge of a potential attacker is. Therefore, it is not possible to effectively avoid linkage and cross-referencing attacks. However, it is possible to make assumptions and estimates to raise privacy thresholds and reduce the risk of disclosure. Whenever possible, an exhaustive mapping of all the data sources spread across different Cloud services and providers (and respective measurement of the anonymization levels) can reduce the likelihood of such attacks.

- Data location and retention – Significantly related to the previous point, it is hard to deal with this aspect. Data curators may be unaware of other datasets available elsewhere. Therefore, it is necessary to make conservative assumptions and estimates for linkage attacks. Cloud service providers have different policies concerning data retention and location. Moreover, the scalable and dynamic nature of the processing, storage, and networking services highly affects the usage of traditional privacy metrics. Moreover, data curators applying such metrics must comply with enforced policies and regulations in the applicable regions.
- Application schemes and guidelines – There are application schemes and guidelines that help to provide uniformity in data privacy. There is also related work that provides insight on the topic and offers guidelines such as parameters, attackers, and data inputs, as shown by Wagner and Eckhoff [14]. Nevertheless, there is not yet a norm to follow. In the Cloud, one should not only comply with data privacy thresholds by anonymizing data, but also by probing the different services' components and modules to keep track of data and service changes and act accordingly.

Regardless of the purpose of data privacy operations and methods (e.g., compliance with regulation, data mining, security, and research), setting and defining sensitive attributes to anonymize, monitoring Cloud services' changes, estimating attackers' background knowledge, developing applications and services, and complying with regulation still face one common issue: the trade-off between privacy and utility. Nevertheless, efforts have been made in different fields. *Privacy by Design* (PbD) is taken into account while developing services and applications in Europe [52], while other methodologies are followed in different regions. Cloud service providers also have more privacy-driven policies. Moreover, research on PETs is active, and the scientific community keeps making progress in privacy metrics and methods.

Given the characteristics of Cloud environments, the previously discussed privacy metrics (Table 5) fail to cover all the relevant aspects. They mainly rely on data analysis and dismiss the surrounding factors. Nevertheless, they can still be used together with additional indicators. In cases where no specific data transformation occurs, but rather transactions, such as an exchange of private information between a user and a service provider, it is possible to devise metrics based on such service providers' properties. Such metrics as trustworthiness scores derived from appropriate indicators (e.g., security features), previous incident history (e.g., previous data breaches), and privacy mechanisms adopted (e.g., in communications, privacy policies, and transparency) can provide additional input for data owners sharing information with specific service providers. In turn, since data owners' privacy awareness increases, they can make appropriate decisions regarding their data.

Not all cloud services are identical, run in similar infrastructures, or employ the same security mechanisms. Therefore, the aforementioned trustworthiness scores are usually assessed from different perspectives. Sule *et al.* [138], for instance, propose fuzzy logic algorithms to assess trustworthiness levels based on characteristics from physical, infrastructure, platform, and software layers (e.g., *Secure Shell* (SSH), *Secure Sockets Layer* (SSL), *Intrusion Detection System* (IDS), *Virtual Machines* (VMs), and other characteristics). A downside of this approach is the computational cost associated with computing, collecting, and storing scores across nodes.

To overcome the previously described limitation, Zhang *et al.* [139] proposed a domain-based trust model that can reduce the aforementioned overhead by storing trust scores within the same domain and with trusted third-party nodes. The proposed model not only reduces computational overhead but also offers higher detection ratio of malicious nodes. Nevertheless, it depends on trustworthy nodes, and at some moment in time, they can be corrupted. As such, another work [140] further improved the approach by proposing a double-blind anonymous evaluation-based trust model that not only discards the usage of trusted third parties but also prevents malicious attacks in cloud computing.

The following section analyzes privacy tools. Some of the tools adopt the previously described privacy metrics, while others implement privacy models and respective privacy mechanisms and algorithms.

V. PRIVACY TOOLS

There is quite a choice of privacy software and tools available for different purposes. Tools are available not only to perform data anonymization operations but also to implement privacy metrics. They operate on different data types, data formats, and scenarios, thus providing different solutions for different needs. This section analyses the different types of anonymization tools and their characteristics and discusses their applicability in the Cloud.

A. ANONYMIZATION TOOLS

Anonymization tools are designed to provide the means to anonymize different datasets with different characteristics. These tools implement algorithms such as those described in Section III (e.g., k-anonymity and t-closeness). Examples of such tools and their characteristics are described next.

Open Anonymizer [141] is a Java tool designed to protect sensitive data with generalization. This feature, based on the concept of k-anonymity and l-diversity, allows for the creation of data twins that mask the identity of individuals.

AnonTool [142] is an open-source tool developed in C programming language that provides easy, flexible, and efficient functions that can be used to anonymize live traffic or packet traces in the *libpcap* file format. It supports several formats such as IP, *Transmission Control Protocol* (TCP), *User Datagram Protocol* (UDP), *Hyper Text Transfer Protocol* (HTTP), *File Transfer Protocol* (FTP), and Netflow.

To be able to anonymize a wider variety of data logs, new solutions were proposed. Li *et al.* [143] introduced Converter and ANonymizer for Investigating Netflow Events *(CANINE)*: an anonymization tool aiming at the privatization and conversion of different NetFlow[1] formats. The tool is coded in Java with a user-friendly *Graphical User Interface* (GUI), giving the possibility to click and choose the method to use, such as truncation, random permutation, or prefix-preserving.

With the intent of anonymizing *Process Accounting* (PA) logs, *SCRUB-PA* was proposed. It is one of the four modules of the SCRUB infrastructure. *SCRUB-tcpdump* is based and built on tcpdump. Yurcik *et al.* [144] designed it to provide the application of multi-level anonymization to packet traces, allowing for the management of packet traces while protecting sensitive information from being disclosed. *SCRUB-PA* is based on the Java code used in CANINE, and this module developed by Luo *et al.* [145] is intended for the anonymization of Process Accounting logs. Aiming for a similar outcome, Liu *et al.* [146] also show how business processes can be mined resorting to their proposed privacy-preserving framework. *SCRUB-NetFlows* is a NetFlow anonymization tool developed by Yurcik *et al.* [147] to fix the flaws found in previous tools and uses several options to anonymize the fields of standard NetFlows. *SCRUB-Alerts* anonymizes intrusion detection system alerts, for example, firewall or virus alerts.

Xu *et al.* [148] developed a tool that allows data curators to anonymize network traces by applying a prefix-preserving technique. *Crypto-PAn* is a cryptography-based method, where the data curators provide the tool with a secret key. With the same key, consistency is achieved in multiple network traces, which means that the same IP address in different traces is anonymized with the same resultant IP address. The algorithm uses bitwise anonymization, and the privatized IP addresses depend on previous anonymizations. Therefore, it leads to a security flaw. As the anonymized IP addresses share a common prefix with the private addresses, if one can de-anonymize one IP address, all the other addresses with the same prefix are affected. Nevertheless, in scenarios where injection attacks are not likely to happen, this option is preferable due to the maintenance of IP structures and the possibility of anonymization across different locations (with key sharing).

Kristoff [149] later proposed a Perl port derived from Crypto-PAn: *IPanonymous*. This tool provides a one-to-one mapping of the private to the anonymized IP address and supports prefix-preserving (cf. Section III), consistency across traces (over time and location), and cryptography-based anonymization. The logic is similar to that of Crypto-PAn, and it can provide consistency in the process. Using the same key will guarantee consistent results with different implementations.

[1]NetFlow is a feature that was introduced on Cisco routers that provides the ability to collect IP network traffic as it enters or exits an interface.

TCPdprive is a lightweight tool, developed by Farah [150], that anonymizes data by eliminating confidential information from packet traces collected from a network. It eliminates sensitive information by replacing sensitive fields with fabricated information, avoiding the reconstruction of sensitive information. It works on *tcpdump* '-w' files and supports different levels of anonymization, from 0 to 99, with 99 being the level where the information is released and 0 being the most secure level. Some of its limitations are, for instance, the system compatibility (SunOS, Solaris, and FreeBSD) or the non-preservation of subnet broadcast information.

Based on TCPdprive, Plonka [151] developed *Ip2anonIP* to turn IP addresses into hostnames or anonymous IP addresses. This tool provides the option of adding some arbitrary fields. However, it can take hours to prepare a data set of a single day.

B. ANONYMIZATION TOOLS THAT ALSO SUPPORT METRICS

The previously described anonymization tools focused on settings associated with each particular algorithm it implements. Nevertheless, some tools implement anonymization algorithms as well as privacy metrics that are not algorithm-dependent. As such, there is higher flexibility, and different metrics can be used to support the anonymization process. Such tools are analyzed next.

A modular command-line UNIX tool named *FLAIM* was proposed by Slagell *et al.* [66]. This framework primes for being particularly modular and not bound by specific types of logs to be anonymized. Data curators or the system administrators can also tune the trade-off between information loss and anonymization level. Since the framework includes several anonymization techniques such as truncation and prefix-preserving, it supports a broader range of applications.

Not all solutions are freeware or under open-source licenses. *PARAT* [154], rebranded to Privacy Analytics Eclipse, is a commercial solution for privacy and anonymization. It offers a solution similar to those referred to before (anonymization algorithms, risk, and data utility metrics) but focuses on medical data from a professional and commercial point of view.

Poulis *et al.* [155] proposed a system for evaluating the efficiency and effectiveness of anonymization algorithms: *SECRETA*. Having the possibility of choosing which algorithms to evaluate, the analysis is interactive and progressive. The results are displayed with the attribute statistics and several data utility indicators in a summarized and graphical form. It supports different operation modes and invokes one or more instances of the anonymization module with the specified algorithm and parameters. The evaluator module collects the anonymization results and forwards them to the experimentation module. From there, results are forwarded to the plotting module (for graphical visualization) or exported.

Datafly [156] is a tool developed to privatize medical records. It works by generalizing, suppressing, inserting, or removing information without losing the useful details

TABLE 6. Privacy tools characteristics.

Tools	Supports	GUI	Freeware	Metrics	Source
Open Anonymizer	JDBC	✓	✓	✗	[141]
AnonTool	IP, NetFlow, etc.	✗	✓	✗	[142]
IPanonymous	IP	✗	✓	✗	[149]
CANINE	NetFlow	✓	✓	✗	[143]
SCRUB-PA	PA logs	✓	✓	✗	[145]
Crypto-PAn	IP	✗	✓	✗	[148]
TCPdprive	tcpdump	✗	✓	✗	[150]
IP2anonIP	tcpdump	✗	✓	✗	[151]
ARX	CSV, Excel, JDBC	✓	✓	✓	[152]
sdcMicro	CSV, SPSS, R, rdf, etc.	✓	✓	✓	[153]
FLAIM	PA logs, tcpdump, NetFlow, netfilter	✗	✓	✓	[66]
PARAT	CSV, JDBC	✓	✗	✓	[154]
SECRETA	TXT, CSV	✓	✓	✓	[155]
Datafly	JDBC	✗	✓	✓	[156]
μ-Argus	CSV	✓	✓	✓	[157]
Utility-Aware Tool	CSV	✓	✗	✓	[158]

contained in the data and guaranteeing *k-anonymity*. To our knowledge, there is no publicly available implementation at the time of this writing.

Hundepool and Willenborg [157] designed *μ-Argus* for creating safe micro-data files. This tool implements a greedy algorithm developed by Sweeney [86]. First, the data curator specifies a value for k and selects how sensitive an attribute is within a range between 0 and 3 (not identifying and identifying, respectively). The rare (i.e. unsafe) combinations are detected by testing two and three combinations of attributes. Generalization and cell suppression are used to eliminate unsafe combinations of attributes. As μ-Argus suppresses data at the cell level, the output data usually contains all the tuples, but with deleted values in some cells. As it tests two and three combinations, the algorithm does not assure *k-anonymity* due to the possibility of finding unique combinations. As the *k-anonymity* requirement is not enforced on suppressed values, it becomes vulnerable to linking and inference attacks.

Operating slightly differently from most tools available, Wang *et al.* [158] proposed a *Utility-Aware Visual Approach for Anonymizing Multi-Attribute Tabular Data*. This long name describes an approach that combines multiple methods and models for data anonymization (e.g., k-anonymity, l-diversity, and t-closeness). It is different from most since it is a user-oriented solution that aims to help users identifying disclosure risks by proposing appropriate methods for the process. The authors use Privacy Exposure Risk Trees to visually guide users through the process.

The *ARX Data Anonymization Tool* is commonly mentioned in the literature and is widely used due to its features. First introduced by Prasser and Kohlmayer [152], it is capable of analyzing data utility and re-identification risks. Moreover, it supports various privacy models (such as *k-anonymity*, *l-diversity*, and *t-closeness*), semantic privacy models (e.g., differential privacy), data transformation techniques (e.g., generalization, suppression, and top/bottom coding), and global and local recoding.

sdcMicro (Statistical Disclosure Control Methods for Anonymization of Microdata and Risk Estimation) is an 'R' package[2] developed by Matthias *et al.* [153] to generate anonymized data. It also includes metrics and estimation processes, which provide better and more complete data analysis. It supports a wide variety of techniques to visualize and apply in the anonymization process. Compared with other tools such as μ-Argus, it supports a broader range of techniques.

Table 6 summarizes the privacy tools discussed, taking into account the inclusion or the lack of a supporting GUI, licensing costs, and the inclusion of privacy metrics.

C. PRIVACY TOOLS FOR THE CLOUD

Some of the previously analyzed privacy tools could be adapted to accommodate the Cloud's dynamics and characteristics. Others were designed from scratch for Cloud scenarios. Next, we analyze some of those tools.

Matsunaga *et al.* [160] proposed using a general anonymization policy in Cloud and Big Data platforms. The objective was to move towards an *Ontology-Based definition of Data Anonymization Policy for Cloud Computing and Big Data* in order to standardize the use of anonymization policies and share a universal agreement on data anonymization structure. It is one step further towards uniformity. By defining which attributes are sensitive, quasi-identifiers, and key attributes, the proposed ontology describes which methods (e.g., suppression or generalization) to use and what the associated regulation is (e.g., HIPAA or *Payment Card Industry Data Security Standards* (PCI-DSS)).

PRIVA as a Service (PRIVAaaS), a toolkit offering a set of libraries for providing privacy, was proposed by Basso *et al.* [161]. The toolkit works based on anonymization methods, such as generalization, suppression, and encryption, and policies, such as PIPEDA, GDPR, and HIPAA. The process is focused on three types of attributes: (1) key attributes (e.g.,

[2]R is a language and environment for statistical computing and graphics [159].

name and social security number), (2) quasi-identifiers (e.g., birth date and zip code), and (3) sensitive information (e.g., salary and credit card information).

There are also tools such as the *privacy-preserving framework for outsourcing location-based services to the cloud* [162], which are context-specific (i.e. location). Others, such as *PMDP* (A Framework for Preserving Multiparty Data Privacy in Cloud Computing) [163], focus on data shared across multiple servers. Nevertheless, although there are not many solutions designed for Cloud scenarios and applications, there is potential to improve existing privacy models and metrics. It is possible to pursue this goal by doing further research, updating existing software, and taking advantage of the Cloud paradigm—instead of disregarding current policies and regulations, adapting to Cloud and Big Data paradigms.

There are also security and performance aspects related to Cloud scenarios. Toch *et al.* [164] discussed privacy in the scope of cybersecurity. Such systems often demand detailed personal information from attack logs. This information sharing may put privacy and rights at stake if a proper balance is not found. It is important to weigh and study the pros and cons of giving away personal or private information to such systems in exchange for a promise of security and performance in the Cloud.

In the scope of data anonymization, ETL (Extract Transform Load [165]) processes used for anonymizing data can be computationally heavy. The aforementioned tools and metrics can benefit from the distributed computation. Moreover, data curators using these mechanisms as a service could rely on predefined regulations and policies offered by service providers.

VI. OPEN ISSUES AND RESEARCH CHALLENGES

In the previous sections, the characteristics of different privacy concepts, threats, PETs, metrics, and tools for the Cloud were presented. All the mentioned approaches possess a unique set of strengths and weaknesses. Moreover, their applicability to the Cloud also poses different challenges and open issues. Some of the open issues from algorithms, models, and infrastructure are scientific, while others such as laws and regulations are more political. Nevertheless, it is possible to group open issues and research directions into respective categories. Next, the open issues and the challenges concerning Cloud applicability are identified.

A. PRIVACY ALGORITHMS

The main purpose of using PETs and anonymizing data is data publishing. Extracting information from data is highly valuable, for instance, for marketing or research purposes. Therefore, protecting private data or PII is essential. In Sections III-D and V, we mentioned approaches, such as the ones proposed by Basso *et al.* [99] and Prasser and Kohlmayer [152], that focus on this point. Regardless of the improvements and novelty of such methods, it seems hard to reach an international consensus on the matter. The issues in these cases are closely related to insufficient cross-border legislation and policies. Next, we analyze the domains where we have identified open issues:

- Data publishing – Re-identification is a significant issue (and risk) in the scope of data publishing. By cross-referencing and linking information, attackers may gain access to private information. It is difficult for data curators to transform data while preventing the linkage with external information and the loss of data utility.
- Attribute classification – Usually, attributes are classified manually as sensitive, quasi-identifiers, or identifiable. It is a process that demands expert knowledge but still fails to avoid cross-referencing or linkage attacks. This also represents a barrier for automated operations suitable in Cloud environments. It is hard to define and classify all the attributes and, at the same time, prevent attacks and allow automation.
- Standardization – Closely related to regulations, there is preliminary work on standardization. There are various proposals for providing data privacy in the literature, for instance, data anonymization methods or enhanced encryption mechanisms. However, very few consider the Cloud perspective. Nevertheless, in both cases, there is a lack of consistency in standards.
- Data life cycle – Another issue to consider is the life cycle of data. Static datasets should gradually reduce with time, giving place to dynamic data. Several data updates, related datasets, different data sources, locations, and owners should be considered when privacy operations and assurances are to be provided. Cloud service providers and their respective *Data Protection Officers* (DPO) should be able to provide assurances. However, it is currently hard to define strategies and develop procedures that comprehensively cover such aspects.
- Security, anonymity, and performance – These concepts are conflicting. There are trade-offs between secure and anonymous communications, as well as anonymity and performance. When considering the three components in parallel, the difficulty in finding a proper balance rises even further. Proposing PETs that are, at the same time, secure, anonymous, fast, and efficient is a considerable research challenge.

B. PRIVACY METRICS

Although many metrics already have normalized formulations (e.g., Pearson's Correlation Coefficient [64] and Cosine Similarity [166]), this is an important aspect to consider in future research. By having normalized metrics, it is possible to perform objective analysis and allow the direct comparison and classification of datasets or systems. The search for efficient algorithms for optimal trade-offs between data privacy and data utility is ongoing. Furthermore, it seems only natural that recent developments in AI will contribute to the optimization of trade-offs and attribute classification.

Next, the domains where we have identified open issues are discussed:

- Anonymization versus data utility – This trade-off is a recurrent challenge. In data publishing, it is computationally hard to transform data while providing optimal data privacy and utility. There is a substantial number of possible combinations of data transformations. Thus, it would require tremendous amounts of computational power and memory to analyze all the outcomes and always achieve the best results. The research challenge here is on how to efficiently compute optimal trade-off values.
- Normalization – There is a large choice of metrics and models used to calculate privacy and risk thresholds. However, some metrics and models are not bound and therefore have subjective interpretations of classifications. For instance, l-diversity, t-closeness, and m-invariance range from zero to infinity. As such, subjective evaluations and classifications are possible.

C. PRIVACY TOOLS

The algorithms and methods used for data anonymization perform operations across the datasets. Many are stand-alone implementations, while others are part of more comprehensive solutions such as ARX [152] or sdcMicro [153]. Although possible, it is not easy to implement such algorithms and methods in ways that entirely take advantage of Cloud platforms. The domains where we have identified open issues are the following:

- Development, implementation, and access – PETs, privacy metrics, and models in Cloud environments are not common. Nevertheless, some improvements and enrichment of available solutions have appeared. Many originated in academic research and turned into further refined open-source projects such as ARX Data Anonymization Tool or sdcMicro. However, there are still issues to overcome, such as the low community-accessible offer of solutions beyond prototypes and demonstrations. For example, the LKC-privacy model and the $\alpha\beta$-algorithm are still being integrated with the ARX Data Anonymization Tool, and, like many others, it does not offer publicly known implementations.
- Probing services and data – Probing services or applications about privacy-related details is not an issue, but rather a development to be done as future work in most cases. Cloud services are often scattered across clusters in different servers and locations. With the variety of services running and processing potentially sensitive information, privacy monitoring tools must probe diverse and sometimes decentralized services, which can be challenging.
- Distributed deployment and computation – Cloud infrastructures horizontally and vertically scale instances of applications and services. Despite a few developments and prototypes (e.g., ARX Data Anonymization Tool), most privacy tools are not designed to support and take advantage of the Cloud's life cycle and overall architecture. Accordingly, it is still an issue to fully take advantage of Cloud architecture.
- Runtime Privacy – The trend of service and microservice containerization needs to be considered in runtime privacy. Cloud services and applications should be continuously monitored at runtime and upon deployment. Unless tailor-made probes are developed, it is a significant challenge to have adaptive mechanisms, due to the number and variety of available services and architectures.

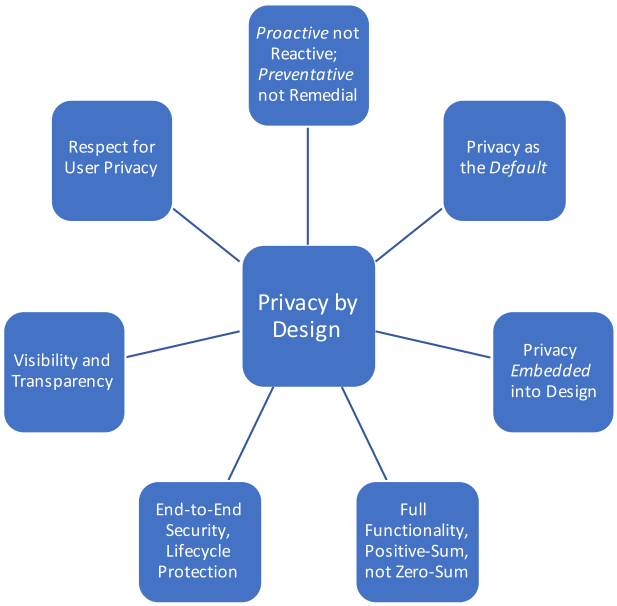

FIGURE 4. Seven principles of privacy by design.

D. PRIVACY BY DESIGN

The concept of Privacy by Design (cf. to [167] and Section VI-D) contributes to consistent privacy monitoring since specification and development aspects are taken into account in the assessment. In Europe, the longer-term enforcement of the GDPR is expected to bring further improvements. Nevertheless, it is necessary to further research privacy mechanisms for the Cloud, for instance, designing and developing adequate adaptive monitoring services.

PbD recommends designing systems accounting for the inclusion of privacy and data protection at each stage of development, rather than as an addition. Figure 4 shows the seven principles recommended to reach PbD. There are, however, several challenges when designing systems with privacy by design. For instance, as privacy can be ambiguous and fuzzy, it can be difficult to protect, there is no transversal methodology for systematic privacy enforcement (except the recently active GDPR [52] for European citizens), and there

is not much knowledge about the privacy benefits or risks practiced by the companies.

According to Shapiro [167], to have PbD, it is necessary to clearly define and produce privacy models and mechanisms early in each development context. As there are security engineers and technical security disciplines, having privacy disciplines and engineers would help to create privacy models and tools that are robust and systematically and transparently implementable.

Despite what was mentioned before, there is one factor that still considerably influences privacy: human behavior. Research conducted by Spiekermann on the behavioral economics of privacy [168] concluded that, regardless of the guidelines, people make irrational decisions and underestimate long-term privacy risks.

Companies have started to adopt such concepts after a period of formulating and consistently defining privacy policies and requirements. For instance, Microsoft adopted it when developing new products or services [53]. In other areas, such as data publishing, there was also room for improvement. For instance, Monreale *et al.* [169] have shown analytic processes such as mobility data publishing, distributed analytical systems, and *Global System for Mobile Communications* (GSM) profiling that minimize (or in some cases prevent) privacy harm and still achieve adequate trade-offs between privacy and utility.

Langheinrich [170] estimates that personal data collection is expected to continue throughout the years. With such conditions, with possible growth rates and data collection means, privacy might not be readily assured if proper measures are not taken. As such, social and technological contexts, as well as privacy laws and regulations, need to change and adapt. Moreover, the multitude of services, micro-services, and applications available in the Cloud are examples of ideal targets to employ privacy by design starting at the earliest development stages.

E. PRIVACY REGULATIONS

Although regulations are in place (e.g., HIPAA in Mercuri [49] or GDPR [52]), there is a lack of international cooperation in this scope. It is not easy to provide uniformity and consistency due to different legislation and politics in different regions. Nevertheless, aggregation of existing regulations and respective compliance should be possible. Next, the domains where we have identified open issues are discussed:

- The ambiguity of concepts and differentiation – Contextual integrity or social norms are examples of distinct privacy concepts, which can lead to ambiguity. Simultaneously, these different interpretations are not always subject to the scrutiny of regulations and policies.
- Conflicting regulations in different regions – HIPAA and GLBA in the USA, GDPR in Europe, PIPEDA in Canada, and CPPDPA in China are examples of distinct regulations across the world. When many Cloud services operate globally, this is a challenge for system developers and data curators, as data and operations are scattered across different regions.
- Data location – It is not always easy to keep track of data origin and location in Cloud services operating with several data centers worldwide. Since the applicability of existing regulations (e.g., GDPR) also depends on the country of origin of the data owner, data location is an important aspect to consider.

Despite the issues and challenges identified, improvements are being made across all of the areas mentioned. Every month, researchers publish completely novel research. Either by enhancing the shortcomings of existing methods or by following different approaches, the outcomes certainly add value to these areas.

The technical advancements, along with regulatory reforms, positively contribute towards more privacy. Nevertheless, while privacy increases, the difficulty for real-world implementation and regulation compliance also increases for many businesses. This difficulty is often a factor that is not given enough attention. For instance, companies that fail to implement such regulations are fined; business models that are incompatible with higher privacy standards (e.g., data collection in exchange for service) require profound changes in the services or products offered.

VII. CONCLUSION AND FINAL REMARKS

The massification of Cloud Services, along with increased privacy awareness, drives an increase in privacy concerns. Information is collected from various sources. Data is processed in many locations, and somewhat different laws regulate it. Therefore, it is relevant to survey the central privacy concepts currently adopted and the PETs, metrics, and tools available in the scope of data anonymization.

This survey provided a background on privacy concepts, threats, and regulations and analyzed its applicability to the Cloud. The functionalities and characteristics of PETs were discussed. The applicability of such mechanisms, tools, and privacy metrics in Cloud contexts was analyzed, and the current open issues and future research challenges were discussed.

Despite the diversity of algorithms, metrics, and tools, no solution fits all purposes. The combined use of such mechanisms and metrics should be the ultimate and ideal scenario. It is relevant to determine the characteristics of such mechanisms (to be able to use the most appropriate ones) and comply with different regulations from different regions. Our review is a contribution to guiding and assisting other privacy (or non-privacy) researchers. Moreover, it can help make informed choices about using PETs, metrics, and tools in the Cloud.

REFERENCES

[1] J. Kirk. (2014). *JP Morgan Chase Says Breach Affected 83M Customers*. Accessed: Nov. 2019. [Online]. Available: https://www.computerworld.com/article/2691246/jpmorgan-chase-says-breac%h-affected-83m-customers.html

[2] J. Dunn. (2018). *Uber Data Breach Aided by Lack Multi-Factor Authentication*. Accessed: Nov. 2019. [Online]. Available: https://nakedsecurity.sophos.com/2018/02/08/uber-data-breach-aided-by-m%ulti-factor-authentication-weakness/

[3] D. Lazar, H. Chen, X. Wang, and N. Zeldovich, "Why does cryptographic software fail?: A case study and open problems," in *Proc. 5th Asia–Pacific Workshop Syst.*, New York, NY, USA, 2014, p. 17, doi: 10.1145/2637166.2637237.

[4] (2019). *Cracking and Hacking Encryption Algorithms Sheds*. [Online]. Available: http://mycrypto.net/encryption/encryption_crack.html

[5] A. Yip, X. Wang, N. Zeldovich, and M. F. Kaashoek, "Improving application security with data flow assertions," in *Proc. 22nd Symp. Oper. Syst. Princ.*, New York, NY, USA, 2009, pp. 291–304, doi: 10.1145/1629575.1629604.

[6] T. Lee. (2013). *NSA-Proof Encryption Exists. Why Doesn't Anyone Use It?—The Washington Post*. [Online]. Available: https://www.washingtonpost.com/news/wonk/wp/2013/06/14/nsa-proof-encryp%tion-exists-why-doesnt-anyone-use-it/?utm_term=.c4368e9e9306

[7] D. Mayers, "Unconditional security in quantum cryptography," *J. ACM*, vol. 48, no. 3, pp. 351–406, May 2001, doi: 10.1145/382780.382781.

[8] N. P. Smart and F. Vercauteren, "Fully homomorphic encryption with relatively small key and ciphertext sizes," in *Public Key Cryptography*, P. Q. Nguyen and D. Pointcheval, Eds. Berlin, Germany: Springer, 2010, pp. 420–443.

[9] C. Gentry, A. Sahai, and B. Waters, "Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based," in *Advances in Cryptology*, R. Canetti and J. A. Garay, Eds. Berlin, Germany: Springer, 2013, pp. 75–92.

[10] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proc. 3rd ACM workshop Cloud Comput. Secur. Workshop*, New York, NY, USA, 2011, pp. 113–124, doi: 10.1145/2046660.2046682.

[11] J. Gantz and D. Reinsel. (2012). *The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. Accessed: Nov. 2019. [Online]. Available: https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe%-in-2020.pdf

[12] L. Sweeney, "Weaving technology and policy together to maintain confidentiality," *J. Law, Med. Ethics*, vol. 25, nos. 2–3, pp. 98–110, Jun. 1997. [Online]. Available: http://journals.sagepub.com/doi/10.1111/j.1748-720X.1997.tb01885.x

[13] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010, doi: 10.1145/1749603.1749605.

[14] I. Wagner and D. Eckhoff, "Technical privacy metrics: A systematic survey," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–38, Jun. 2018, doi: 10.1145/3168389.

[15] F. Shirazi, M. Simeonovski, M. R. Asghar, M. Backes, and C. Diaz, "A survey on routing in anonymous communication protocols," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 51:1–51:39, Jun. 2018, doi: 10.1145/3182658.

[16] J. Domingo-Ferrer, O. Farràs, J. Ribes-González, and D. Sánchez, "Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges," *Comput. Commun.*, vol. 140, pp. 38–60, May 2019.

[17] J. Tang, Y. Cui, Q. Li, K. Ren, J. Liu, and R. Buyya, "Ensuring security and privacy preservation for cloud data services," *ACM Comput. Surv.*, vol. 49, no. 1, pp. 1–39, Jul. 2016.

[18] S. Singh, Y.-S. Jeong, and J. H. Park, "A survey on cloud computing security: Issues, threats, and solutions," *J. Netw. Comput. Appl.*, vol. 75, pp. 200–222, Nov. 2016.

[19] N. Kaaniche and M. Laurent, "Data security and privacy preservation in cloud storage environments based on cryptographic mechanisms," *Comput. Commun.*, vol. 111, pp. 120–141, Oct. 2017.

[20] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum, "Privacy and contextual integrity: Framework and applications," in *Proc. IEEE Symp. Secur. Privacy*, Washington, DC, USA, 2006, pp. 184–198, doi: 10.1109/SP.2006.32.

[21] J. Daintith. (2019). *A Dictionary Computing*. Accessed: Nov. 2019. [Online]. Available: http://www.encyclopedia.com/doc/1O11-anonymization.html

[22] F. A. P. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding—A survey," *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, Jul. 1999.

[23] R. Wiles, G. Crow, S. Heath, and V. Charles, "The management of confidentiality and anonymity in social research," *Int. J. Social Res. Methodol.*, vol. 11, no. 5, pp. 417–428, Dec. 2008.

[24] M. Cragin, B. Heidorn, C. Palmer, and L. Smith. (2007). *An Educational Program on Data Curation*. Accessed: Nov. 2019. [Online]. Available: http://hdl.handle.net/2142/3493

[25] D. Solove, *Understand. Privacy*. Cambridge, MA, USA: Harvard University Press, 2009, vol. 173. [Online]. Available: https://books.google.pt/books?id=KCNPPgAACAAJ

[26] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, Oct. 2002.

[27] B. Krishnamurthy and C. E. Wills, "On the leakage of personally identifiable information via online social networks," in *Proc. 2nd ACM workshop Online Social Netw.*, New York, NY, USA, 2009, pp. 7–12, doi: 10.1145/1592665.1592668.

[28] S. Polettini. (2003) *A Note on The Individual Risk of Disclosure*. [Online]. Available: http://www3.istat.it/dati/pubbsci/contributi/Contributi/contr_2003/2003%_13.pdf

[29] J. Drake. (2019). *Four Types Invasion Privacy*. Accessed: Nov. 2019. [Online]. Available: https://legalbeagle.com/8068982-four-types-invasion-privacy.html

[30] W. L. Robison, *Digitizing Privacy*. Cham, Switzerland: Springer, 2018, pp. 189–204, doi: 10.1007/978-3-319-74639-5_13.

[31] F. L. T. Reuters. (2018). *Invasion Privacy*. Accessed: Nov. 2019. [Online]. Available: https://injury.findlaw.com/torts-and-personal-injuries/invasion-of-priv%acy.html

[32] D. Butler, "Data sharing threatens privacy," *Nature*, vol. 449, no. 7163, pp. 644–646, 2007.

[33] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," *CoRR*, vols. 5, p. 24, Dec. 2006. [Online]. Available: http://arxiv.org/abs/cs/0610105

[34] L. J. Trautman and P. C. Ormerod, "Corporate directors' and officers' cybersecurity standard of care: The yahoo data breach," *Am. UL Rev.*, vol. 66, p. 1231, Dec. 2016.

[35] A. J. Burns and E. Johnson, "The evolving cyberthreat to privacy," *IT Prof.*, vol. 20, no. 3, pp. 64–72, May 2018.

[36] J. Isaak and M. J. Hanna, "User data privacy: Facebook, cambridge analytica, and privacy protection," *Computer*, vol. 51, no. 8, pp. 56–59, Aug. 2018.

[37] UpHuard. (2019). *Losing Face: Two More Cases of Third-Party Facebook App Data Exposure*. Accessed: Aug. 2020. [Online]. Available: https://www.upguard.com/breaches/facebook-user-data-leak

[38] C. Cimpanu. (2020). *Microsoft Discloses Security Breach of Customer Support Database*. Accessed: Aug. 2020. [Online]. Available: https://www.zdnet.com/article/microsoft-discloses-security-breach-of-cu%stomer-support-database/

[39] *Plastics—Determination of Fracture Toughness—Linear Elastic Fracture Mechanics (LEFM) Approach*, International Organization for Standardization, Geneva, CH, Standard, Sep. 2009. [Online]. Available: http://standards.iso.org/ittf/PubliclyAvailableStandards/c041933_ISO_IE%C_27000_2009.zip

[40] E. Gachanga, M. Kimwele, and L. Nderu, "Sensitivity based anonymization with multi-dimensional mixed generalization," in *Proc. 13th Int. Conf. Digit. Inf. Manage. (ICDIM)*, Piscataway, NJ, USA, Sep. 2018, pp. 168–172.

[41] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "An improved differential privacy algorithm to protect re-identification of data," in *Proc. IEEE Canada Int. Hum. Technol. Conf. (IHTC)*, Toronto, ON, Canada, Jul. 2017, pp. 133–138.

[42] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim, "Data synthesis based on generative adversarial networks," *Proc. VLDB Endowment*, vol. 11, no. 10, pp. 1071–1083, Jun. 2018.

[43] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L -diversity: Privacy beyond k -anonymity," *ACM Trans. Knowl. Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.

[44] Federal Deposit Insurance Corporation. (2019) *FDIC: Privacy Act Issues under Gramm-Leach-Bliley*. [Online]. Available: https://www.fdic.gov/consumers/consumer/alerts/glba.html

[45] U. Code. (1999). *Title V of the Gramm-Leach-Bliley Act's (GLBA)*. Accessed: Nov. 2019. [Online]. Available: https://www.gpo.gov/fdsys/pkg/USCODE-2011-title15/pdf/USCODE-2011-title%15-chap94-subchapI.pdf

[46] A. Singh and K. Chatterjee, "Cloud security issues and challenges: A survey," *J. Netw. Comput. Appl.*, vol. 79, pp. 88–115, Feb. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804516302983?via%%3Dihub
[47] M. Sergey. (2018). *Data Protection in Russian Federation: Overview—Practical Law*. Accessed Nov. 2019. [Online]. Available: https://uk.practicallaw.thomsonreuters.com/2-502-2227__lrTS=2018041913%0106547&transitionType=Default&contextData=(sc.Default)
[48] G. Greenleaf and H.-L. Chen. (May 2012). *Data Privacy Enforcement in Taiwan, Macau, and China*. Accessed: Nov. 2019. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2118332
[49] R. T. Mercuri, "The HIPAA-potamus in health care data security," *Commun. ACM*, vol. 47, no. 7, pp. 25–28, Jul. 2004, doi: 10.1145/1005817.1005840.
[50] Senate of the United States. (2018). *Clarifying Lawful Overseas Use of Data Act or the CLOUD Act*. Accessed: Nov. 2019. [Online]. Available: https://www.hatch.senate.gov/public/_cache/files/6ba62ebd-52ca-4cf8-9bd%0-818a953448f7/ALB18102(1).pdf
[51] M. Moon. (2018). *President Signs Overseas Data Access Bill Into Law*. Accessed: Nov. 2019. [Online]. Available: https://www.engadget.com/2018/03/24/cloud-act-law/
[52] E. Parliament and E. U. Council, "Regulation (EU) 2016/ 679 OF the European parliament and of the council - of 27 apr. 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 (gen," *Off. J. Eur. Union*, vol. 59, p. 88, Dec. 2016. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R067%9&qid=1490179745294&from=en
[53] Microsoft. (2014). *Protecting Data Privacy Cloud*. Accessed: Nov. 2019. http://download.microsoft.com/download/2/0/a/20a1529e-65cb-4266-8651-1b%57b0e42daa/protecting-data-and-privacy-in-the-cloud.pdf
[54] U.S. Department of Commerce. (2000). *Safe Harbor Privacy Principles*. Accessed: Nov. 2019. [Online]. Available: http://webarchive.loc.gov/all/20150410181019/
[55] S. Gibbs. (2015). *The Guardian—What is 'Safe Harbour' Why Did EUCJ Just Declare it Invalid?*. Accessed: Nov. 2019. [Online]. Available: https://www.theguardian.com/technology/2015/oct/06/safe-harbour-europea%n-court-declare-invalid-data-protection
[56] M. Hansen, J.-H. Hoepman, and M. Jensen, "Readiness Analysis for the Adoption and Evolution of Privacy Enhancing Technologies," European Union Agency For Network And Information Security, Heraklion, Greece, Tech. Rep., 2015. [Online]. Available: www.enisa.europa.eu, doi: 10.2824/614444.
[57] J. Clark, P. C. van Oorschot, and C. Adams, "Usability of anonymous Web browsing: An examination of tor interfaces and deployability," in *Proc. 3rd Symp. Usable Privacy Secur.*, New York, NY, USA, 2007, pp. 41–51, doi: 10.1145/1280680.1280687.
[58] T. G. Abbott, K. J. Lai, M. R. Lieberman, and E. C. Price, "Browser-based attacks on tor," in *Privacy Enhancing Technology*, N. Borisov and P. Golle, Eds. Berlin, Germany: Springer, 2007, pp. 184–199.
[59] B. Schneier. (2013). *Attacking Tor: How the NSA Targets Users' Online Anonymity|US News|The Guardian*. Accessed: Nov. 2019. [Online]. Available: https://www.theguardian.com/world/2013/oct/04/tor-attacks-nsa-users-onl%ine-anonymity
[60] H. Shulman, "Pretty bad privacy: Pitfalls of DNS encryption," in *Proc. 13th Workshop Privacy Electron. Soc.*, New York, NY, USA, Nov. 2014, pp. 191–200, doi: 10.1145/2665943.2665959.
[61] B. Zhou, J. Pei, and W. Luk, "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM SIGKDD Explor. Newslett.*, vol. 10, no. 2, p. 12, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1540276.1540279
[62] A. Pfitzmann and M. Hansen, "A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management," *Tech. Univ. Dresden*, vol. 34, pp. 1–98, Dec. 2010. [Online]. Available: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml%5Cnhttp://dud.inf.t%u-dresden.de/literatur/Anon_Terminology_v0.34.pdf
[63] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," *PLoS ONE*, vol. 6, no. 12, Dec. 2011, Art. no. e28071, doi: 10.1371/journal.pone.0028071.
[64] K. Mivule and B. Anderson, "A study of usability-aware network trace anonymization," in *Proc. Sci. Inf. Conf. (SAI)*, London, U.K., Jul. 2015, pp. 1293–1304.
[65] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*. Phoenix, AZ, USA: Courier Corporation, 2004.
[66] A. Slagell, K. Lakkaraju, and K. Luo, "Flaim: A multi-level anonymization framework for computer and network logs," in *Proc. 20th Conf. Large Installation Syst. Admin.*, Berkeley, CA, USA, 2006, pp. 63–67. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267793.1267799
[67] Y. Smamash, "Truncation method of reduction: A viable alternative," *Electron. Lett.*, vol. 17, no. 2, pp. 97–99, Jan. 1981.
[68] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Inf. Sci.*, vol. 231, pp. 83–97, May 2013.
[69] V. Rastogi, D. Suciu, and S. Hong, "The boundary between privacy and utility in data publishing," in *Proc. 33rd Int. Conf. Very Large Data Bases*, Vienna, Austria, 2007, pp. 531–542. [Online]. Available: http://dl.acm.org/citation.cfm?id=1325851.1325913
[70] R. Pérez-Laínez, C. De Pablo-Sánchez, and A. M. Iglesias, "Anonimytext: Anonimization of unstructured documents," in *Proc. Int. Conf. Knowl. Discovery Inf. Retr.*, Setubal, Portugal, 2008, pp. 284–287.
[71] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar, "On the privacy preserving properties of random data perturbation techniques," in *Proc. 3rd IEEE Int. Conf. Data Mining*, Washington, DC, USA, 2003, p. 99. [Online]. Available: http://dl.acm.org/citation.cfm?id=951949.952160
[72] E. Boschi and B. Trammell, *Ip Flow Anonymization Support*, document RFC 6235, May 2011.
[73] J. L. Carter and M. N. Wegman, "Universal classes of hash functions," *J. Comput. Syst. Sci.*, vol. 18, no. 2, pp. 143–154, Apr. 1979.
[74] C. Dwork, "Differential Privacy: A Survey of Results," in *Theory and Applications of Models of Computation*. Berlin, Germany: Springer, 2008, pp. 1–19. [Online]. Available: http://link.springer.com/10.1007/978-3-540-79228-4_1
[75] J. Domingo-Ferrer, F. Sebé, and J. Castellà-Roca, "On the security of noise addition for privacy in statistical databases," in *Privacy in Statistical Databases*, J. Domingo-Ferrer and V. Torra, Eds. Berlin, Germany: Springer 2004, pp. 149–161.
[76] K. Chen and L. Liu, *A Survey of Multiplicative Perturbation for Privacy-Preserving DataMining*. Boston, MA, USA: Springer, 2008, pp. 157–181, doi: 10.1007/978-0-387-70992-5_7.
[77] J. J. Kim, J. J. Kim, W. E. Winkler, and W. E. Winkler, "Multiplicative noise for masking continuous data," Statistical Res. Division, US Bureau of the Census, Washington, DC, USA, Tech. Rep. 2003-01, 2003.
[78] T. Dalenius and S. P. Reiss, "Data-swapping: A technique for disclosure control," *J. Stat. Planning Inference*, vol. 6, no. 1, pp. 73–85, Jan. 1982.
[79] S. Agrawal, J. R. Haritsa, and B. A. Prakash, "FRAPP: A framework for high-accuracy privacy-preserving mining," *Data Mining Knowl. Discovery*, vol. 18, no. 1, pp. 101–139, Feb. 2009, doi: 10.1007/s10618-008-0119-9.
[80] L. Kencl and M. Loebl, "DNA-inspired information concealing: A survey," *Comput. Sci. Rev.*, vol. 4, no. 4, pp. 251–262, Nov. 2010.
[81] C. Dwork, "Differential privacy and the US census," in *Proc. 38th ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, New York, NY, USA, 2019, p. 1, doi: 10.1145/3294052.3322188.
[82] J. Li, J.-J. Yang, Y. Zhao, B. Liu, M. Zhou, J. Bi, and Q. Wang, "Enforcing differential privacy for shared collaborative filtering," *IEEE Access*, vol. 5, pp. 35–49, 2017.
[83] S. E. Fienberg, A. Rinaldo, and X. Yang, "Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables," in *Proc. Int. Conf. Privacy Stat. Databases*. Cham, Switzerland: Springer, 2010, pp. 187–199.
[84] P. Silva, L. Kencl, and E. Monteiro, "Data privacy protection—Concealing text and audio with a dna-inspired algorithm," in *Proc. 12th Int. Conf. Auto. Infrastruct., Manage. Secur.*, B. Stiller, Ed. Munich, Germany: IFIP, Jun. 2018, pp. 46–59. [Online]. Available: http://www.aims-conference.org/2018/AIMS-2018-Proceedings.pdf#page=54
[85] V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Alberta, GA, Canada: ACM, 2002, pp. 279–288.
[86] L. Sweeney, "Achieving K-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 571–588, Oct. 2002.
[87] C. C. Aggarwal and P. S. Yu, *A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. Boston, MA, USA: Springer, 2008, pp. 11–52, doi: 10.1007/978-0-387-70992-5_2.

[88] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond K-anonymity and L-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Piscataway, NJ, USA, Apr. 2007, pp. 106–115.

[89] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C.-K. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 4, p. 18:1–18:33, Oct. 2010, doi: 10.1145/1857947.1857950.

[90] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *Proc. VLDB Endowment*, vol. 5, no. 11, pp. 1388–1399, Jul. 2012, doi: 10.14778/2350229.2350255.

[91] J. Li, Y. Tao, and X. Xiao, "Preservation of proximity privacy in publishing numerical sensitive data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2008, pp. 473–486.

[92] T. M. Truta, A. Campan, and P. Meyer, "Generating microdata with P-sensitive K-anonymity property," in *Secure Data Manage.* Berlin, Germany: Springer, 2007, pp. 124–141. [Online]. Available: http://link.springer.com/10.1007/978-3-540-75248-6_9

[93] J. Brickell and V. Shmatikov, "The cost of privacy," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, New York, USA, 2008, p. 70. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1401890.1401904

[94] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986. Online]. Available: http://link.springer.com/10.1007/BF00116251

[95] X. Xiao and Y. Tao, "M-invariance: Towards privacy preserving re-publication of dynamic datasets," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2007, pp. 689–700, doi: 10.1145/1247480.1247556.

[96] K. El Emam and F. K. Dankar, "Protecting privacy using k-anonymity," *J. Amer. Med. Inform. Assoc.*, vol. 15, no. 5, pp. 627–637, Sep. 2008. http://academic.oup.com/jamia/article/15/5/627/732733/Protecting-Privac%y-Using-kAnonymity

[97] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*. New York, NY, USA, 2007, p. 665. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1247480.1247554

[98] P. M. Aonghusa and D. J. Leith, "Don’t Let Google know i’m lonely," *ACM Trans. Priv. Secur.*, vol. 19, no. 1, p. 3:1–3:25, Aug. 2016, doi: 10.1145/2937754.

[99] T. Basso, L. Montecchi, R. Moraes, M. Jino, and A. Bondavalli, "PrivAPP: An integrated approach for the design of privacy-aware applications," *Softw., Pract. Exper.*, vol. 48, no. 3, pp. 499–527, Mar. 2018. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/spe.2546/full

[100] D. J. Kelly, R. A. Raines, M. R. Grimaila, R. O. Baldwin, and B. E. Mullins, "A survey of state-of-the-art in anonymity metrics," in *Proc. 1st ACM workshop Netw. Data Anonymization*, New York, NY, USA, 2008, pp. 31–40, doi: 10.1145/1456441.1456453.

[101] F. Kohlmayer, F. Prasser, C. Eckert, and K. A. Kuhn, "A flexible approach to distributed data anonymization," *J. Biomed. Informat.*, vol. 50, pp. 62–76, Aug. 2014. https://www.sciencedirect.com/science/article/pii/S1532046413001937?via%

[102] Y. Lindell, "Secure multiparty computation for privacy preserving data mining," in *Encyclopedia Data Warehousing Mining*. Hershey, PA, USA: IGI Global, 2005, pp. 1005–1009.

[103] P. Bogetoft, D. L. Christensen, I. Damgård, M. Geisler, T. Jakobsen, M. Krøigaard, J. D. Nielsen, J. B. Nielsen, K. Nielsen, J. Pagter, M. Schwartzbach, and T. Toft, "Secure multiparty computation Goes live," in *Financial Cryptography Data Security*, R. Dingledine and P. Golle, Eds. Berlin, Germany: Springer, 2009, pp. 325–343.

[104] A. López-Alt, E. Tromer, and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," in *Proc. 44th Symp. Theory Comput.*, New York, NY, USA, 2012, pp. 1219–1234, doi: 10.1145/2213977.2214086.

[105] K. Eldefrawy, R. Ostrovsky, S. Park, and M. Yung, "Proactive secure multiparty computation with a dishonest majority," in *Security and Cryptography for Networks*, D. Catalano and R. De Prisco, Eds. Cham, Switzerland: Springer, 2018, pp. 200–215.

[106] S. Goldwasser, S. Micali, and C. Rackoff, "The knowledge complexity of interactive proof-systems," in *Proc. 17th Annu. ACM Symp. Theory Comput.*, New York, NY, USA, 1985, pp. 291–304, doi: 10.1145/22145.22178.

[107] F. Li and B. McMillin, "Chapter two—A survey on zero-knowledge proofs," in *Advances in Computers*, vol. 94, A. Hurson, Ed. Missouri, Amsterdam, The Netherlands: Elsevier, 2014, pp. 25–69. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128001615000025

[108] H. Wu and F. Wang, "A survey of noninteractive zero knowledge proof system and its applications," *Sci. World J.*, vol. 2014, May 2014, Art. no. 560484. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24883407

[109] Y. Shinde and A. Vishwa, "Privacy preserving using data partitioning technique for secure cloud storage," *Int. J. Comput. Appl.*, vol. 116, no. 16, pp. 24–27, Apr. 2015.

[110] C. Fu, Z. Yang, X.-Y. Liu, J. Yang, A. Walid, and L. T. Yang, "Secure tensor decomposition for heterogeneous multimedia data in cloud computing," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 247–260, Feb. 2020.

[111] J. Sedayao and I. I. Enterprise Architect, "Enhancing cloud security using data anonymization," Intel Corporation, Mountain View, CA, USA, Tech. Rep., 2012. https://www.intel.co.kr/content/dam/www/public/us/en/documents/best-pra%ctices/enhancing-cloud-security-using-data-anonymization.pdf

[112] H. Pang, J. Shen, and R. Krishnan, "Privacy-preserving similarity-based text retrieval," *ACM Trans. Internet Technol.*, vol. 10, no. 1, pp. 1–39, Feb. 2010, doi: 10.1145/1667067.1667071.

[113] D. Mittal, D. Kaur, and A. Aggarwal, "Secure data mining in cloud using homomorphic encryption," in *Proc. IEEE Int. Conf. Cloud Comput. Emerg. Markets (CCEM)*, Bangalore, India, Oct. 2014, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/7015496/

[114] J. Koneäný, B. McMahan, and D. Ramage, "Federated optimization:Distributed optimization beyond the datacenter," 2015, *arXiv:1511.03575*. [Online]. Available: http://arxiv.org/abs/1511.03575

[115] J. Koneäný, H. Brendan McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*. [Online]. Available: http://arxiv.org/abs/1610.02527

[116] M. H. ur Rehman, K. Salah, E. Damiani, and D. Svetinovic, "Towards blockchain-based reputation-aware federated learning," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Jul. 2020, pp. 183–188.

[117] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, early access, Oct. 1, 2020, doi: 10.1109/TWC.2020.3024629.

[118] C. Fang, Y. Guo, N. Wang, and A. Ju, "Highly efficient federated learning with strong privacy preservation in cloud computing," *Comput. Secur.*, vol. 96, Sep. 2020, Art. no. 101889. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404820301620

[119] M. Hao, H. Li, X. Luo, G. Xu, H. Yang, and S. Liu, "Efficient and privacy-enhanced federated learning for industrial artificial intelligence," *IEEE Trans. Ind. Informat.*, vol. 16, no. 10, pp. 6532–6542, Oct. 2020.

[120] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 911–926, 2020.

[121] R. Hu, Y. Guo, H. Li, Q. Pei, and Y. Gong, "Personalized federated learning with differential privacy," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9530–9539, Oct. 2020.

[122] C. Niu, F. Wu, S. Tang, L. Hua, R. Jia, C. Lv, Z. Wu, and G. Chen, "Billion-scale federated learning on mobile clients: A submodel design with tunable privacy," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2020, doi: 10.1145/3372224.3419188.

[123] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.

[124] N. Hoshino, "Applying pitman's sampling formula to microdata disclosure risk assessment," *J. Off. Statist.*, vol. 17, no. 4, p. 499, 2001.

[125] G. Chen and S. Keller-McNulty, "Estimation of identification disclosure risk in microdata," *J. Off. Statist.*, vol. 14, no. 1, p. 79, 1998.

[126] F. K. Dankar, K. El Emam, A. Neisa, and T. Roffey, "Estimating the re-identification risk of clinical data sets," *BMC Med. Informat. Decis. Making*, vol. 12, no. 1, p. 66, Jul. 2012, doi: 10.1186/1472-6947-12-66.

[127] T. N. Phyu, "Survey of classification techniques in data mining," in *Proc. Int. Multi Conf. Eng. Comput. Sci.*, Hong Kong, vol. 1, 2009, pp. 18–20.

[128] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, no. 6, Jun. 2004, Art. no. 066138.

[129] P. Sedgwick, "Pearson's correlation coefficient," *Bmj*, vol. 345, p. e4483, Dec. 2012.
[130] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 1, pp. 40–48, Nov. 2010, doi: 10.1145/1882471.1882478.
[131] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–1, no. 2, pp. 224–227, Apr. 1979.
[132] L. Muflikhah and B. Baharudin, "Document clustering using concept space and cosine similarity measurement," in *Int. Conf. Comput. Technol. Develop.*, Kuala Lumpur Malaysia, vol. 1, 2009, pp. 58–62.
[133] C. Skinner and D. J. Holmes, "Estimating the re-identification risk per record in microdata," *J. Off. Statist.*, vol. 14, no. 4, p. 361, 1998.
[134] F. F. Blauw and S. von Solms, "Towards quantifying and defining privacy metrics for online users," in *Proc. IST-Africa Week Conf.*, Piscataway, NJ, USA, May 2017, pp. 1–9.
[135] J. Becker and H. Chen. (2009). *Measuring Privacy Risk in Online Social Networks, Web 2.0 Security and Privacy*. [Online]. Available: http://web.cs.ucdavis.edu/~hchen/paper/w2sp2009.pdf
[136] A. Braunstein, L. Granka, and J. Staddon, "Indirect content privacy surveys: Measuring privacy without asking about it," in *Proc. 7th Symp. Usable Privacy Secur.*, New York, NY, USA, 2011, p. 15, doi: 10.1145/2078827.2078847.
[137] E. Shanas, "Descriptive and sampling statistics. john gray peatman," *Amer. J. Sociol.*, vol. 53, no. 5, p. 412, 1948, doi: 10.1086/220229.
[138] M.-J. Sule, M. Li, and G. Taylor, "Trust modeling in cloud computing," in *Proc. IEEE Symp. Service-Oriented Syst. Eng. (SOSE)*, Mar. 2016, pp. 60–65.
[139] P. Zhang, Y. Kong, and M. Zhou, "A domain partition-based trust model for unreliable clouds," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 9, pp. 2167–2178, Sep. 2018.
[140] P. Zhang, M. Zhou, and Y. Kong, "A double-blind anonymous evaluation-based trust model in cloud computing environments," *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Apr. 4, 2019, doi: 10.1109/TSMC.2019.2906310.
[141] K. Star. (2015). *Open Anonymizer*. Accessed: Nov. 2019. [Online]. Available: http://sourceforge.net/projects/openanonymizer/
[142] Technology Methods and Infrastructure for Networked Medical Research (TMF). (2015). *Anon Tool*. Accessed: Nov. 2019. [Online]. Available: http://www.tmf-ev.de/Home
[143] Y. Li, A. Slagell, K. Luo, and W. Yurcik, "Canine: A combined conversion and anonymization tool for processing netflows for security," in *Proc. 10th Int. Conf. Telecommun. Syst. Model. Anal.*, vol. 21. Dallas, TX, USA: ICTSMA, 2005, pp. 1–9.
[144] W. Yurcik, C. Woolam, G. Hellings, L. Khan, and B. Thuraisingham, "SCRUB-tcpdump: A multi-level packet anonymizer demonstrating privacy/analysis tradeoffs," in *Proc. 3rd Int. Conf. Secur. Privacy Commun. Netw. Workshops*, 2007, pp. 49–56.
[145] K. Luo, Y. Li, C. Ermopoulos, W. Yurcik, and A. J. Slagell, "SCRUB-PA: A multi-level multi-dimensional anonymization tool for process accounting," *CoRR*, vols. 4, p. 19, Dec. 2006. [Online]. Available: http://arxiv.org/abs/cs/0601079
[146] C. Liu, H. Duan, Q. Zeng, M. Zhou, F. Lu, and J. Cheng, "Towards comprehensive support for privacy preservation cross-organization business process mining," *IEEE Trans. Services Comput.*, vol. 12, no. 4, pp. 639–653, Jul. 2019.
[147] W. Yurcik, C. Woolam, G. Hellings, L. Khan, and B. Thuraisingham, "Measuring anonymization privacy/analysis tradeoffs inherent to sharing network data," in *Proc. IEEE Netw. Operations Manage. Symp.*, Bahia, Brazil, 2008, pp. 991–994.
[148] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon, "Prefix-preserving ip address anonymization: Measurement-based security evaluation and a new cryptography-based scheme," in *Proc. 10th IEEE Int. Conf. Netw. Protocols*, Washington, DC, USA: IEEE Computer Society, 2002, pp. 280–289. [Online]. Available: http://dl.acm.org/citation.cfm?id=645532.656186
[149] J. Kristoff. (2019) *IP: Anonymous—Perl Port of Crypto-PAn to Provide Anonymous IP Addresses*. [Online]. Available: https://metacpan.org/pod/IP::Anonymous
[150] T. Farah, "Algorithms and tools for anonymization of the Internet traffic," Ph.D. dissertation, School Eng. Sci., Fac. Appl. Sci., Simon Fraser Univ., Burnaby, BC, Canada, 2013. [Online]. Available: http://www2.ensc.sfu.ca/~ljilja/cnl/pdf/Thesis_tanjila_final.pdf
[151] D. Plonka. (2003). *IP2anonIP*. Accessed: Nov. 2019. [Online]. Available: http://pages.cs.wisc.edu/~plonka/ip2anonip/
[152] F. Prasser and F. Kohlmayer, *Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool*. Cham, Switzerland: Springer, 2015, pp. 111–148, doi: 10.1007/978-3-319-23633-9_6.
[153] M. Templ, A. Kowarik, and B. Meindl, "Statistical disclosure control for micro-data using theRPackagesdcMicro," *J. Stat. Softw.*, vol. 67, no. 4, pp. 1–36, 2015.
[154] (2019). *Risk-based De-identification Software and Services for HIPAA Compliance*. Accessed: Dec. 2019. [Online]. Available: https://privacy-analytics.com
[155] G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, *SECRETA: A Tool for Anonymizing Relational, Transaction RT-Datasets*. Cham, Switzerland: Springer, 2015, pp. 83–109, doi: 10.1007/978-3-319-23633-9_5.
[156] L. Sweeney, *Datafly: A System for Providing Anonymity in Medical Data*. Boston, MA, USA: Springer, 1998, pp. 356–381, doi: 10.1007/978-0-387-35285-5_22.
[157] A. Hundepool and L. Willenborg, "Argus, software packages for statistical disclosure control," in *Compstat*, R. Payne and P. Green, Eds. Berlin, Germany: Physica-Verlag, 1998, pp. 341–345.
[158] X. Wang, J.-K. Chou, W. Chen, H. Guan, W. Chen, T. Lao, and K.-L. Ma, "A utility-aware visual approach for anonymizing multi-attribute tabular data," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 351–360, Jan. 2018. [Online]. Available: http://ieeexplore.ieee.org/document/8019828/
[159] Comprehensive R Archive Network. (2019). *R: What is R*. [Online]. Available: https://www.r-project.org/about.html
[160] R. Matsunaga, I. Ricarte, T. Basso, and R. Moraes, "Towards an ontology-based definition of data anonymization policy for cloud computing and big data," in *Proc. 47th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. Workshops (DSN-W)*, Jun. 2017, pp. 75–82. http://ieeexplore.ieee.org/document/8023701/
[161] T. Basso, R. Moraes, N. Antunes, M. Vieira, W. Santos, and W. Meira, "PRIVAaaS: Privacy approach for a distributed cloud-based data analytics platforms," in *Proc. 17th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGRID)*, Piscataway, NJ, USA, May 2017, pp. 1108–1116, doi: 10.1109/CCGRID.2017.136.
[162] X. Zhu, E. Ayday, and R. Vitenberg, "A privacy-preserving framework for outsourcing location-based services to the cloud," Dept. Inform., Univ. Oslo, Oslo, Norway, Tech. Rep. 35645, 2019. [Online]. Available: https://www.duo.uio.no/handle/10852/60435
[163] J. Li, J. Wei, W. Liu, and X. Hu, "PMDP: A framework for preserving multiparty data privacy in cloud computing," *Secur. Commun. Netw.*, vol. 2017, pp. 1–7, Jan. 2017.
[164] E. Toch, C. Bettini, E. Shmueli, L. Radaelli, A. Lanzi, D. Riboni, and B. Lepri, "The privacy implications of cyber security systems: A technological survey," *ACM Comput. Surv.*, vol. 51, no. 2, p. 36:1–36:27, Feb. 2018, doi: 10.1145/3172869.
[165] P. Vassiliadis, "A survey of extract–transform–load technology," *Int. J. Data Warehousing Mining*, vol. 5, no. 3, pp. 1–27, Jul. 2009.
[166] A. Singhal, "Modern information retrieval: A brief overview," *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, Jan. 2001.
[167] S. S. Shapiro, "Privacy by design: Moving from art to practice," *Commun. ACM*, vol. 53, no. 6, p. 27, Jun. 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1743546.1743559
[168] S. Spiekermann, "The challenges of privacy by design," *Commun. ACM*, vol. 55, no. 7, p. 38, Jul. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2209249.2209263
[169] A. Monreale, S. Rinzivillo, F. Pratesi, F. Giannotti, and D. Pedreschi, "Privacy-by-design in big data analytics and social mining," *EPJ Data Sci.*, vol. 3, no. 1, p. 10, Dec. 2014. [Online]. Available: http://www.epjdatascience.com/content/3/1/10
[170] M. Langheinrich, "Privacy by design—Principles of privacy-aware ubiquitous systems," in *Ubicomp 2001: Ubiquitous Computing*, G. D. Abowd, B. Brumitt, and S. Shafer, Eds. Berlin, Germany: Springer, 2001, pp. 273–291.

PAULO SILVA received the M.Sc. degree in communications, services and infrastructures. He is currently pursuing the Ph.D. degree with the Doctoral Program in Information Science and Technology of the University of Coimbra. He is currently a Software Engineer. He is also with the Center for Informatics and Systems of the University of Coimbra.

He has over ten publications in journals and international conferences. He has also participated in several European initiatives and projects. His main research interests are data privacy protection and security services for cloud computing.

EDMUNDO MONTEIRO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering and the Habilitation degree in informatics engineering from the University of Coimbra (UC), Portugal, in 1996 and 2007, respectively. He is currently a Full Professor with UC.

He has participated in many European initiatives and projects. He has authored over 200 publications in books, journals, book chapters, and international conferences. He has also co-authored nine international patents. His research interests are computer networks, Internet security, cloud networking, and wireless communications. He is a member of the Editorial Board of *Wireless Networks* and ITU-FET journals.

PAULO SIMÕES (Senior Member, IEEE) is currently an Associate Professor with the Department of Informatics Engineering, UC, and a Senior Researcher with the Laboratory of Communications and Telematics.

He was also co-founder of two technological spin-off companies that currently employ more than 300 people. He has been involved in several European- and industry-funded research projects, with both technical and management activities. His main research interests are security, network management, and critical infrastructure protection. He has over 150 journal and conference publications in these areas.