

Received April 21, 2022, accepted May 6, 2022, date of publication May 16, 2022, date of current version May 23, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3175219

Toward Privacy Preservation Using Clustering Based Anonymization: Recent Advances and Future Research Outlook

ABDUL MAJEED¹, SAFIULLAH KHAN²,
AND SEONG OUN HWANG¹, (Senior Member, IEEE)

¹Department of Computer Engineering, Gachon University, Seongnam-si 13120, Republic of Korea

²Department of IT Convergence Engineering, Gachon University, Seongnam-si 13120, Republic of Korea

Corresponding authors: Seong Oun Hwang (sohwang@gachon.ac.kr) and Abdul Majeed (ab09@gachon.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) under the High-Potential Individuals Global Training Program funded by the Korean Government through Ministry of Science and ICT (MSIT) under Grant 2021-0-01532 (50%), and in part by the National Research Foundation of Korea (NRF) funded by the Korean Government through MSIT under Grant 2020R1A2B5B01002145 (50%).

ABSTRACT With the continuous increase in avenues of personal data generation, privacy protection has become a hot research topic resulting in various proposed mechanisms to address this social issue. The main technical solutions for guaranteeing a user's privacy are encryption, pseudonymization, anonymization, differential privacy (DP), and obfuscation. Despite the success of other solutions, anonymization has been widely used in commercial settings for privacy preservation because of its algorithmic simplicity and low computing overhead. It facilitates unconstrained analysis of published data that DP and the other latest techniques cannot offer, and it is a mainstream solution for responsible data science. In this paper, we present a comprehensive analysis of clustering-based anonymization mechanisms (CAMs) that have been recently proposed to preserve both privacy and utility in data publishing. We systematically categorize the existing CAMs based on heterogeneous types of data (tables, graphs, matrixes, etc.), and we present an up-to-date, extensive review of existing CAMs and the metrics used for their evaluation. We discuss the superiority and effectiveness of CAMs over traditional anonymization mechanisms. We highlight the significance of CAMs in different computing paradigms, such as social networks, the internet of things, cloud computing, AI, and location-based systems with regard to privacy preservation. Furthermore, we present various proposed representative CAMs that compromise individual privacy, rather than safeguarding it. Besides, this article provides an extended knowledge (e.g., key assertion(s), strengths, weaknesses, clustering methods used in the anonymization process, and %age improvements in quantitative results) about each technique that provides a clear view of how much this topic has been investigated thus far, and what are the research gaps that seek pertinent solutions in the near future. Finally, we discuss the technical challenges of applying CAMs, and we suggest promising opportunities for future research. To the best of our knowledge, this is the first work to systematically cover current CAMs involving different data types and computing paradigms.

INDEX TERMS Privacy, utility, anonymization, personal data, clustering, social networks, differential privacy, pseudonymization, encryption, de-anonymization, responsible data science, artificial intelligence.

I. INTRODUCTION

With the rapid advances in information and communications technologies, personal data have become an economic resource that can assist data owners (hospitals, banks,

The associate editor coordinating the review of this manuscript and approving it for publication was Sedat Akleyek¹.

insurance companies, social networking service providers, etc.) in fulfilling the needs/expectations of their affiliates in a seamless manner. With the huge proliferation of pervasive computing and digital tools, data owners are obtaining huge and varied amounts of personal data for financial gain. In recent years, collections of personal big data (i.e., private data produced in the daily lives/work of individuals) have

become valuable assets in the data market, and have replaced oil as the most economical resource [1]. The huge amount of collected personal data often encompasses information about an individual's demographics, spatial-temporal activities, photographs, finances, political/religious views, interests, hobbies, social circle information, and medical status, to name just a few types. Outsourcing the collected data to analytics firms/companies in order to extract relevant information regarding consumers can help companies sustain a competitive advantage, but privacy problems are the main hurdle in doing so [2]. Due to privacy issues, companies often prefer not to outsource their consumer/customer data to legitimate information consumers for knowledge discovery. The three common privacy issues that can occur as a result of data outsourcing based on users' attributes are disclosures of identity, sensitive information, and memberships [3].

According to a survey in the United States [4], unique identification of individuals is possible at very different percentages based on the following combinations of three attributes:

- zip code (five-digits), date of birth, gender → **87%**
- place of residence, date of birth, gender → **50%**
- country of origin, date of birth, gender → **18%**

User attributes such as date of birth, gender, zip code, race, and country of origin are called quasi-identifiers (QIDs). These QIDs in personal data can increase the chances of disclosing identities and corresponding sensitive attributes (SAs) [5]. To address these privacy issues, personal data are usually anonymized before publication. The technical solutions for protecting an individual's privacy in personal data handling are obfuscation, encryption, anonymization, and pseudonymization. However, due to low computing overhead and algorithmic simplicity, anonymization has been used extensively in commercial settings for privacy-preserving data publishing (PPDP) and was recently legislated in some advanced countries [6]. It employs many anonymization operations, such as generalization, suppression, randomization, slicing, and derived records, in order to strike a balance between privacy and utility in PPDP.

Primarily, most anonymization approaches are applied to tabular/relational data. The well-known anonymization approaches applied to tabular data are k -anonymity [7], ℓ -diversity [8], and t -closeness [9]. These models showed remarkable results in terms of privacy preservation in the early days. However, they proved unsuccessful against certain contemporary privacy threats, and many refinements have been proposed to upgrade them [10], [11]. Some other developments (a.k.a. utility enhancements) have emerged in parallel to meet the needs of data analysts by keeping most data characteristics as close as possible to the original. For example, in 2006, differential privacy (DP) [12] was proposed for dynamic scenarios (e.g., query-answer). Afterwards, researchers extended the anonymization concepts from tabular data to social networking (SN) data in order to protect

user privacy in graph publishing [13], [14]. For example, the k -anonymity concept for tabular data was modified to k -degree anonymity in order to preserve privacy in social graph $G(U, V)$, where U denotes SN users, and V is the set of edges modeling the relationship between users [15]. In recent years, anonymization approaches have been rigorously applied to diverse data formats (matrixes, tables, graphs, text, traces, multimedia, documents, etc.) for privacy preservation under multiple computing paradigms, such as the internet of things (IoT), artificial intelligence (AI) environments, and cloud computing. In this paper, we focus on clustering-based anonymization mechanisms (CAMs) that have shown remarkable improvements over traditional approaches in preserving both privacy and utility in recent years.

Previous reviews related to PPDP have covered important aspects, such as relational/graph anonymization techniques, privacy models and their extensions, anonymization operations, data anonymity frameworks, privacy-protection tools, and evaluation metrics used by the PPDP mechanisms. Rajendran *et al.* [16] discussed the strengths and weaknesses of three famous anonymity models: k -anonymity, ℓ -diversity, and t -closeness. Tran and Hu [17] provided a systematic review of big data analytics that preserves privacy. Other authors have discussed many generic privacy-preserving approaches for data querying, data publishing, and data mining. A few surveys have been published on outsourcing SN users' data while preserving privacy [18], [19]. Some authors have discussed various ways for preserving node/edge privacy when sharing G with third parties. Sharma *et al.* [20] discussed privacy concerns and corresponding privacy-preserving techniques for big data. Majeed and Lee [21] presented a detailed review of anonymization approaches that were applied to tabular and graph data. Cunha *et al.* [22] discussed various anonymization approaches for different data types, and provided a detailed taxonomy of privacy protection mechanisms and tools. Recently, a survey on privacy preservation in social media networks was published [23]. Although we fully affirm the key findings of previous reviews, the concepts/approaches covered in those reviews were limited, and CAMs were not covered thoroughly. To the best of our knowledge, none of the existing reviews covered CAMs that have been proposed for different computing paradigms and heterogeneous data formats. To close the gap, this paper presents a comprehensive review of anonymization techniques that employ clustering concepts while converting raw data into anonymized data. The major contributions of this article to the PPDP field are summarized as follows.

- We summarize the key findings of state-of-the-art (SOTA) clustering-based anonymization mechanisms that have been proposed for the effective resolution of privacy and utility trade-offs in PPDP.
- We systematically categorize existing CAMs into heterogeneous data formats, including SN (i.e., social graphs), relational (i.e., tabular), transactional (i.e., set), and trace, and present an up-to-date, thorough review

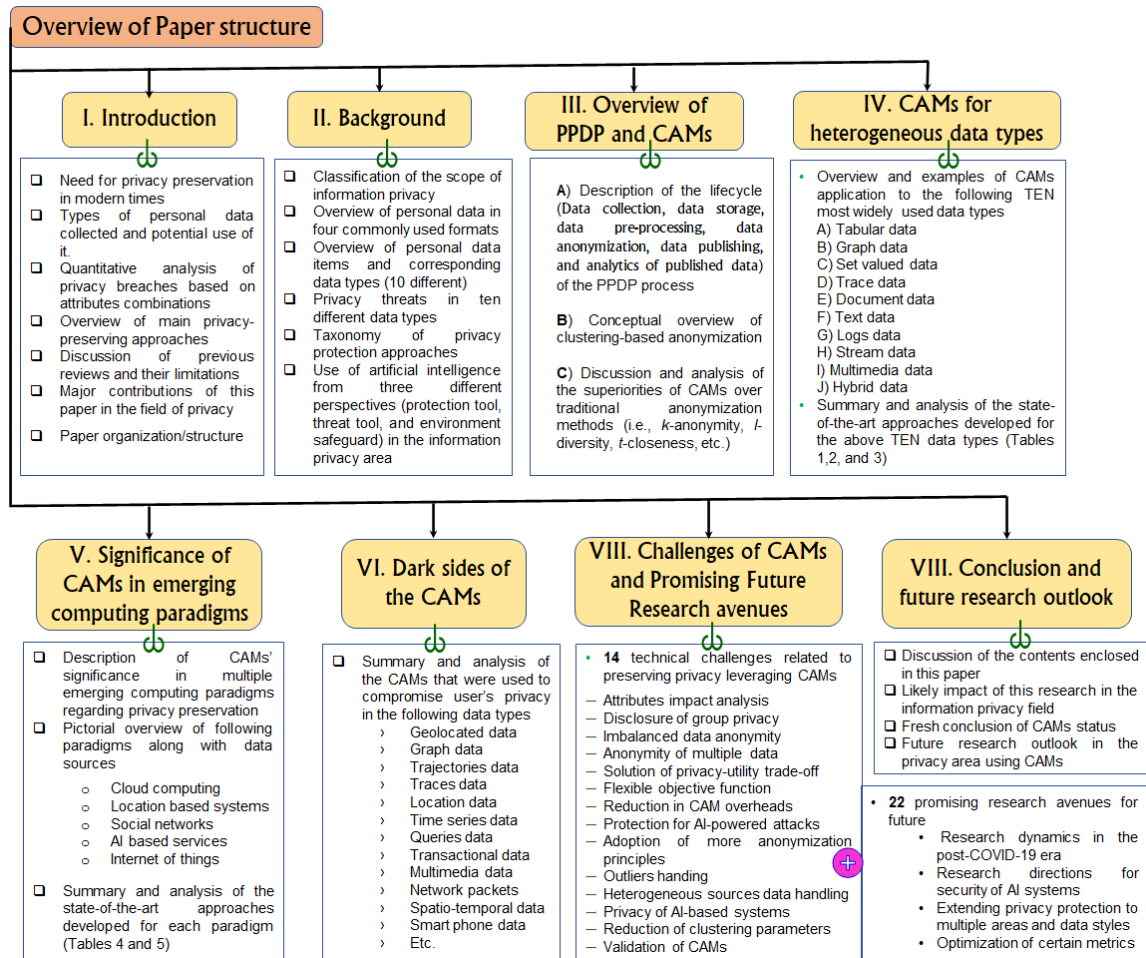


FIGURE 1. Overview of the paper's structure (i.e., main sections/headings and key contents enclosed under each section).

of the latest anonymization techniques and metrics employed for their evaluations.

- We describe the role of CAMs regarding privacy preservation in different computing paradigms, such as cloud computing, the IoT, location-based services, AI-powered services, and application-specific SN scenarios (community clustering, collaborative filtering, privacy-aware recommendations, graph mining, etc.) that remained unexplored in the recent literature.
- This paper highlights various representative CAMs that are exploited by malevolent adversaries in order to compromise user privacy from published data (i.e., unique re-identification across SN sites, SA disclosures, and inferring private data).
- We discuss many latest and real-time computing (RTC) applications that leverage personal data, and highlight privacy development in each RTC application.
- We present the technical challenges in protecting user privacy by using CAMs, and list potential avenues for future research to address contemporary privacy threats.
- This is the first work focusing on CAMs from a broader perspective that can provide a solid foundation for future developments in the PPDP area.

The remainder of this article is organized as follows. Section II presents the background of the privacy concept, on personal information enclosed in multiple data types, on well-known privacy threats, on privacy protection techniques and their operations, and on the role of machine learning (ML) techniques in the information privacy domain. Section III presents an overview of the PPDP process, conceptual overview of CAMs, and superiority of CAMs over traditional anonymization methods. Section IV provides an overview of the 10 most widely used data formats and the corresponding SOTA CAMs for each data format. Section V discusses the significance of CAMs in the emerging computing paradigms. Section VI highlights the dark side of CAMs in terms of privacy breaches. Then, we discuss the challenges of CAMs and suggest promising avenues for future research in Section VII. Finally, we conclude the paper in Section VIII. Figure 1 demonstrate the high level structure of this survey paper.

As shown in Figure 1, we categorize the structure of this survey paper based on the complexity of information in a sequential manner (i.e., the information complexity increase down the order). For example, in Section II, we present the basic knowledge about the subject matter including the scope

of privacy as a whole, ten different data types in which personal data is usually represented, and privacy threats based on the data types (i.e., edge disclosure can occur only in a graph data), the taxonomy of major privacy-enhancing technologies, and role of AI in privacy area from three perspectives. In Section III, we demonstrate an overview of the system where CAMs are used followed by their working principle. In Section IV, we present an overview of different data styles and CAMs applications on them. Later, we analyze SOTA CAMs used for each data type with a detailed analysis of each technique. In Section V, we show the CAMs application in multiple computing paradigms along with a detailed analysis. Basically, Sections IV and V show the bright sides of CAMs. In Section VI, we show the dark sides of the CAMs along with the critical analysis of each study. In Section VII, we highlight open challenges and future research opportunities in detail. Finally, we summarize the key points of this article and the conclusion of CAMs in Section VIII.

II. BACKGROUND

Privacy has countless shades/definitions and is very subjective (i.e., the perception of it varies from individual to individual) [24]. In simple words, privacy is about safeguarding private information against prying eyes (a.k.a. public access) [25]. Privacy is regarded as one of the fundamental human rights and is vital for autonomy, individualism, and self-respect. The scope of privacy can be classified into four distinct categories, as shown in Figure 2. This review focuses on the first category (information privacy), which includes systems/infrastructures that gather, store, analyze, utilize, and disseminate personal data.

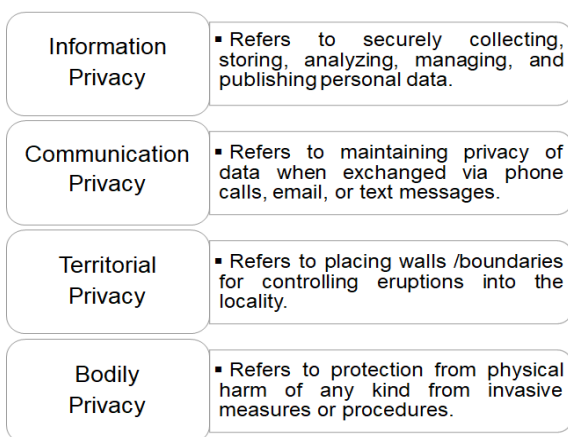


FIGURE 2. Description of the scope of privacy (adapted from [26]).

Personal data can be represented in different formats, including tables, graphs, text, sets, and images. For example, SN data are frequently modeled/represented with graphs. Moreover, hospitals/clinics mostly store and process personal data in a tabular form. Superstores usually manage consumer/customer data in set-valued form. In contrast, some sectors handle personal data in a continuous fashion called streams. Figure 3 presents a generic overview of

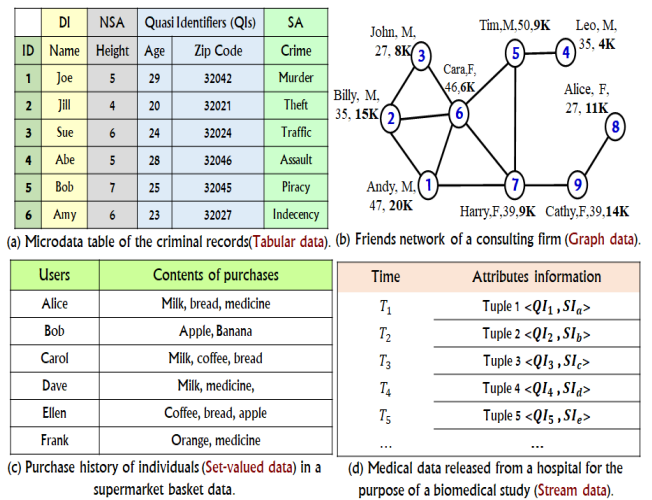


FIGURE 3. Overview of user data in four different styles (e.g., tables, graphs, sets, and streams).

the four different types/styles in which personal data are encompassed. In some cases, the same personal data can be consistently modeled in multiple formats. For instance, SN users' data can be presented in both graphs and tables. Personal information that needs privacy preservation can be of different types (diseases, photos, income, etc.) and can be encompassed in any one of the above data formats (tables, traces, set-valued, etc.). We present a detailed overview of personal information enclosed in different data types/styles in Figure 4. These can be classified as unstructured, semi-structured, and structured [22].

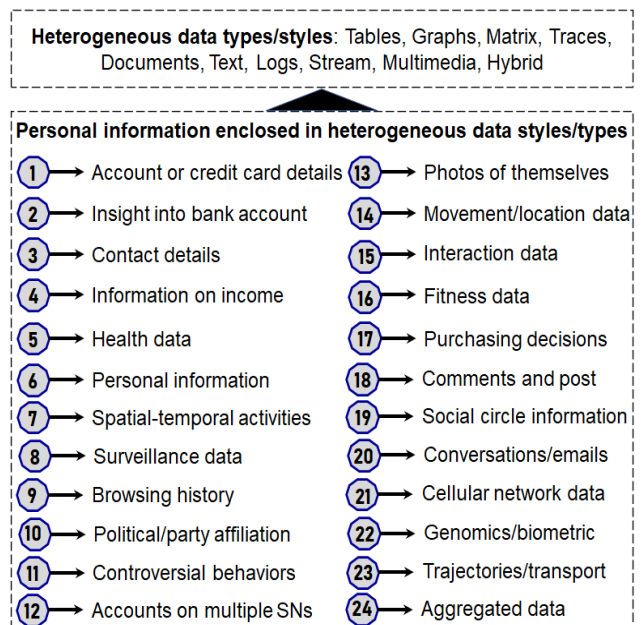


FIGURE 4. Generic overview of different types of data and personal information enclosed in them.

Privacy threats can also vary depending upon the style of the data and the corresponding personal information enclosed

in each style (a.k.a. type). We provide a brief overview of privacy threats that can be executed on different data styles as follows.

- **Table:** identity disclosure, SA disclosure, membership disclosure, privacy-intrusive pattern revelation, group privacy theft, association rules extraction, etc.
- **Graph:** node re-identification, connection/ relationship disclosure, edge/vertex label disclosure, affiliation disclosure, multiple SN account disclosure, community label disclosure, etc.
- **Matrix/Set:** sensitive itemset disclosures, purchase history theft, financial status disclosure, spatial-temporal activities disclosure, transport usage data disclosure, etc.
- **Traces/Logs:** location disclosure, trajectory disclosures, mobility pattern disclosures, spatial-temporal stay points disclosure, web-search disclosure, sensitive place visit disclosure, interaction disclosures, etc.
- **Documents:** intimate details of someone’s life, medical/prescription history disclosure, income tax exposure, personal data disclosure, genomics data disclosure, etc.
- **Text:** intent disclosures, opinion disclosures, political party affiliation disclosure, personal preferences disclosure, social circle information disclosure, content disclosure, etc.
- **Stream:** diagnosis history, illegitimate data aggregation, stalking of individuals, targeted profiling, patterns in web searches, interest disclosures, mobility disclosure, location disclosure, etc.
- **Multimedia:** facial privacy disclosures (a.k.a. identity disclosure), SA disclosure, appearance disclosure, political affiliation disclosure, sensitive/controversial place visit disclosures, sensitive information predictions, itemset disclosures, surveillance data disclosure, hidden profiling, etc.
- **Hybrid:** multiple and intrusive high-privacy disclosures mentioned in the above data styles.

To safeguard user privacy against prying eyes, multiple privacy protection approaches have been proposed for the secure collection, processing, analysis, utilization, and publication of personal data. We present a taxonomy of famous approaches in Figure 5, along with their concise descriptions and main operations. The main operations performed in each approach have benefits/liabilities in terms of computing complexity, conceptual simplicity, robustness, effectiveness in the privacy/usefulness trade-off, number of iterations, and resource utilization. For example, suppression and generalization operations have a distinct impact on privacy and utility, respectively. The former provides a higher level of privacy, but no utility for information consumers. In contrast, the latter sustains better utility and privacy in anonymized data. In addition, cryptography-based operations are mostly slow, but enable trans-border data flow, and provide rigorous privacy guarantees. These operations have been widely used in interactive scenarios (e.g., the IoT, SN, edge/cloud computing). Obfuscation-based approaches are highly useful in preserving the privacy of geo-spatial data (i.e.,

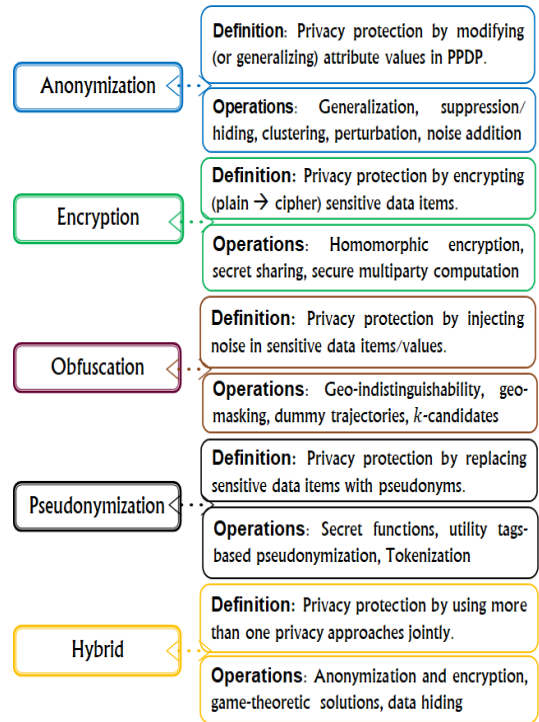


FIGURE 5. Taxonomy of privacy protection approaches used for safeguarding an individual’s privacy.

mobility and trajectories) by incorporating a fair amount of noise. The operations performed by pseudonymization-based approaches assist in hiding sensitive data by replacing them with pseudonyms. These approaches are mainly preferred in vehicular networks and smart-home environments. Finally, the hybrid approaches perform multiple operations, jointly considering the type of data, the characteristics of the attributes, and the objectives of privacy/utility in order to meet privacy/utility expectations [27]. All these approaches have been widely used in preserving both privacy and utility in different computing paradigms.

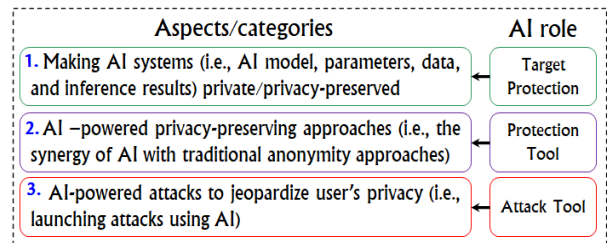


FIGURE 6. Significance of AI in information privacy domain (adapted from [28]).

In recent years, AI approaches have opened up new challenges and opportunities in the privacy protection domain. On one hand, they have enhanced the capabilities of existing privacy-preserving approaches in effectively preserving a user’s privacy. On the other hand, they have become a target of malevolent adversaries, and can still allow disclosure of sensitive information. Majeed et al. [29] applied AI concepts to improve performance from the traditional anonymity

approach to privacy and utility preservation. In contrast, Park and Lim [30] proposed the idea of securing federated learning (FL) using homomorphic encryption. In the coming years, privacy-preserving approaches will benefit from AI-based approaches, and vice versa. In line with this trend, synergy between AI and privacy-preserving approaches can be categorized from three aspects (as shown in Figure 6).

Although many SA-specific, data-specific, application/threat-specific, domain-specific, attack-specific, sector-specific, and AI-based privacy-preserving approaches have been devised, clustering-based privacy-preserving approaches have improved traditional anonymization in different contexts. Therefore, the remainder of this review solely explores clustering-based anonymization approaches/developments in the context of PPDP.

III. OVERVIEW OF PRIVACY PRESERVING DATA PUBLISHING AND CAMs

In this section, we discuss the overview of PPDP and CAMs. Specifically, we discuss the life cycle of PPDP, the basic concepts of CAMs, and the superiority of CAMs over traditional anonymization algorithms.

A. DESCRIPTION OF THE LIFE CYCLE OF PPDP

The typical PPDP process encompasses six steps, all of which, along with their execution order, are shown in Figure 7. In Step A, appropriate data are collected from relevant individuals. Examples of data collection are account-opening procedures in a bank, or a check-up from a diagnostic center. In both of these scenarios, some basic information (i.e., QIs) as well as sensitive information (i.e., SAs), is obtained. Subsequently, the collected data are stored in safe repositories/databases for further analysis (Step B). Storage can be in graph form (e.g., SN data) or tabular form (e.g., hospital/bank data) depending upon the nature of the collected data. Due to the recent advancements in technology, storage capacity has become sufficiently large, and all types of data can be stored for utilization in multiple contexts. In Step C, pre-processing is applied to the collected data. During this step, the data are cleaned (outliers and missing values are removed, formatting and type checking is performed, and redundant records are removed). In Step D, the cleaned data from Step C are anonymized. During data anonymization, the original data are modified to preserve privacy, leaving the anonymized dataset useful for analysis. In Step E, anonymized data are published for analysis and data mining. In the final step, analytics is applied to the published data to extract useful information for hypothesis generation/verification.

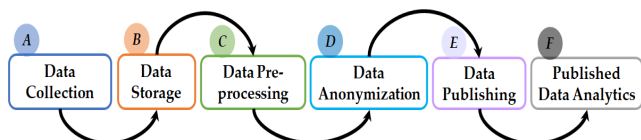


FIGURE 7. Overview of the life cycle of the PPDP process.

A conceptual overview of the anonymization process applied on raw data for PPDP is demonstrated in Figure 8.

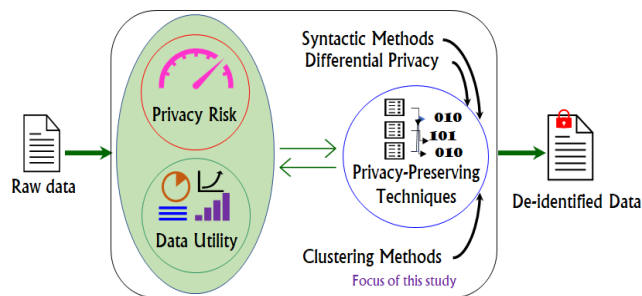


FIGURE 8. High level overview of the anonymization process (i.e., converting raw data → de-identified data) used for PPDP, and main focus of this study.

B. DESCRIPTION OF THE CLUSTERING BASED ANONYMIZATION APPROACHES USED FOR PPDP

Thus far, many anonymization approaches have been proposed to address privacy and utility issues in PPDP-leveraging clustering concepts. We illustrate a generic overview of the clustering concept in Figure 9. The anonymization of clusters is mainly the same as anonymizing QI groups (a.k.a. equivalence classes) in the traditional anonymization approaches (k -anonymity, ℓ -diversity, t -closeness, and their extensions). The CAMs have been extensively studied in the recent literature for privacy preservation due to improved privacy and utility results. Furthermore, the anonymized data produced by the CAMs are helpful for secondary purposes (e.g., demography-based disease analysis, policy-making, future event predictions).

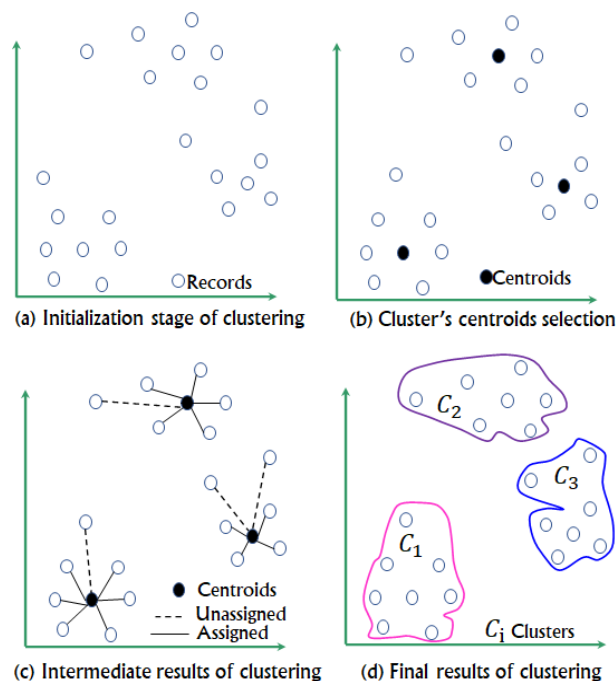


FIGURE 9. Generic overview of the clustering concept.

C. THE SUPERIORITY OF CLUSTERING-BASED APPROACHES OVER TRADITIONAL ANONYMIZATION APPROACHES

The clustering-based approaches have revolutionized the information privacy domain in many aspects. For instance, the k -anonymity model enforces a constraint on the number of people in a QI-group/class, and usually retains k people in the QI group. In contrast, CAMs can remove such hard constraints and can keep the same people in the cluster, regardless of their strengths in using the similarity/distance concept. The mathematical expression used to compute similarity (S) between two users (or between a user and the cluster center) is given in Eq. 1:

$$S(u_1, u_2) = \frac{\sum_{i=1}^p u_{1i} \times u_{2i}}{\sqrt{(\sum_{i=1}^p u_{1i})^2} \times \sqrt{(\sum_{i=1}^p u_{2i})^2}} \quad (1)$$

where i denotes the QIs, and p represents the total number of QIs. From Eq. 1, S values between users and cluster centers can be computed, and clusters can be formed.

In CAMs, multiple checks are performed for each record in order to find the best-matching cluster that ensures homogeneity in clusters. In contrast, the traditional anonymization approaches usually assign records to a QI group by performing a single check (leading to imprecise utility results in most cases). Since traditional anonymization approaches often ignore similarity/distance concepts while making the groups/classes, the generalization intervals are very wide, which can lead to false hypothesis generation in the end. In contrast, CAMs employ distance/similarity concepts, and therefore, the possibility of false hypothesis generation is relatively low. Furthermore, CAMs can control the issue of over-generalization, and an anonymized dataset produced by them has better utility and privacy. Analytical and data mining tasks can be performed with sufficient accuracy. In contrast, traditional anonymization often leads to imprecise analysis results by introducing heavier changes in the anonymized data. Furthermore, CAMs have the ability to control the heavier changes during data anonymization, which can pave the way for better resolution of privacy versus utility.

The anonymized data produced by CAMs enable better understanding of differences among commonalities, and of commonalities among the differences. Furthermore, CAMs are vital for enhancing the performance of knowledge-based systems/applications. CAMs have the ability to control inaccurate decision-making, and they enable a better understanding of patterns/trends from the anonymized data. In addition, CAMs are flexible, meaning they can be applied to different data styles with minor modifications. CAMs can yield consistent performance with different data styles and domains. CAMs have the ability to produce promising results in big data platforms such as MapReduce, Spark, and Hadoop. Recently, CAMs have also been extensively applied to unsuitable/imbalanced data in order to meet analytics

demands [31]. In the coming years, application areas for CAMs are likely to expand to many domains.

IV. CLUSTERING-BASED ANONYMITY MECHANISMS FOR HETEROGENEOUS DATA TYPES/STYLES

In this section, we describe the effectiveness of CAMs on heterogeneous data, and we present SOTA approaches for each data type. We chose 10 representative data styles for the analysis: tables, graphs, matrixes, traces, documents, text, streams, logs, multimedia, and hybrids. We discuss basic concepts with an example of each style before discussing SOTA CAMs in Table 1.

A. CAMs FOR TABULAR DATA

Most data owners, such as banks, hospitals, and insurance companies, maintain their patient/customer/subscriber data in tabular form. Data storage, analysis, utilization, and distribution are relatively easier in tabular form, compared to other styles. A table, T , is a combination of rows and columns. Each row of T provides complete information about an individual, whereas a column is for one item (e.g., age) concerning the individuals. A generic overview of a common structure of T for a sample of 9000 individuals is shown in Eq. 2:

$$T_{users,attributes} = \begin{pmatrix} u_i & QI_1 & QI_2 \cdots & QI_p & S \\ u_1 & v_{QI_1} & v_{QI_2} \cdots & v_{QI_3} & v_1 \\ u_2 & v_{QI_1} & v_{QI_2} \cdots & v_{QI_3} & v_2 \\ u_3 & v_{QI_1} & v_{QI_2} \cdots & v_{QI_3} & v_1 \\ \dots & \dots & \dots & \dots & \dots \\ u_N & v_{QI_1} & v_{QI_2} \cdots & v_{QI_3} & v_n \end{pmatrix} = \begin{pmatrix} u_i & QI_1 = age & QI_2 = sex & QI_p = race & S = disease \\ 1 & 29 & M \cdots & Black & Flu \\ 2 & 38 & F \cdots & White & Rhinitis \\ 3 & 59 & F \cdots & White & Cancer \\ \dots & \dots & \dots & \dots & \dots \\ 9000 & 37 & M \cdots & Black & Leukemia \end{pmatrix} \quad (2)$$

where each row represents complete information about a user, including basic attributes (i.e., QIDs) as well as SAs. Moreover, each column represents one item (e.g., age or salary) related to all users.

Many approaches have been proposed to anonymize T . Well-known anonymization models (e.g., k -anonymity, ℓ -diversity, t -closeness, and their extensions) were primarily applied to tabular data only. Later, they were extended to other styles of data. CAMs have improved various drawbacks in these models with regard to computing efficiency, privacy, and utility preservation. Figure 10 presents an overview of tabular data anonymization using the clustering concept. The original T to be anonymized with the clustering technique is shown in Figure 10 (a). In Figure 10 (b), the clustering technique has been applied to T , and corresponding clustered results are shown. As seen in Figure 10 (b), record placement

(a) Original data table to be anonymized

Quasi Identifiers (QIs)				SA Info
Education	Race	Sex	Age	Salary
Bachelors	White	M	39	> 50K
Bachelors	White	M	50	≤ 50K
HS-grad	White	M	38	≤ 50K
11 th	Black	M	53	> 50K
Bachelors	Black	F	28	≤ 50K
Masters	White	F	37	> 50K
9 th	Black	F	49	≤ 50K
HS-grad	White	F	52	> 50K
Masters	White	F	31	≤ 50K
Bachelors	White	M	42	> 50K

(b) Original data after clustering process

Quasi Identifiers (QIs)				SA Info
Education	Race	Sex	Age	Salary
Bachelors	White	M	39	> 50K
Bachelors	White	M	42	> 50K
Bachelors	White	M	50	≤ 50K
HS-grad	White	M	52	> 50K
HS-grad	White	M	38	≤ 50K
Masters	White	F	37	> 50K
Masters	White	F	31	≤ 50K
Bachelors	Black	F	28	≤ 50K
11 th	Black	M	53	> 50K
10 th	Black	F	49	≤ 50K

(c) Original data after being anonymized

Quasi Identifiers (QIs)				SA Info
Education	Race	Sex	Age	Salary
Bachelors	White	M	39-42	> 50K
Bachelors	White	M	39-42	> 50K
*	White	M	50-52	≤ 50K
*	White	M	50-52	> 50K
*	White	*	37-38	≤ 50K
*	White	*	37-38	> 50K
High	*	F	28-31	≤ 50K
High	*	F	28-31	≤ 50K
Low	Black	*	49-53	> 50K
Low	Black	*	49-53	≤ 50K

FIGURE 10. Overview of tabular data anonymization using clustering concepts (adapted from Ref. [32]).

has been changed, and users have been grouped into different clusters. In the last step, anonymized data T' was generated. T' can be outsourced to information consumers for analytical/data-mining purposes.

B. CAMs FOR GRAPH DATA

Social network data are usually modeled/represented with the help of a social graph. Social graph G can contain n users, and each user can have m edges/connections with other users. Multiple ways exist to represent SN user data. In Figure 11, we illustrate the four most widely used representations of SN data via G . Anonymization approaches generally modify the structure of G to preserve both privacy and utility. The anonymization approaches devised for one type of representation cannot be directly applied to another type of G . The five main approaches used for privacy-preserving SN-data publishing are G modification, G generalization/clustering, DP-based approaches to G anonymization, privacy-aware G computation, and hybrid G anonymity methods [33].

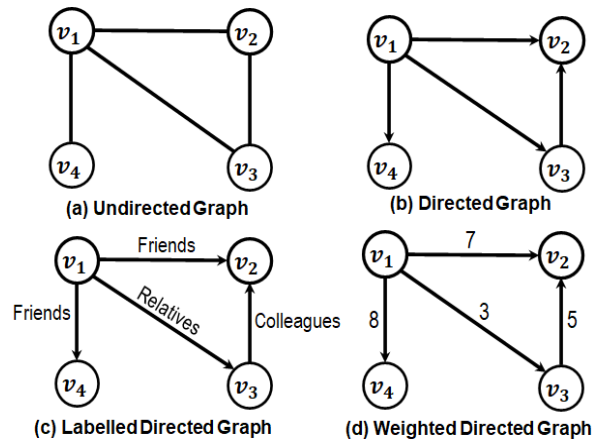


FIGURE 11. Overview of SN data representation using different versions of G (adapted from [21]).

In this work, our focus is on clustering-based anonymization, and therefore, we discuss concepts and examples related to clustering-based approaches.

CAMs usually partition G into various non-overlapping clusters, and then generalize the clusters to either super nodes or edges. An overview of clustering-based anonymization of G is shown in Figure 12.

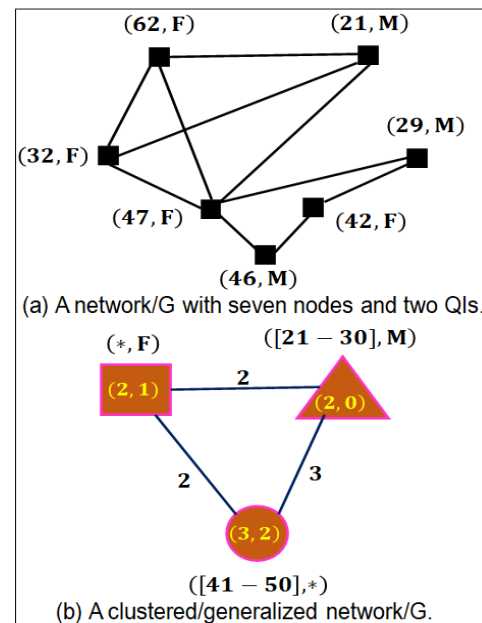


FIGURE 12. Overview of G anonymization using a CAM.

In Figure 12, G encompasses seven vertices and two QIs used as input for CAMs. The CAM partitions this G into three non-overlapping clusters by exploiting similarities between QIs. Finally, a generalized/anonymized G is obtained with three super nodes. For the sake of simplicity, we denote only three clusters with distinct shapes. The ordered pair of numbers in each super node represents the numbers of users and intra-cluster edges. Zero in a super node indicates no edge/connection between the users. Recently, there has

been an increasing focus on developing CAMs for data encompassed in G .

C. CAMs FOR SET-VALUED/MATRIX DATA

Superstores usually store and process user data in set-valued/matrix form. The complete set of data in this regard is known as a transactional database. These transactional datasets contain multiple records, called transactions, which encompass a set of items (e.g., products purchased or diagnosis codes). Such datasets have higher applicability in biomedical studies, e-commerce, and recommender systems. Many approaches have been developed for anonymizing set-valued data [34]. We demonstrate an overview of set-valued data anonymization with the clustering concept in Figure 13. In Figure 13 (a), original data to be anonymized are shown, whereas Figure 13 (b) shows the anonymized data. Due to the significant advancements in recommender systems, transactional database anonymization has become a hot research topic.

<i>tid</i>	<i>items</i>	<i>tid</i>	<i>items</i>
T_1	$i_1 i_2 i_7$	T_1	$(i_1 i_2) (i_7)$
T_2	$i_2 i_7$	T_2	$(i_1 i_2) (i_7)$
T_3	$i_3 i_5$	T_3	$(i_3 i_4) (i_5)$
T_4	$i_4 i_6 i_7$	T_4	$(i_3 i_4) (i_6) (i_7)$
T_5	$i_5 i_7$	T_5	$(i_5) (i_7)$

(a) Original data (b) Anonymized data

FIGURE 13. Overview of set-valued data anonymization using a CAM (adapted from [35]).

D. CAMs FOR TRACE DATA

Web searches, network usage, and mobility data are mostly collected, stored, and processed in trace form. Trace data hold spatial-temporal and detailed information about individuals. The anonymization of trace data has become a hot research topic, especially in the era of COVID-19. The contact tracing apps used in this pandemic mainly store individuals' movements for contact tracing purposes. Furthermore, trace data have been widely used in analytics and recommender systems. However, anonymization of trace data is very challenging due to the existence of multiple fields and high dimensionality. In the trace data, there exist multiple fields such as time of day, protocol, IP address, and many other fields. We demonstrate an overview of clustering-based anonymization of trace data with a single field (e.g., the IP address) in Figure 14. C_1 and C_2 refer to cluster 1 and cluster 2, respectively.

E. CAMs FOR DOCUMENT DATA

Sensitive data, such as medical histories, newspapers, conversations, reports, agreements, etc., are mostly enclosed in document form. In recent years, anonymization of document data has become a very hot research topic [37], [38]. Various techniques, from natural language processing (named

Original IP	Clustering	Anonymized IP
<i>SRC_IP</i>	<i>SRC_IP</i>	<i>SRC_IP</i>
10.50.50.12	10.50.50.12	10.50.50.17
10.200.21.122	10.50.50.20	10.50.50.17
10.200.21.174	10.50.60.20	10.50.60.17
10.50.60.20	10.200.21.133	10.200.21.143
10.200.21.133	10.200.21.174	10.200.21.143
10.50.50.20	10.200.21.122	10.200.21.143

C_1 (rows 1-3), C_2 (rows 4-6)

FIGURE 14. Overview of trace data anonymization using a CAM (adapted from [36]).

entity recognition) combined with clustering concepts (e.g., k -means) are employed to anonymize textual data of documents. We present in Figure 15 anonymization of text data in a document format.

Original document	Anonymized document
Dear Dr. Blue, Your patient, Mr. John Brown, stayed in our service from 05/05/1999 to 05/08/1999. Mr. Brown, 72 year old, has been admitted to emergency on 05/05/1999. His tests for the cytomegalovirus and the EBV were negative. Therefore, Dr. George Green performed an abdominal CT scan. Mr. Brown will be followed in ambulatory by Dr. Green...	Dear Dr. <PER1>, Your patient, Mr. <PER2>, stayed in our service from to <DATE 1> to <DATE2= DATE1+3>. Mr. <PER2>, 72 year old, has been admitted to emergency on <DATE1>. His tests for the cytomegalovirus and the EBV were negative. Therefore, Dr.<PER3> performed an abdominal CT scan. Mr. <PER2> will be followed in ambulatory by Dr. <PER3> ...

FIGURE 15. Overview of documents data anonymization using a CAM (adapted from [39]).

F. CAMs FOR TEXT DATA

With the rapid adoption of SN across the globe, text data including posts/comments encompass a variety of personal data that need privacy preservation from malevolent adversaries. Due to the inclusion of personal data in texts and blogs, privacy preservation has become challenging for SN service providers. Similarly, privacy preservation in clinical text by detecting and anonymizing sensitive data items has also become a vibrant area of research in recent years [40]. We present an overview of text data anonymization in Figure 16. CAMs usually anonymize cluster names and other sensitive data items from multiple texts.



Original text	Anonymized text
 Name and surname: Dinka Kovac Address: Stupine B17, Tuzla The reason for visit: influenza The name of healthcare institute: UKC, Tuzla	 Name and surname: <Empty> Address: <Empty> The reason for visit: influenza The name of healthcare institute: UKC, Tuzla

FIGURE 16. Overview of text data anonymization using a CAM (adapted from [41]).

G. CAMs FOR LOGS DATA

Web searches, website usage, communication frequency, and SN usage data are mostly collected, stored, and processed in

log form. Web search logs are useful in many respects, but present the possibility of misuse. Since logs are distinct, compared to other data styles, many aspects (such as diversity) can easily lead to privacy breaches and hidden data collection. This necessitates the need to develop anonymization methods and solutions specific to this data style/environment. CAMs are highly applicable to log data for privacy preservation [42]. We demonstrate an overview of log data anonymization in Figure 17. In clustering-based anonymization, similar attributes are combined to effectively address the privacy-utility trade-off, and storage capacity [43].

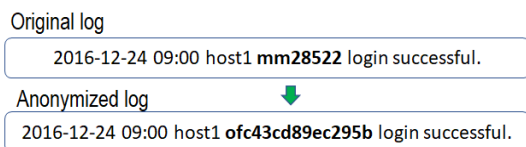


FIGURE 17. Overview of logs data anonymization using a CAM.

H. CAMs FOR STREAM DATA

With the emergence of the cloud and the edge computing paradigm, many real-time services have been developed for healthcare and intelligent prediction. In these environments, data are usually collected in real time. The main term used to denote this kind of data is stream. Stream data have many potential benefits in time-sensitive and IoT-based applications. Privacy preservation in stream data has been extensively studied in recent years [44]–[46]. Stream data are usually collected in tuples; hence, privacy preservation is more challenging, compared to other types of data [47]. We present an overview of stream data anonymization using a CAM in Figure 18. With stream data, anonymization approaches generally employ the widening concept during conversion of raw data into anonymized data.

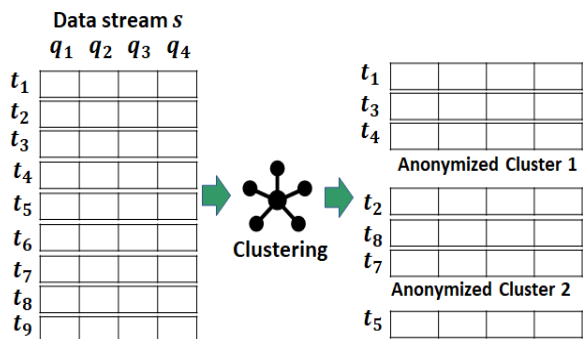


FIGURE 18. Overview of stream data anonymization using a CAM (adapted from [48]).

I. CAMs FOR MULTIMEDIA DATA

With the rapid developments of SN services, multimedia data (i.e., images and video) have become a rich source of interaction among people. People increasingly

use multimedia for a variety of purposes, such as sharing events, historical places visited, workplace activities, photographs, etc. Due to a significant rise in multimedia data generation and consumption, privacy protection has become necessary on multiple platforms. We present an overview of image data anonymization using CAM in Figure 19. Apart from image data, video also needs privacy protection from prying eyes [49]. Hence, privacy protection of multimedia data has been extensively studied in recent times.



FIGURE 19. Overview of image data anonymization using a CAM (adapted from [50]).

J. CAMs FOR HYBRID DATA

In hybrid data, more than one style is used to represent personal data. For example, SN data can be modeled with the help of tables and graphs. The anonymization of hybrid data can be performed just like it is with an individual data style. However, the number of operations can be higher with the hybrid data style, compared to the single data style. Mohapatra and Patra [112] discussed clustering-based anonymization of hybrid data (e.g., tables and graphs). The proposed approach has the ability to represent data in both table and graph. Recently, some CAMs have focused on securing personal data with higher dimensional hybrid data [113]–[115]. With the rapid developments of many digital infrastructures, hybrid data have been increasingly used in knowledge-based systems, and therefore, privacy preservation is more urgent.

In recent years, many CAMs have been developed for each data style explained above, achieving multiple objectives (privacy-preserving data analytics, data mining, analytical tasks, securing IoT-based infrastructure, and securing personal data from AI-based systems/manipulations). We provide a systematic analysis of SOTA CAMs used for different data styles in Table 1.

The SOTA CAMs listed in Table 1 have resolved many privacy-related issues in different contexts while guaranteeing utility from the anonymized data. Furthermore, these approaches were extensively used for privacy preservation in different computing paradigms, such as edge computing, the IoT, cloud computing, and SN. In Table 2, we present an in-depth analysis in terms of strength and weaknesses of

TABLE 1. Summaries and comparisons of recently proposed SOTA CAMs for heterogenous data types.

Ref.	Data type	Key Assertion (e.g., problem solved) in the context of privacy preserving data publishing	Clustering concept used
Zheng et al. [51]	Table	Improved k -anonymity using clustering by choosing cluster position reasonably to improve utility.	Community clustering
Kabiru et al. [52]	Table	Improved privacy utility trade-off using dimensionality reduction techniques in clustering	Self organising maps
Khan et al. [53]	Table	better privacy preservation over sanitization-based methods using clustering concepts	Hierarchical clustering
Priitani et al. [54]	Table	Ensured better privacy in healthcare sector using clustering concepts and by improving t -closeness	single identity clustering
Zouinina et al. [55]	Table	Bette privacy preservation in PPDP via constrained clustering and topological collaborative clustering	Self organising maps
Zheng et al. [56]	Table	Significant reduction in information loss and privacy issues by using clustering -based anonymity	K means algorithm
Ashkouti et al. [57]	Table	Bette privacy preservation in PPDP via constrained clustering and topological collaborative clustering	K means algorithm
Abbasi et al. [58]	Table	Addressed scalability and utility issues in PPDP involving high dimensional data	K -means++ method
Onesimu et al. [59]	Table	Ensured privacy protection against five active privacy attacks by using k -anonymity and clustering	bottom-up clustering
Yan et al. [60]	Table	Lowered the influence of outliers on the clustering process and improved data availability in PPDP	Weighted K -member algorithm
Rupali et al. [61]	Graph	Ensured high-level of privacy compared to SOTA methods using clustering concept on SN data	K means algorithm
Zhang et al. [62]	Graph	Minimize the information loss in graph data publishing using k -degree and clustering concepts	Partitional clustering
Shakeel et al. [63]	Graph	Provided strong privacy guarantees against active attacks while sustaining the usefulness of data	Node clustering
Chen et al. [64]	Graph	Maintained the balance between utility of mobile SN data and performance of anonymization	Density-based clustering
Debasis et al. [65]	Graph	Provided better protection against identity disclosure and maintained higher utility	Hierarchical clustering
Skarkala et al. [66]	Graph	Ensured strong privacy preservation in weighted graphs using k -anonymity and clustering	node and edge clustering
Langari et al. [67]	Graph	Provided better privacy in four SN data such as Google+, Facebook, YouTube, and Twitter	K -member fuzzy clustering
Reza et al. [68]	Graph	Provided superior results in information loss, privacy, and time complexity in G anonymization	Edge clustering
Kumar et al. [69]	Graph	Resolved the privacy and utility trade-off in anonymizing SN data enclosed in a G form	Fuzzy clustering
Heidari et al. [70]	Graph	Analyzed the distribution of nodes in clustering process to improve running time and utility of G	Sequential clustering
Wang et al. [71]	Set-valued	Provided effective resolution of privacy and information loss using community clustering concepts	Community clustering
Shyue et al. [72]	Set-valued	Provided experimental analysis of different clustering techniques on set-valued data	Gray Sort Clustering
Divanis et al. [73]	Set-valued	Devised a method that can fulfil diverse privacy and utility requirements in PPDP	agglomerative clustering
Awad et al. [74]	Set-valued	Provided higher usefulness of anonymized transactional databases in term of utility rules	ant-based clustering
Awad et al. [75]	Set-valued	Devised a method for higher knowledge extraction from anonymized data for future analysis	ant-based clustering
Barakat et al. [76]	Set-valued	Provided a mechanism for detecting quantitative privacy breach on real-world disassociated datasets	Partitional clustering
WANG et al. [77]	Set-valued	Provided a first solution towards personalized privacy preservation in transactional data anonymization	k -member clustering
Can et al. [78]	Set-valued	Proposed a mechanism for lowering information loss while anonymizing data	agglomerative and hierarchical
Meisam et al. [79]	Trace	Proposed a viable method for preventing sensitive information leakage from trace	Partitional clustering
Fan et al. [80]	Trace	Ensured strong privacy in network traffic synthesis using DP concepts and clustering	similarity-aware clustering
MEISAM et al. [81]	Trace	Retained higher utility in trace anonymization without compromising privacy via grouping approach	Multi-view grouping
Aleroud et al. [82]	Trace	Provided a robust solution for preserving privacy and utility in publicly sharing trace data	k -means clustering
Ahmed et al. [83]	Trace	Addressed privacy-utility trade off, and it excel prior techniques in attack prediction accuracy	k -means clustering
Velarde et al. [83]	Trace	Provided a new anonymization method for real life traffic traces while preserving utility and privacy	k -means clustering
Shaham et al. [84]	Trace	Provided an ML based approach for higher utility provision in spatial-temporal databases	k means algorithm
Mahajan et al. [85]	Documents	Developed a framework for carrying out search over published information in the cloud environments	EM and k means
Li et al. [86]	Documents	Implemented a prototype system for privacy preservation in sharing medical documents with 3rd party	recursive partitional clustering
Li et al. [87]	Documents	Developed a system to ensure privacy in documents such as pathology reports and clinical narratives	DP-based grouping
Kong et al. [88]	Documents	Devised a system for preserving privacy of contents and their extracted features in documents	Supervised clustering
Garat et al. [89]	Documents	Developed a prototype for privacy preservation of sensitive data in legal documents	agglomerative clustering
Li et al. [90]	Text	Developed a prototype system for privacy preservation of medical data enclosed in text form	k -means clustering
Lima et al. [91]	Text	Developed a web-based framework named, <i>HITZALMED</i> for privacy preservation in clinical text	unsupervised clustering
Liu et al. [92]	Text	Developed a practical approach for frequent pattern mining with higher utility in the PPDP	similarity-based clustering
Liu et al. [93]	Text	Preserved privacy without compromising data utility in anonymizing incomplete medical data	k -member algorithm
Ghemri et al. [94]	Text	Ensured better utility and privacy while converting raw text data into anonymized data for data sharing	k means clustering
Chen et al. [95]	Logs	Developed a system for privacy preservation in logs data using clustering-based generalization concept	Hierarchical clustering
Garcia et al. [96]	Logs	Developed privacy-preserving method for logs data that retain univariate statistics for data mining	Semantic similarity
Yuvaraj et al. [97]	Logs	Developed a model for accuracy enhancement of anonymized data without compromising data quality	Deep adaptive clustering
Meng et al. [98]	Logs	Developed a system for improving personalized search by preserving utility and privacy in logs data	Semantic clustering
Pamies et al. [99]	Logs	Developed a method for query logs protection from adversaries while improving the service searches	Semantic clustering
Ullah et al. [100]	Logs	Developed a framework for privacy preservation of web searches via clustering based anonymization	k means clustering
Nasab et al. [46]	Stream	Developed a framework for privacy preservation that can handle both numerical and categorical data	Adaptive clustering
Kumar et al. [47]	Stream	Developed a method for privacy preservation in IoT scenario using stream clustering concepts	Stream clustering
Veron et al. [101]	Stream	Ensured privacy of location data in cloud/edge computing settings using stream clustering concept	Stream clustering
Yang et al. [102]	Stream	Developed anonymization technique for distributed data in sensor network for privacy preservation	Similarity-aware clustering
Patil et al. [44]	Streams	Developed a privacy preserved system for crime related news identification by mining live data	Supervised clustering
Tekli et al. [103]	Stream	Addressed the correlation problem using clustering concept in transactional streams with better privacy	(k, l)-clustering
Honda et al. [104]	Images	Developed a practical anonymity approach for crowd movement analysis with privacy guarantees	Fuzzy clustering
Yang et al. [105]	Images	Developed a practical privacy preservation framework for facial recognition in online services	Eigenface algorithm
Zhang et al. [106]	Video	Developed a framework for anonymizing any class of objects of interest using clustering approach	Semantic segmentation
Le et al. [107]	Images	Designed a methodology and full system to improve and adjust the privacy-utility trade-off in images	StyleGAN and clustering
Grossel et al. [108]	Video	Developed an anonymization pipeline for effectively preserving privacy in video data	Semantic segmentation
Ren et al. [109]	Images	Developed a complete anonymizer for privacy preservation of people's action in human image data	Semantic segmentation
Deivanai et al. [110]	Hybrid	Developed a practical method for privacy preservation in a multiparty environment with hybrid data	Records clustering
Bazai et al. [111]	Hybrid	Developed a multi-dimensional anonymization scheme for effective resolution of privacy and utility	Spark clustering

each SOTA approach listed in Table 1. This comprehensive analysis of the approaches stated in Table 2 can pave the way for understanding existing developments as well as improving them from multiple perspectives. In Table 2, we categorized the nature of each existing study into one of the three types, theoretical, practical, and/or conceptual. Theoretical studies have not been deployed to some real-world scenarios, and limited experiments were conducted to prove their persuasiveness against major privacy threats. In contrast, practical approaches have been deployed to some real-world case, and their evaluation was performed rigorously using real-world benchmark datasets. Furthermore, in most practical approaches, attention was paid to both metrics (i.e., privacy and utility). Lastly, conceptual studies have presented only proof-of-concepts or ablation analysis, and their efficacy through detailed experiments is yet to be investigated.

The evaluation metrics employed by CAMs can vary based on data style, attack scenario, and target application. We present in Table 3 a generic overview of the metrics employed by SOTA CAMs. Apart from the famous privacy and utility metrics listed in Table 3, some SOTA CAMs have improved the performance-time, scalability, resource consumption, and other data-related issues in the PPDP process (noise, imbalance, dimensionality, outliers, etc.) [125]–[129]. Furthermore, some studies also used application-specific metrics to quantify the level of privacy and utility [130]–[132]. In many real-world cases, the privacy measured by one evaluation metric may not be monotonic. Hence, some approaches have suggested employing a metrics suite (i.e., multiple metrics), rather than relying on one or two metrics, while evaluating performance of anonymization methods [116], [133]. With the rapid increase in the diversity

TABLE 2. Detailed analysis (i.e., strengths and weaknesses) of the SOTA studies presented in Table 1.

Ref.	Study nature	Strengths	Weaknesses
Zheng et al. [51]	Practical	Exploits the distribution of QIs to enhance the data quality using an improved clustering concept	SA disclosure is possible because the proposed method does not consider the diversity of SA's values
Kabiru et al. [52]	Practical	Ensures higher data utility for data mining by increasing problem size (i.e., more attributes)	The privacy breaches can be higher due to the linkage attacks with auxiliary data (i.e., voter list)
Khan et al. [53]	Practical	Ensures minimal changes in data anonymization and better data re-construction	Prone to background knowledge and other practical attacks (i.e., skewness, table linkage, etc.)
Pritam et al. [54]	Conceptual	Better privacy preservation using enhanced t -closeness concept	Poor utility due to suppression operation and more diversity in SA's values
Zouinina et al. [55]	Theoretical	Constrained cluster-based k -anonymization with significantly reduced hand engineering	Prone to the disclosure of personal information and data re-construction attack
Zheng et al. [56]	Practical	Effective re-resolution of privacy-utility trade-off in data publishing scenarios	Fails to provide privacy and utility when data is highly imbalanced (distribution is uneven)
Ashkoufi et al. [57]	Practical	Higher utility of data by using improved k -diversity in big data environments	Prone to skewness attack and less applicability to highly imbalanced datasets
Abhassi et al. [58]	Practical	A low cost anonymization method with $1.5 \times$ reduction in IL and $3.5 \times$ reduction in time	Deletes less frequent data items that may hinder knowledge discovery process
Onsimu et al. [59]	Practical	Guarantees user's privacy at the data collection time in healthcare sectors using clustering	Prone to data reconstruction as well as table/record linkage with the data available at auxiliary sources
Yan et al. [60]	Practical	Efficiently anonymize data in the presence of outliers and lowers the IL as well as clustering effort	Prone to attribute disclosure by not considering the diversity of SA's values
Rupali et al. [61]	Practical	Privacy protection of three different elements (i.e., node, edge, and attributes) of social networks	Degradation of information availability by enforcing strict privacy parameters (i.e., k, l, t)
Zhang et al. [62]	Practical	Strong privacy protection in neighborhood attacks using k -anonymity approach on graphs	Fails to provide robust privacy in subgraph attacks as well as graph matching
Shakeel et al. [63]	Practical	Strong privacy protection in mutual friend attack scenarios	Prone to sensitive information disclosure in attributed social networks
Chen et al. [64]	Practical	Strong privacy protection against contemporary privacy threats in graph data	Feasibility test were conducted on relatively small-sized graphs (privacy analysis is not stated)
Dehaisi et al. [65]	Practical	Strong protection against identity disclosure using k -degree anonymity concept	Poor utility due to the addition of edges from outside in order to fulfill k -degree requirements
Skarkaka et al. [66]	Conceptual	Strong privacy guarantees against identity, attributes, and edge weight	The feasibility evaluation was carried out on relatively small graphs and preliminary analysis is given
Langari et al. [67]	Practical	Robust anonymization of graph data using hybrid clustering and satisfying all syntactic methods	Prone to nodes' attributes and links privacy breaches by not considering background knowledge
Reza et al. [68]	Practical	Efficient anonymization of graph data by fulfilling (k, l) -anonymity properties	Limited applicability to other types (i.e., directed, weighted, attributed, etc.) of the graphs
Kumar et al. [69]	Practical	Anonymizes graphs with better utility for social network analysis and graph mining tasks	The proposed method has a very high computational complexity, and privacy issues via graph linkage
Heidari et al. [70]	Practical	Strong privacy guarantees in graph data anonymization using k -edge-connected subgraph clustering	Prone to identity and attribute disclosure in attributed social networks
Wang et al. [71]	Conceptual	Better privacy guarantees in publishing personal data using g uncertainty model	Excessive disclosure of the sensitive transaction by not ensuring sufficient diversity in the SA's values
Shyue et al. [72]	Practical	Strong privacy protection in transactional data using sensitive k -anonymity with tuple delete/Add	Subject to important data items deletion that can hinder data analytics and mining
Divanis et al. [73]	Practical	Unified framework that satisfies multiple privacy requirements and incurs less IL	Less applicability to heterogeneous data types, and sensitive itemset disclosure
Awad et al. [74]	Theoretical	Provides higher utility for certain itemset in transactional data using anti-based clustering	The vulnerability analysis of selected itemset is not provided that may expose one/group privacy
Awad et al. [75]	Practical	Creates a neighbor dataset for knowledge discovery/extraction purposes using utility rules	The vulnerability analysis of selected itemset is ignored that may impact one/group privacy
Barakat et al. [76]	Practical	Executed a privacy attack on k^m -anonymity model that can expose the privacy of some user explicitly	Utility analysis and formal proof of the privacy breach on large scale datasets are not provided
WANG et al. [77]	Practical	Sufficient protection for a group of people who have distinct privacy-related preferences in data	Prone to lower utility on special-purpose metrics (i.e., accuracy, precision, recall, F_1 scores, etc.)
Can et al. [78]	Practical	Ensures protection based on distinct privacy-related preferences provided by users to control anonymity	Can lead to higher information loss if data is imbalanced, and values of most QIs are close
Meisam et al. [79]	Practical	Preserving both privacy and utility by creating k views of the trace data	Can lead to higher computing cost when the dataset is large, utility can be poor when data is skewed
Fan et al. [80]	Practical	Effectively preserve the privacy of network flows data by creating synthetic data using GANs	Can lead to higher utility loss when offset between original and synthetic data is high
Meisam et al. [81]	Practical	Preserves privacy of important fields in trace data using pseudonyms and Multiview approach	Yields higher computing complexity by creating multiple views of data, and prone to linking attack
Aleroud et al. [36]	Practical	A DP-based prototype to address the privacy-utility trade-off in network trace data	Subject to personal information disclosure in the presence of auxiliary information
Ahmed et al. [82]	Practical	Strong privacy protection of critical fields in network logs data using condensation-based approach	Prone to low utility results on special purpose metrics (i.e., accuracy, F_1 , etc.) of data mining
Velarde et al. [83]	Practical	Practical solution for anonymization of traffic trace data with better privacy using entropy approaches	Yields poor utility when most of the data belongs to distinct regions and # of fields are large
Shaham et al. [84]	Practical	Strong privacy protection in location data sharing, and applicable to medical records and web analysis	Less resilience against knowledge graph-powered attacks as well linking attack using auxiliary data
Mahajan et al. [85]	Conceptual	Enables keyword searches on encrypted data with better privacy using k -means clustering approach	Does not provide provable utility in terms of information loss, accuracy, and F_1 score on diverse data
Li et al. [86]	Practical	A low-cost solution for extracting and anonymizing sensitive data items from documents	Lack of validation and testing on real-world (i.e., PHH data) and large scale medical documents
Li et al. [87]	Practical	Developed a practical solution for identifying, summarizing, and report generation from health data	Prone to repeated query attacks using same noise for some queries, and can reveal true values
Kong et al. [88]	Practical	Privacy preservation of documents and multiple data items including features, metadata, and text	Evaluation was conducted on static data, there exist a possibility of true value disclosure
Garat et al. [89]	Practical	A corpus-based method for privacy preservation of court documents and sensitive data items in them	Requires a very large # of documents (e.g., up to 80K) for good performance, and complexity is high
Li et al. [90]	Practical	Robust privacy protection of medical data by concealing potentially identifying health data items	Poor utility when original data to be anonymized is in scattered form, and values are highly dissimilar
Lima et al. [91]	Practical	A robust three-step privacy-preserving solution for documents data which can be used in medical data	Does not provide additional support for languages other than clinical text written in Spanish language
Liu et al. [92]	Practical	A practical approach for bag-valued data with better data utility using semantic similarity concept	Prone to identity, SA, and membership disclosure by not identifying the vulnerable data items
Liu et al. [93]	Practical	Ability to anonymize personal data in which some fields values are missing using clustering concepts	Can lead to the disclosure of identities as well as membership when the adversary has known data
Ghemri et al. [94]	Conceptual	Ensures the analytics results/statistics remain same when computed from original/anonymized data	Formal analysis and validation was performed in limited aspects, fewer records were used in tests
Chen et al. [95]	Practical	Robust and efficient solution for anonymizing query logs data with better utility and privacy	Prone to identity and itemset disclosure when a large # of queries mapped to same user
Garcia et al. [96]	Practical	Strong privacy preservation of personal data using dependant and independent attributes information	Limited applicability to categorical data, and disclosure of multivariate statistics during data analytics
Yuvuraj et al. [97]	Practical	Strong privacy preservation of individual using both anonymization and cryptography approaches	Higher computing cost, and limited evaluation against major privacy risks (i.e., data reconstruction)
Meng et al. [98]	Practical	Restricts personal information disclosure without sacrificing data utility in web search data	Yields limited analysis of data, and prone to higher privacy leakage without doing sensitivity analysis
Famies et al. [99]	Practical	Privacy preservation of query logs by anonymizing sensitive data items using dynamic analysis	Efficacy in real-time environments has not been tested, and computing complexity is much higher
Ullah et al. [100]	Practical	A practical solution towards search queries anonymization to protect privacy of users in search engines	Limited scalability tests were performed as the efficacy of method was tested with 1000 users only
Nasab et al. [46]	Practical	Ensures better privacy in anonymizing real-time IoT data of two types (i.e., categorical & numerical)	Disclosure of SA is possible with higher probability by not considering the diversity of SA values
Kumar et al. [47]	Conceptual	Strong preservation of user's privacy in transactional data mining via sliding window addition concept	Less applicability to diverse datasets and prone to linkage attacks in the presence of auxiliary data
Veron et al. [101]	Practical	Better privacy protection of location data using three technologies in cloud computing environments	May lead to poor data utility when obtained data is noisy and contain skewed values for some QIs
Yang et al. [102]	Conceptual	Better resolution of the privacy-utility trade-off in data stemming from IoT environments	Anonymized data may hinder the knowledge discovery as well as new hypothesis generation
Patil et al. [44]	Practical	Privacy preservation of user in mining news data streams related to crimes using k -anonymity concept	Yields constrained analysis of data and prone to identity disclosure via background knowledge
Tekli et al. [103]	Practical	Strong privacy protection when original data contains more than one records about the same individual	Poor data utility by enforcing strict privacy parameters (e.g., k, l) during anonymization
Honda et al. [104]	Conceptual	A strong privacy preservation method for facial image data using k -anonymity concept	Validation was limited to only fewer tests that may hinder solution's progress in complex cases
Yang et al. [105]	Conceptual	Strong privacy protection of facial images in real-time scenarios over encrypted outsourced dataset	Prone to individual and group privacy disclosures by linking extracted features with open data
Zhang et al. [106]	Practical	Ensure privacy guarantees in videos data using blurring algorithm that hides salient parts	May lead to privacy disclosure if background contains sensitive information about individual
Le et al. [107]	Practical	Robust and practical solution for adjusting the degree of privacy-utility while anonymizing image data	May yield inconsistent results on data that is partially anonymized due to policies or regulations
Grossel et al. [108]	Practical	An intelligent mechanism to selectively anonymize selective parts in image data for privacy protection	Less applicable to medical environments due to very high diversity in data styles and templates
Ren et al. [109]	Practical	Strong privacy preservation of facial attributes without sacrificing action detection accuracy	Prone to higher utility loss, and identity and habits disclosure via linkage attack with auxiliary data
Deivana et al. [110]	Conceptual	Better privacy protection by classifying data and selective anonymization	Extensive comparisons with existing methods are missing, and formal aspects are weak
Bazai et al. [111]	Practical	Efficient implementation of Mondrian algorithm on Spark framework	Computing complexity is significantly high, and is prone to diversity-based attacks

TABLE 3. Summaries of evaluation metrics used by CAMs.

Category	Famous Evaluation Metrics	Rep. Studies
Privacy	Disclosure risk, probabilistic disclosure for SA/identity, record linkage, table linkage, privacy-sensitive (PS) rules protection, inference preservation, thwarting prediction against SA/attributes, community privacy preservation, statistical disclosure control, distribution disclosure, trajectory info hiding, vertex and edge disclosure protection, entropy leakage, association rule hiding, content hiding, and privacy preservation from active attacks, etc.	[21], [116], [117], [118], [119]
Utility	Information loss, distortion, accuracy, F -measure, coverage of generalized values, discernability metrics, information theoretic metrics, size of equivalence classes, degree preservation, precision, normalized mutual information, recall, clustering coefficient, KL -divergence, queries' accuracy, network resilience, centralities, amount of knowledge, authority score, effective diameter, and data mining tasks, etc.	[21], [120], [121], [122], [123], [124]

of privacy threats, and the ever-changing landscape of attacker capabilities, the development of accurate privacy and utility evaluation metrics has become more urgent than ever. Lastly, we present a quantitative analysis (e.g., average results) of SOTA studies included in each category (i.e., tables/graphs) in terms of privacy preservation and utility enhancement in Figure 20. From the analysis, it can be observed that CAMs can improve the privacy and

utility results significantly. The higher improvements in the utility results are due to the distance/similarity concepts adoption in the clustering process. Through quantitative analysis of each study, we found that the lowest and highest values of the utility improvements were 5% and 90%, respectively. In contrast, the lowest and highest improvements in privacy results were 3.1% and 35 %, respectively.

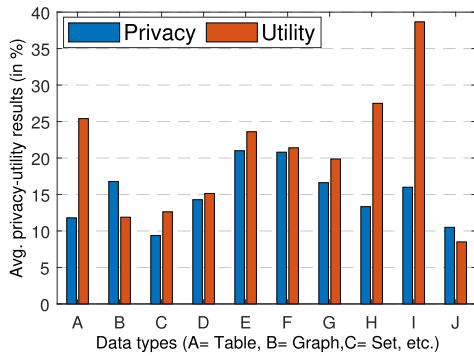


FIGURE 20. Quantitative analysis of CAMs proposed for different data types.

V. SIGNIFICANCE OF CAMs IN DIFFERENT COMPUTING PARADIGMS

In this section, we emphasize the significance of CAMs in multiple computing paradigms with regard to privacy preservation in different contexts. For example, in SN, CAMs not only help with privacy preservation in data publishing, but they are also used to support multiple applications involving personal data (e.g., community detection/clustering, information diffusion, privacy-aware graph

computation, and sensitive topic diffusion, to name a few). Hence, it is vital to provide thorough perspectives on CAMs in different emerging computing paradigms along with recent SOTA approaches. We demonstrate an overview of the five emerging computing paradigms along with concise details of data sources in Figure 21. In Table 4, we describe the significance of CAMs for practical applications/services in each computing paradigm shown in Figure 21.

As shown in Table 4, CAMs have played a vital role in multiple emerging computing paradigms in different contexts. Many sectors benefit from CAMs, including health-care, SN service providers, recommender systems, third-party apps, data mining infrastructures, intelligent services, multi-party computations, policymakers, researchers, and cloud-based services. In the coming years, CAMs can play a vital role in preserving privacy of AI-based systems, such as federated learning, swarm learning, and federated analytics. The synergy of CAMs with these emerging technologies can protect AI-based systems and the associated data (e.g., models, parameters, and underlying data). Furthermore, CAMs have been increasingly applied to heterogeneous data formats for privacy preservation and utility enhancements.

Apart from the strengths and weaknesses, we present a quantitative analysis (e.g., average results) of SOTA studies included in each computing paradigm in terms of privacy preservation and utility enhancement in Figure 22. From

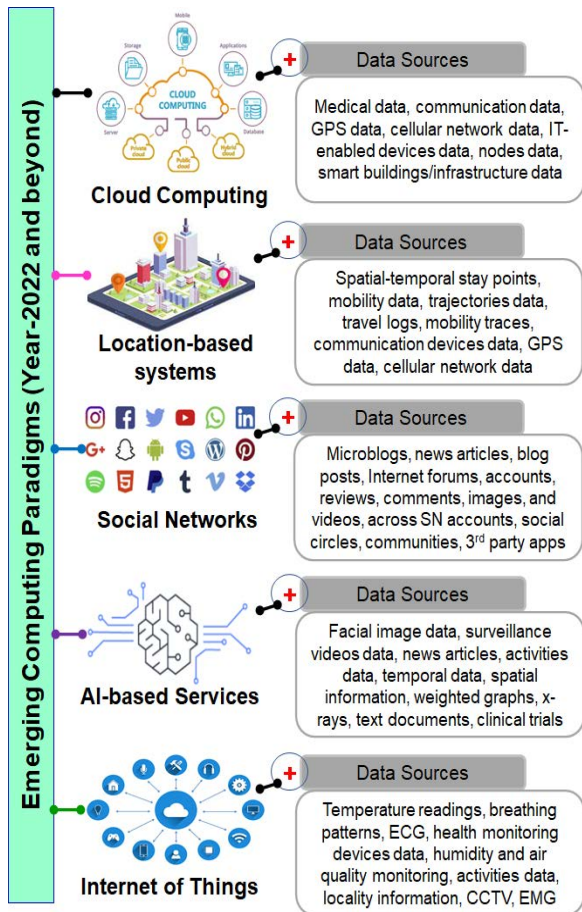


FIGURE 21. Overview of emerging computing paradigms.

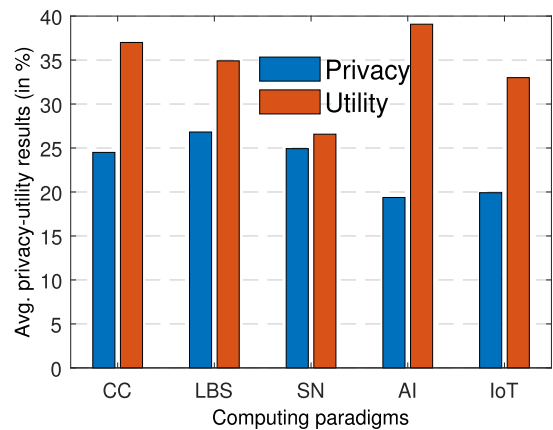


FIGURE 22. Quantitative analysis of CAMs proposed for different computing paradigms.

the analysis, it can be observed that CAMs can improve both the privacy and utility results significantly. Through quantitative analysis of each study, we found that the lowest and highest values of the utility improvements were 5.3% and 96%, respectively. In contrast, the lowest and highest improvements in privacy results were 2.1% and 75 %, respectively.

The technical knowledge presented in Sections IV and V was discovered in six ways, (i) detailed analysis of the experiment section of many SOTA studies, (ii) analysis of datasets used in the evaluation of CAMs, (iii) analysis of the recent anonymity tools, (iv) use cases designed to compare

TABLE 4. Significance of CAMs in multiple emerging computing paradigms (mainly regarding privacy preservation).

Computing Paradigm	Objective Achieved	Clustering Type Used	Practical Application/Service	Rep. Studies
Cloud Computing	Reduction in the cost of data storage	K -Means clustering	Data mining	Logeswari et al. [134]
	Scalability in heterogeneous data anonymization	Parallel Clustering	Big data handling	Usha et al. [135]
	Control on proximity privacy breaches	t -ancestors clustering	Medical applications	Zhang et al. [136]
	Protected from similarity and inference attacks	(G, S) clustering	Data mining and analysis	Nayahi et al. [137]
	Ensured the higher confidentiality of data	KNN(G, S) clustering	Privacy-preserving keyword search	Singh et al. [138]
	Privacy-preserving cluster analysis	k -means clustering	Collaborative data analysis	Lekshmy et al. [139]
	Privacy-preserving healthcare data sharing	ECDSA algorithm	Healthcare industry	Jayaraman et al. [140]
	Maintained privacy of data in cloud environment	Dragonfly (DF) algorithm	Data analytics	Madan et al. [141]
	Data protection from 3rd parties	Adaptive Dragon-PSO	Data analytics	Madan et al. [142]
	Privacy-preserved data storage and processing	Fuzzy C means	Medical applications	Shanmuga et al. [143]
Location-based Services	Privacy preservation in moving objects data	k -member clustering	Query answering	Abul et al. [144]
	Privacy protection of location information	Hana clustering	Query answering	Fei et al. [145]
	Privacy mechanism to control locations tracking	Hierarchical clustering	Location-based services	Lee et al. [146]
	Strong privacy protection in geo-matching attacks	k -implicit clustering	Location-based services	Niu et al. [147]
	Protection of location privacy in data	Clustering k -anonymity	Mobile applications	Yao et al. [148]
	Protection of location and query privacy	Enhanced clustering cloak	Location based services	Lin et al. [149]
	Strongly protected users' location privacy	Spatial cloaking	Location-based services	Zhang et al. [150]
	Maintain privacy of location, choices, & comments	Spatial clustering	Communication apps	Altuwaiyan et al. [151]
	Privacy protection of trajectory data	Greedy clustering	Location-based services	Mahdaviar et al. [152]
	Privacy protection of spatiotemporal locations data	K -means clustering	Location-based services	Dritsas et al. [153]
Social Networks	Personalized privacy preservation in trajectories	K -means clustering	Location-based services	Chen et al. [154]
	Privacy protection in labelled and undirected G	SaNGreeA clustering	Graph mining	Ros et al. [155]
	Privacy protection of SN graph structure adaptively	K -means clustering	Graph analytics	GU et al. [156]
	Privacy of structural properties' values of G	Structural clustering	Graph mining	Truta et al. [157]
	Privacy and structure preservation of communities	SaNGreeA clustering	Community clustering	Campan et al. [158]
	Privacy protection of community data	Partitioned clustering	Community detection/hiding	Chen et al. [159]
	Privacy preservation in answering queries from G	structured clustering	Query-analysis	Ghosh et al. [160]
	Privacy protection of user's connections in a G	Partitioned clustering	Graph analytics	Yu et al. [161]
	Privacy protection of multiple properties of G	k -clustering and SFLA	Structural analysis of G	Gazalian et al. [162]
	Preservation of structural properties of G	Node clustering	Information retrieval	Zhang et al. [163]
AI-based services	Protection from identity disclosure in G	Sub-graph clustering	Hidden knowledge discovery	Liu et al. [164]
	Privacy protection of node, edge, and attributes in G	Enhanced equi-cardinal	Graph analytics	Siddula et al. [165]
	Strong privacy protection of users with flexibility	Policies clustering	Location based SN services	Sai et al. [166]
	Privacy preservation of social connections in G	Edge clustering	Graph analytics	Gao et al. [167]
	Preserving privacy of sensitive labels in G	Node clustering	Graph mining	Yuan et al. [168]
	Strong privacy preservation in IoT scenarios	Federated averaging	Collaborative analytics	Zhao et al. [169]
	Privacy preservation in traffic prediction scenarios	Ensemble clustering	Traffic predictions	Liu et al. [170]
	Privacy preservation of consumers' personal habits	Ring architecture	Smart grid technologies	Badra et al. [171]
	Privacy preservation of smart meter data	k -means clustering	Electricity usage pattern	Wang et al. [172]
	Privacy preservation of structure of customers	ANNs based clustering	Insurance companies	Ghahramani et al. [173]
Internet of Things	Lowering privacy risks in data analysis	Self-organising maps	Large-scale analysis	Mohammed et al. [174]
	Privacy preservation in edge computing	Federated k -Means	Collaborative analytics	Kumar et al. [175]
	Privacy preservation in globally similar data	fuzzy c -means	Collaborative analytics	Stallmann et al. [176]
	Privacy preservation of sensitive rules	Objects clustering	Data mining tasks	Rajesh et al. [177]
	Fostering data re-usability for analysis	Subspace clustering	Effective data mining	Virupaksha et al. [178]
	Privacy preservation in high dimensional data	Multi-label clustering	Cloud-based apps	Bollaa et al. [179]
	Privacy preservation in federated learning	Socially-aware clustering	Wireless networks	Khan et al. [180]
	Privacy-utility trade-off resolution in PDP	Subspace-based clustering	Data mining tasks	Virupaksha et al. [181]
	Privacy preservation of users and clusters	k -means strategy	Healthcare systems	Guo et al. [182]
	Privacy preservation of sensors network data	k -means clustering	Intrusion detection	Zhu et al. [183]
Data Anonymization	Privacy protection of data in healthcare systems	Clustering strategy	Healthcare industry	Almusallam et al. [184]
	Privacy preservation of interval data	Distributed k -means	Cloud-based apps	Huang et al. [185]
	Privacy preservation of real-time and critical data	Graph-based clustering	IoT-based healthcare solution	Elhoseny et al. [186]
	Privacy preservation of sensor nodes data	Layered clustering	collaborative processing	Kumar et al. [187]
	Privacy preservation of events in geo-textual data	Hybrid clustering	Social IoT apps	Shuja et al. [188]
	Ensures privacy of activity information & location	Hybrid clustering	Healthcare/Smart homes	Ogtonbayar et al. [189]
	Privacy preservation of IoT data in cloud computing	Interaction clustering	Cloud security	Patil et al. [190]
	Privacy preservation in patient's data sharing	K -medoid clustering	Medical healthcare-IoTs	Ullah et al. [191]
	Prevents privacy leakage of vehicles data	Node clustering	Internet of vehicles	Li et al. [192]
	Privacy preservation in scattered data	Spatial clustering	Multi-cloud platform	Liu et al. [193]

each study with prior ones, (v) critiques already published in survey article about CAMs, and (vi) analyzing deficiencies in the working methodology of published studies.

VI. DARK SIDE OF CLUSTERING-BASED ANONYMIZATION MECHANISMS

Although CAMs have demonstrated more effectiveness in preserving privacy and utility than traditional anonymization approaches, they can also be used to jeopardize individual privacy. For example, plenty of methods have been proposed based on clustering concepts that can either re-identify people or assist in inferring private information from anonymized graphs/tables. Analysis of the dark side of CAMs can pave the way to securing personal data against prying eyes in a more practical way. Figure 23 presents a generic overview of de-anonymization of published data and inferring SAs from that data.

Zhang et al. [194] described an identity-revelation method based on attributes from the anonymized G . The authors achieved accuracy of up to 80% in de-anonymization of

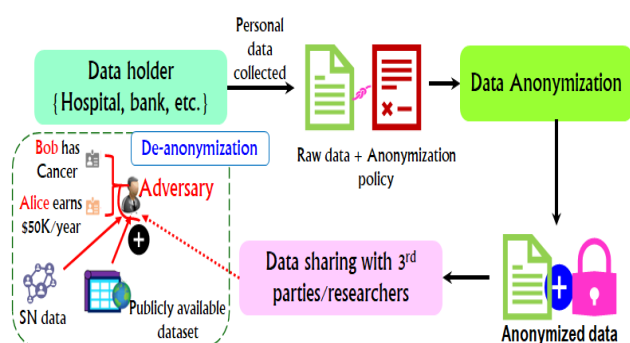


FIGURE 23. Overview of the data de-anonymization process.

identities. Similarly, some de-anonymizing approaches have used the community-clustering concept to group users, and users' de-anonymization was subsequently successful [195]. Some approaches have jointly used clustering and attribute information to correctly identify individuals from privacy-preserved published graphs [196]. Shao et al. [197]

TABLE 5. Detailed analysis (i.e., strengths and weaknesses) of the SOTA studies presented in Table 4.

Computing Paradigm	Rep. Studies	Strengths	Weaknesses	Study nature
Cloud Computing	Logeswari et al. [134]	Efficient clustering and enable collaborate learning of medical data with privacy preservation	Evaluation was carried out using only 2K records and formal aspects are not discussed	Conceptual
	Usha et al. [135]	Effective solution towards privacy-utility trade-off in big data environment using parallelism	Prone to data leakage by combining multiple SA from heterogeneous data and high complexity	Practical
	Zhang et al. [136]	Highly efficient and scalable solution towards big data anonymization using k -ancestor clustering	Weaker assumptions towards data availability at external sources which can lead to disclosures	Practical
	Nayahi et al. [137]	Strong resistance to major privacy threats, and flexible tailoring to any domain and datasets	Utility of anonymized data can be poor due to the higher noise addition in some cases	Practical
	Singh et al. [138]	Robust solution towards maintaining the confidentiality of the data in cloud environments	Missing discussion about classifying about what constitutes as sensitive data in the cloud	Theoretical
	Lekshmy et al. [139]	Cloud-simulator based implementation to preserve users privacy in data mining tasks	Pre-mature convergence when data is skewed, and high difficulty in selecting parameters' values	Practical
	Jayaraman et al. [140]	Robust solution for maintaining the integrity of confidential data in cloud environments	The computing complexity of the method used to identify likely security incidents is very high	Theoretical
	Madan et al. [141]	Optimization of the utility-privacy trade-off using fitness function of dragonfly algorithm	Prone to identity and SA disclosure when data is imbalanced, and values range is small	Practical
	Madan et al. [142]	Better utility of anonymized data under many constant anonymization parameters in the cloud	Prone to skewedness, homogeneity, and similarity attacks by not using the SA's values diversity	Practical
	Shammuga et al. [143]	Effective protection of medical big data by jointly using clustering and access control	Less reliable query results for data mining tasks, and low accuracy on special purpose metrics	Practical
Location-based Services	Abul et al. [144]	Strong privacy preservation in moving databases using k -anonymity and co-localization	Insufficient to resist location disclosure of a user group, and prone to hidden profiling of users	Practical
	Fei et al. [145]	Low-cost schema to be used in client-server settings for privacy protection of trajectory data	Poor utility and misleading analytics results in most cases by adding dummy location data	Practical
	Lee et al. [146]	An efficient solution for lowering identity and SA disclosures in location data using clustering	Lack of assumption regarding the auxiliary data availability, and poor data utility	Practical
	Niu et al. [147]	Robust answers to dynamic queries without compromising user's privacy using tags/sequences	In some cases, the data can be rendered useless due to higher noise addition (i.e., ϵ is low)	Practical
	Yao et al. [148]	Personalized privacy preservation using k -anonymity based clustering (CK) in location services	Computing complexity is high, data utility can be low when # of preferences are large	Practical
	Lin et al. [149]	Strong privacy preservation by decoupling the requested contents and location position	Inability to address group privacy issues as well as data reconstruction attack at server-side	Practical
	Zhang et al. [150]	Strong privacy preservation of queries data using semantic information and cloaking concept	Prone to intent and SA disclosure by using the combine information of data and location	Conceptual
	Altuwaiyan et al. [151]	Strong privacy preservation of location data, request contents, and user's positions	Limited experimental evaluation was performed, and comments sensitivity analysis is not given	Conceptual
	Mahdavian et al. [152]	Personalized privacy protection of location data considering moving objects requirements	Prone to identity and SA disclosure when certain users specify lower privacy preferences	Practical
	Dritsas et al. [153]	Strong privacy protection of mobile users data by defining a new metric (i.e., vulnerability)	Prone to identity and SA disclosures when cluster cannot meet the diversity requirements	Practical
Social Networks	Chen et al. [154]	Sufficient protection of basic and sensitive data items against background knowledge attacks	Prone to record, SA, and table linkage due to the existence of auxiliary data	Practical
	Ros et al. [155]	Generic solution towards SN data mining with privacy protection for recommendation purposes	Prone to linkage attack, and limited applicability to other graph's types (directed, weighted, etc.)	Practical
	GU et al. [156]	Strong resistance against background knowledge attacks and limits SA/identity disclosure	Prone to higher utility loss by introducing heavier changes in the structure of graph	Practical
	Truta et al. [157]	Fostering social mining and graph analytics by making less changes in the graph structure	Prone to community privacy disclosure, and SA in the case of attributed SN	Practical
	Campan et al. [158]	Ensure strong protection against community privacy disclosure, and control heavier changes	Prone to individual's privacy disclosure when # of users in each community are large	Practical
	Chen et al. [159]	Personalized privacy preservation of users in the community by link perturbation techniques	Decrease the reliability of extracted information from graph due to noises in the form of links	Practical
	Ghosh et al. [160]	Strong defense against identity, SA, and membership attacks by encrypting graph structure	Prone to higher complexity when graph size is large, and induce heavier changes in graph	Practical
	Yu et al. [161]	Provides strong protection against linkage attack and preserves the integrity of sensitive edges	Can lead to node privacy breaches as well as the SA in attributed social network graph	Practical
	Gazalian et al. [162]	Protection against multiple threats such as Identity, SA, membership, and graph linkage	Prone to poor utility in SN mining and analysis, structural modifications are very high	Practical
	Zhang et al. [163]	Strong protection in 1-neighbourhood attacks in privacy-preserving graph data publishing	Heavier changes in the anonymized graph in order to meet the constraints values	Practical
AI-based services	Liu et al. [164]	Strong defense against identity disclosure problem using k -possible anonymity concept	Prone to SA and membership attacks by not ignoring the diversity of sensitive information	Practical
	Siddula et al. [165]	Strong privacy protection of the whole network against linkage attack from auxiliary data	Can lead to the disclosure of identity, SA, and membership by not considering user-level privacy	Practical
	Sai et al. [166]	Provides customized settings for better control of privacy preservation in location data of SNS	Prone to a # of attacks for the users who set loose privacy settings or unaware about privacy	Practical
	Gao et al. [167]	Ensures privacy protection of nodes and edges by adding exponential noises in the data	Yields lower utility when graph size is relatively small and prone to data reconstruction	Practical
	Yuan et al. [168]	Ensures strong protection against attributes inference attacks and is applicable to diverse graphs	Largely destroys the structure of the graph by enforcing the strict parameters (i.e., k and f)	Practical
	Zhao et al. [169]	Strong privacy protection of users data, parameter, and communication in dynamic scenarios	Prone to data derivation, prediction, and re-construction attacks when the adversary is in system	Practical
	Liu et al. [170]	Strong privacy protection in collaborative machine learning without data dissemination	Prone to parameters/data reconstruction attacks and wrong utility analysis during analytics	Practical
	Badra et al. [171]	Strong protection of user's billing information in energy sector using encryption	Can lead to higher computing complexity and do not assist in searching from encrypted data	Conceptual
	Wang et al. [172]	Effective solutions towards privacy preservation in heterogeneous data using federated k -means	Prone to privacy breaches when the data is relatively small and belong to an identical sector/domain	Practical
	Ghahramani et al. [173]	Strong solution towards hiding privacy-sensitive patterns in insurance company data	User- and group level privacy breaches cannot be effectively prevented from the segmented area	Practical
Internet of Things	Mohammed et al. [174]	Strong privacy preservation of users from high dimensional social networks data	Prone to the disclosure of group privacy as well meta data of graph that can be used to infer SA	Practical
	Kumar et al. [175]	Privacy protection of the individual as well group metadata and original data across platforms	Less applicability to diverse data types and prone to identity, SA, and membership disclosures	Practical
	Stallmann et al. [176]	Strong privacy protection of local data via clustering in federated learning scenarios	Can lead to the disclosure of local data or gradients by not adding noise during transfer	Practical
	Rajesh et al. [177]	Ensures privacy protection in association rule mining cases via perturbation-based approach	Less resilience against SA reconstruction, derivation, and prediction in medical environments	Practical
	Virupaksha et al. [178]	Overcome the issues of invalid and ineffective data mining results by adding less noise to data	Prone to privacy breaches when the adversary has auxiliary data or background knowledge	Practical
	Bolla et al. [179]	Privacy preservation of SA disclosure by identifying and generalizing sensitive data in clusters	Can lead to infeasible query results as well as inaccurate data-mining/analyses results	Practical
	Khan et al. [180]	Reduction in client-side privacy breaches in federated learning by not centralizing local data	Fails to provide a strong defense against the data poisoning as well reconstruction attacks	Practical
	Virupaksha et al. [181]	Enhances the quality of anonymized data by adding noise along each dimension of the data	Prone to explicit disclosures of user's identity and SA by not changing position of data items	Practical
	Gao et al. [182]	Strong privacy protection at user-level as well as cluster centers-level by using encryption	Failed to provide resilience against probabilistic, skewedness, similarity, and homogeneity attacks	Practical
	Zhu et al. [183]	Higher accuracy in reducing multiple attacks stemming from network traffic using k -means	Failed to address group-privacy issues, and complexity is high when data is high dimensional	Practical
Internet of Things	Annusalam et al. [184]	Detailed discussion of privacy attacks originating in smart healthcare (IoT & edge healthcare)	Limited discussion about the real-time data processing and corresponding privacy attacks	Theoretical
	Huang et al. [185]	Robust privacy guarantees in small and large-scale interval data stemming from IoT devices	Higher computing complexity, limited tests, and less defense against hidden profiling attacks	Practical
	Ehosseny et al. [186]	Effective privacy preservation of real-time data transmitting between IoT devices to gateway	Prone to identity, or SA disclosure by not classifying sensitive/non-sensitive data before sending	Practical
	Kumar et al. [187]	Strong privacy preservation of IoT sensor nodes data without disclosing semantic information	Communication cost is very high and privacy issues can occur when values range is small	Practical
	Shuja et al. [188]	Strong users data privacy protection by not computing similarity/distance among all data points	Fails to handle complex and high dimensional data such as temporal sequences or locality traces	Practical
	Ogbonbaye et al. [189]	Enable anonymization of dynamic, incomplete, and high dimensional data without privacy loss	Some statistics such as vulnerability/utility-levels of data items cannot be computed in real-time	Practical
	Patil et al. [190]	Minimizes privacy violations that can stem from the personal data originating from smart homes	Damage the utility of less sensitive data by not using AI methods to classify the data nature	Practical
	Ullah et al. [191]	Fosters data re-usability by sharing it at a large scale without compromising user's privacy	Prone to identity/SA disclosure when a higher amount of data is available at auxiliary sources	Practical
	Li et al. [192]	Better privacy preservation by solving data island problem by sharing location data of vehicles	Less adoption in real-world cases by not considering the data diversity and heterogeneous styles	Practical
	Liu et al. [193]	Strong protection against SA inference attacks by incentivizing users to form cohesive clusters	Prone to explicit disclosures of identity/SA when # of users in each cluster are significantly small	Practical

proposed a robust de-anonymization method based on structural information from a published graph. Figure 24 illustrates an overview of SN data de-anonymization. In figures 24 (b) and (c), users' location information can be inferred by linking anonymized and crawled networks, respectively. Similarly, clustering concepts are employed to group similar/dissimilar people in order to infer their private information by employing background knowledge or auxiliary graphs.

Due to the rapid developments in SN services, user de-anonymization within an SN site and across SN sites has become a very hot research topic. In line with the trends, we summarize the contributions of clustering-based de-anonymization methods in different computing environments, along with their data types, in Table 6. The analysis presented in Table 6 provides another perspective on CAMs (i.e., de-anonymization of users and their corresponding personal information) that has remained unexplored in the recent literature. By understanding the dynamics of such research from the attackers' perspectives, more secure and resilient anonymity methods can be developed to preserve users' privacy. Furthermore, these kinds of analyses provide a better overview of the research gaps to aid researchers who are working on the defense side.

In Table 6, we compared various methods based on four parameters (i.e., data/items exploited in de-anonymization, objectives achieved in compromising user privacy, clustering concepts employed, and target applications/services). The

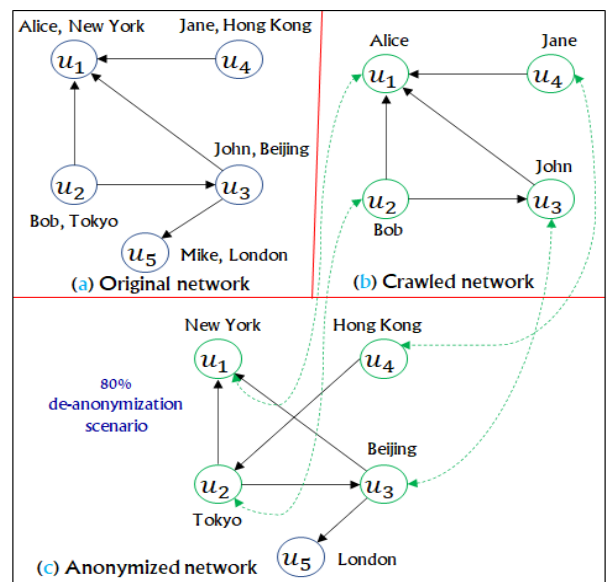


FIGURE 24. Overview of SN de-anonymization by leveraging auxiliary data (adapted from [197]).

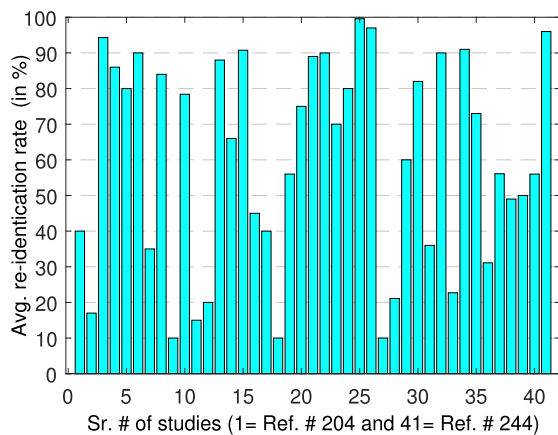
last column of the table included pertinent studies from which detailed contents can be gathered for an in-depth investigation of each method.

Key limitations (i.e., weaknesses) of each study listed in Table 6 are discussed in Table 7. This extended knowledge

TABLE 6. State-of-the-art clustering-based de-anonymization approaches employed to breach privacy.

Data used	Objective Achieved (key strengths)	Clustering Concept Used	Target Application/Service	Representative Studies
Geolocated data	Inference attack on mobility data	Hierarchical clustering	Location-based services	Gambis et al. [198]
SN data (i.e., graph)	Network de-anonymization with higher matches	Node clustering	Social network services	Chiasserini et al. [199]
SN data (i.e., graph)	Network de-anonymization via user relations	Community clustering	Social network services	Chiasserini et al. [200]
SN data (i.e., graph)	Network de-anonymization via graph matching	Node Clustering	Social network services	Chiasserini et al. [201]
SN data (i.e., graph)	Re-identifying users in anonymized G by mapping	Partitional clustering	Social network services	Fu et al. [202]
SN data (i.e., graph)	De-anonymizing users' identities by mapping	Community clustering	Social network services	Fu et al. [203]
Trajectory data	Re-identifying the stay points of users	k -means clustering	Cellular networks	Francia et al. [204]
Model updates data	Re-identification of participating devices data	k -means clustering	Collaborative learning	Orekondy et al. [205]
Trace data	Re-identifying mobility patterns of people	Density-based clustering	Location-based services	Chen et al. [206]
Mobility Traces	Re-identifying people movement	Markov cluster	Location-based services	Murakami et al. [207]
SN data (i.e., graph)	De-anonymization of heterogeneous SNS	Hierarchical clustering	Social network services	Li et al. [208]
Trajectories data	Unique identification of people from data	Density-based clustering	Location-based services	Zhen et al. [209]
Logs data	Privacy breaches from cookie logs data	Hierarchical clustering	Location-based services	Wang et al. [210]
SN data (i.e., graph)	Users re-identification across networks	Sub-graph clustering	Social network services	Zhang et al. [211]
SN data (i.e., graph)	Compromise of link/connection privacy in G	Collaborative clustering	Social network services	Chen et al. [212]
Location data	Identity disclosure from multiple data	Density-based clustering	Location-based services	Ma et al. [213]
SN data (i.e., graph)	Reduction in the anonymity of users	Sub-graph clustering	Social network services	Nilizadeh et al. [214]
Time series data	Loss of privacy due to matching	Statistical matching	Internet of things	Takbiri et al. [215]
Queries data	Identifying people uniquely via queries	Users grouping	Knowledge-based systems	Shirani et al. [216]
SN data (i.e., graph)	Enhancement of success rate of identity disclosure	Seed-based grouping	Social network services	Aliakbari et al. [217]
Transactional data	De-anonymity of people in crypto currency	Heuristic clustering	Blockchain-based services	Xueshuo et al. [218]
SN data (i.e., graph)	User's identification across SNS	Structural clustering	Social network services	Li et al. [219]
Trajectory data	Privacy breaches on location data via communities	Community-based clustering	Location-based services	Wang et al. [220]
Mobility data	Full trajectory recovery in aggregated mobility data	Trajectory clustering	Location-based services	Tu et al. [221]
Multimedia data	Rapid prediction of re-identification risk	Random sampling	Multimedia applications	Yang et al. [222]
Network packets	User recognition based on network traffic	Temporal clustering	Web applications	Miculan et al. [223]
TCP packets	Users re-identification from the web traffic data	Temporal clustering	Web applications	Nardin et al. [224]
GPS trajectories	Users re-identification from the trajectory data	Heuristics-based clustering	Location-based services	Naini et al. [225]
SN data (i.e., graph)	Re-identification of SA from the graph	Structural clustering	Social network services	Tian et al. [226]
Tabular data	User's re-identification from multiple data	IRM clustering	Recommendation apps	Iwata et al. [227]
Call records	Identification of people from anonymized CDR	k means clustering	Cellular networks	Cecaj et al. [228]
CDR and graph	Re-identification of user with 90% accuracy	Community clustering	Mobile SN	Cecaj et al. [229]
Music preferences	Re-identification of user from rating data	k means clustering	Recommendation systems	Hirschprung et al. [230]
Smart contracts	Re-identification of user in blockchain network	Address clustering	Blockchain-based services	Linoy et al. [231]
SN data (i.e., graph)	Re-identification of user from large G	Structural clustering	Social network services	Sharad et al. [232]
Image data	Re-identification of user's faces from larges	Feature-based clustering	Social network services	Acquisti et al. [233]
Masked data	Re-identification of people from masked data	Attribute-based clustering	Cyber-physical systems	Huang et al. [234]
SN data (i.e., graph)	Disclosure of users' private information	Structural clustering	Social network services	Chen et al. [235]
Transaction data	Reconstruction of original transaction data	Semantic-based clustering	Medical studies	Ong et al. [236]
Smartphone data	Accurate recovery of individuals' mobility data	KNN-based clustering	Cellular networks	Lin et al. [237]
Spatiotemporal data	Disclosure of people's mobility patterns/sequences	Swarm-based clustering	Cyber-physical systems	Castro et al. [238]

demonstrates that CAMs can be used to infer the identity/SA of the user with significantly higher %ages using various kinds of data available at external sources (i.e., online repositories, social networks, web searches, internet traffic logs, etc.). However, the utilization of a strong privacy mechanism and lower availability of auxiliary data can restrict the re-identification rate.

**FIGURE 25. Quantitative analysis of clustering-based de-anonymization approaches.**

Apart from the strengths and weaknesses, we present a quantitative analysis (e.g., average re-identification rate results) of SOTA studies included in Table 6 and 7 in Figure 25. From the analysis, it can be observed that de-anonymization approaches can significantly impact the

TABLE 7. Limitations of de-anonymization approaches listed in Table 6.

SOTA study	Key limitation (s)
Gambis et al. [198]	Poor performance on highly skewed (or non i.i.d) training data
Chiasserini et al. [199]	De-anonymization rate drops when # of users in each cluster increase
Chiasserini et al. [200]	Less applicability to other types (i.e., weighted, attributed, etc.) of graphs
Chiasserini et al. [201]	Yields poor performance when total variations in graph structures are high
Fu et al. [202]	Poor convergence when nodes degree or variations in attributes' values is high
Fu et al. [203]	Optimal mapping conditions cannot be met in non-overlapping community cases
Francia et al. [204]	De-anonymization rate drops when anonymization is made through DP model
Orekondy et al. [205]	The performance can be severely impacted if no external data is available
Chen et al. [206]	Yields poor performance in the presence of outliers/misaligned feature space
Murakami et al. [207]	Use different variants (≈ 20) of user's locations to perform de-anonymization
Li et al. [208]	De-anonymization rate drops significantly when no identical graphs are available
Zhen et al. [209]	Extensive comparisons are required to infer SA when most users are similar
Wang et al. [210]	Heavily relies on exogenous records to perform de-anonymization of data
Zhang et al. [211]	Waste of computing time when same users are not available in both graphs
Chen et al. [212]	The algorithmic complexity is high, and dramatically increases with graph size
Ma et al. [213]	Many operations are performed to infer SA, and the solution is not generic
Nilizadeh et al. [214]	Poor performance in the case of overlapping communities, or sparse graph cases
Takbiri et al. [215]	Lack of numerical tests and uses multiple assumptions to perform SA's inference
Shirani et al. [216]	The # of queries can significantly increase when the number of users is very large
Aliakbari et al. [217]	Poor results in terms of matching and computing time when seeds are erroneous
Xueshuo et al. [218]	Significant efforts are needed to convert complex data into structured data
Li et al. [219]	Poor performance when most profile attributes are not visible externally
Wang et al. [220]	Matching rate drops significantly in the presence of noises and mismatches
Tu et al. [221]	Prono to less matching when co-relation among aggregated data is low
Yang et al. [222]	Heavily depends on the auxiliary data to perform user's de-anonymization
Miculan et al. [223]	Prono to poor performance when users perform searches in a dissimilar way
Nardin et al. [224]	# of successful matches decrease when most data is in the encrypted form
Naini et al. [225]	Poor performance when data is overly anonymized with strong anonymity model
Tian et al. [226]	Prono to fewer matches when graph structure is not aligned with external graph
Iwata et al. [227]	Yields poor performance in node-level (one user) de-anonymization from graph
Cecaj et al. [228]	Prono to poor performance when dummy records are present across datasets
Cecaj et al. [229]	Yields infeasible results when most data cannot be collected due to regulation
Hirschprung et al. [230]	Requires extensive matching and analysis to infer the user's identity/SA
Linoy et al. [231]	Yield infeasible results when diversity among users in terms of attributes is high
Sharad et al. [232]	The applicability of approach on other similar datasets is not possible
Acquisti et al. [233]	Cannot guarantee consistent results when images have poor visibility or tilted
Huang et al. [234]	Requires a significantly large amount of data in order to compromise privacy
Chen et al. [235]	Less applicability to other types of social graphs and higher time complexity
Ong et al. [236]	Requires a substantial # of records to be present in Google repository
Lin et al. [237]	Poor performance when most data items are hidden based on privacy policies
Castro et al. [238]	Yields poor results when diversity among features is high (i.e., low similarity)

privacy of users. From the results, the lowest and highest re-identification rate were 10% and 99.6 %, respectively.

VII. CHALLENGES OF CAMs AND FUTURE RESEARCH DIRECTIONS

In this section, we highlight the technical challenges of CAMs regarding user's privacy preservation in recent times, and we

provide promising avenues for future research taking into account emerging computing systems.

A. OPEN CHALLENGES IN PERSEVERING USER PRIVACY LEVERAGING CAMs

Due to the rapid increase in digital-solution use and adoption, privacy protection has become more challenging. Owing to pervasive technology developments, many users are deeply concerned about privacy and the responsible use of their personal information. Since sensitive data of all kinds about an individual's daily activities and schedules can easily be collected now, there is a risk of intimate detail disclosures. The rate of personal data collection is increasing at a significantly rapid pace, and the scale and number of privacy breaches are likely to increase in the coming years. Hence, there is an emerging need to upgrade the existing defense mechanisms and to propose new, sophisticated, privacy-enhancing technologies. In Figure 26, we present a high level description of different open challenges in information privacy domain.

Open Technical Challenges in Information Privacy Domain
Quantification of attributes impact (i.e., vulnerability) on privacy and utility
Lack of resilience against group privacy while preserving individual's privacy
Ensuring privacy and utility guarantees while anonymizing imbalanced data
Flexible anonymity solutions that remain applicable to heterogeneous data types
Effective resolution of privacy-versus-equity trade-off not only privacy-utility
Tailoring the objective function of clustering to utility and privacy expectations
Reducing the computational complexity, # of iterations, and time of the CAMs
Provision of a strong defense against AI-powered attacks through anonymity
Realization of most anonymization principles (i.e., I-diversity) through CAMs
Yielding consistent performance when data contain outliers or highly noisy
Ensuring anonymization of data stemming from multiple sources(i.e., IoT, etc.)
Improving security and privacy of AI-based systems via anonymity techniques
Adaptive configuration of privacy as well as clustering parameters based on data
Improving the validity of anonymity results using four kinds of analysis such as internal, external, statistical, and conclusion validity

FIGURE 26. Overview of open challenges in information privacy domain.

We summarize below the details of fourteen unique technical challenges of CAMs in protecting user privacy at present.

- *Quantifying the impacts of user's attributes on privacy and utility:* Most CAMs give equal weight to all attributes in data from a privacy and utility point of view. However, recent research has shown that each item within an attribute has a distinct impact on privacy and utility [239]. For example, a zip code allows locating someone more accurately than race and/or gender. Similarly, gender is more appropriate for making credit-related decisions, rather than age. Hence, quantifying the impacts of a user's attributes, and ensuring protection based on such statistics in the CAMs, is challenging.
- *Hidden disclosure of group privacy:* With the advent of big data, a new threat to information privacy has emerged, named group privacy [240]. Most existing

CAMs provide strong resilience against privacy threats concerning individual privacy. However, they are prone to hidden disclosure of group privacy. For example, clustering based on k -anonymity concepts can preserve the privacy of one person in a group of k users, but it can inevitably hurt group privacy. Hence, controlling group privacy issues while preserving individual privacy when leveraging CAMs is very challenging.

- *Anonymization of imbalanced data:* Generally, most anonymization methods, including CAMs, work well on balanced data (the distribution of most attribute values is uniform). However, due to the rapid developments in AI (e.g., federated learning) and legal measures enforcement, diverse values regarding individuals cannot be collected, leading to imbalanced datasets. In these datasets, the distribution of most attribute values is not uniform, and anonymization can be highly complex [241], [242]. In such circumstances, preserving privacy while sustaining high utility from data anonymized using CAMs is very challenging.
- *Applicability to heterogeneous types of data:* Most CAMs were designed for specific scenarios/applications, and extension to diverse types of data is not straightforward. For example, CAMs proposed for a single SA cannot be directly applied to multiple-SA scenarios. Similarly, CAMs proposed for tables cannot be straightforwardly applied to directed graphs. Hence, making each CAM efficient and applicable to diverse data formats is very challenging.
- *Effective resolution of the privacy-equity trade-off:* In the recent past, utility and privacy were regarded as two conflicting goals. Optimizing for utility can degrade privacy, and vice versa. A lot of research has been conducted to resolve this universal trade-off [243]–[245]. Recently, due to significant advancements in AI techniques, a new trade-off, named privacy-equity, has emerged that can lead to biased and inaccurate decision making about some minor groups [246]. However, solving the privacy-equity trade-off with CAMs is very challenging.
- *Tailoring the objective function of clustering to privacy and utility expectations/goals:* In most cases, the objective function of CAMs usually focuses on grouping similar data items in order to lessen the heavier changes in anonymized data. By doing so, only one metric (e.g., utility) can be improved, and privacy issues such as identity and SA disclosures inevitably occur [56]. How to make the objective function aware of both utility and privacy goals/expectations is very challenging.
- *Reducing the computational complexity of CAMs:* Generally, the clustering process encompasses multiple iterations and many hyperparameters, leading to higher computing complexity while processing high-dimensional datasets. In anonymization, the clustering process usually adopts some anonymity requirements as well (i.e., k users in a cluster/class); hence, computation

complexity increases drastically [247]. Although some efforts have been devoted to lowering the computing complexity in CAMs [248], [249], reducing the computing burdens of CAMs on high-dimensional and large datasets is still very challenging.

- *Ensuring sufficient resilience against AI-powered attacks:* In recent years, due to the proliferation of AI-based systems, privacy breaches have increased significantly because traditional anonymization mechanisms cannot ensure sufficient resilience against AI-powered attacks [250]–[253]. AI-powered attacks can be launched to disclose identities, SAs, and memberships from large and complex datasets with the help of hyperparameter tuning [254]. Hence, there is an emerging need to integrate AI concepts in the anonymization approaches for effective resolution of privacy and utility. However, integrating AI concepts in CAMs to safeguard the privacy of individuals from multiple perspectives is very challenging.
- *Adaptation of CAMs to more anonymization principles:* In the published literature, most CAMs have created synergy with the k -anonymity concept in order to preserve user privacy in different settings [255]. Moreover, the k -anonymity concept is relatively weak at resisting many contemporary privacy threats. Therefore, establishing synergy in CAMs with more anonymization principles (e.g., ϵ -DP) has become more urgent than ever. However, establishing synergy between CAMs and other sophisticated anonymization principles is challenging due to the many differences in algorithm designs. Furthermore, guaranteeing the construct validity of these synergies is challenging due to higher variations in personal data formats across domains/applications.
- *Consistent performance in the presence of outliers:* The presence of outliers (out-of-range values) in the data can significantly increase the complexity of the anonymization process, and the resulting anonymized dataset can yield poor utility. Most traditional algorithms, such as k -anonymity, ℓ -diversity, and t -closeness, cannot guarantee consistent performance when the original data encompass outliers [256]. Furthermore, CAM performance on data that contain outliers can be degraded, and convergence cannot be achieved in a reasonable time. Recently, some CAMs have been proposed to efficiently detect outliers and minimize their impact on the clustering process [257], [258]. However, devising low-cost CAMs that can perform well on data with outliers is still very challenging and requires further development from the research community.
- *Heterogeneous source data anonymization using CAMs:* In some real-world computing environments (e.g., the IoT, IoMT, and IIoT), a huge amount of data is collected from heterogeneous sources for analytical purposes. These data play a vital role in pattern extraction leading to effective and accurate decision making. However, anonymization mechanisms based on clustering concepts are paramount in such environments in order to alleviate privacy concerns [135]. Recently, some parallel clustering algorithms have been devised to address data diversity and heterogeneity issues during anonymization [259]–[261]. Moreover, the application of CAMs on data originating from heterogeneous sources is challenging due to the huge diversity in data formats and correlations between tuples. In recent years, personal data anonymization originating from different devices in the form of distributed streams has become a popular research topic [262], [263]. However, the application of CAMs to such data is challenging due to temporal differences in the stream order.
- *Privacy preservation of AI-based systems/ infrastructures through CAMs:* In recent years, there has been an increasing focus on privacy preservation of AI-based systems such as federated learning, deep learning, and centralized machine learning [264]–[268]. These systems have become the target of malevolent adversaries and require privacy preservation of the model's parameters, workflow, and underlying data. The DP approach has been extensively investigated in preserving privacy of AI-based systems/ infrastructures [269]–[273]. However, the application of CAMs in order to preserve AI-based system/infrastructure privacy is challenging due to the fundamental differences in workflows and data types.
- *Adaptive configuration of clustering and privacy parameters in CAMs:* Most CAMs developed for privacy preservation require configuration of clustering (e.g., the number of clusters, the number of iterations, and the optimizing strategy) as well as anonymization parameters (the number of users in a cluster, the similarity/dissimilarity threshold, the value ranges, etc.). These have a significant impact on privacy preservation and utility enhancement, and careful selection of parameters is vital to lowering the complications from the anonymization process. However, devising CAMs with as few parameters as possible without compromising privacy and utility is challenging. In addition, applying optimization strategies to select these parameter values in order to optimize the clustering process is challenging due to the differences in data styles or application features.
- *Verification/validity of internal, external, statistical, and construct validity in CAMs:* Most CAMs that have been developed so far are threat-, domain-, and attack-specific. Hence, their internal, external, statistical, and construct validity cannot be guaranteed in most generic scenarios. Moreover, due to various parameters and optimization goals, validation of external, statistical, and construct validity in CAMs is challenging. In addition, quantifying the defence level accurately at the time of anonymization is also challenging owing to inadequate knowledge of an attacker's expertise.

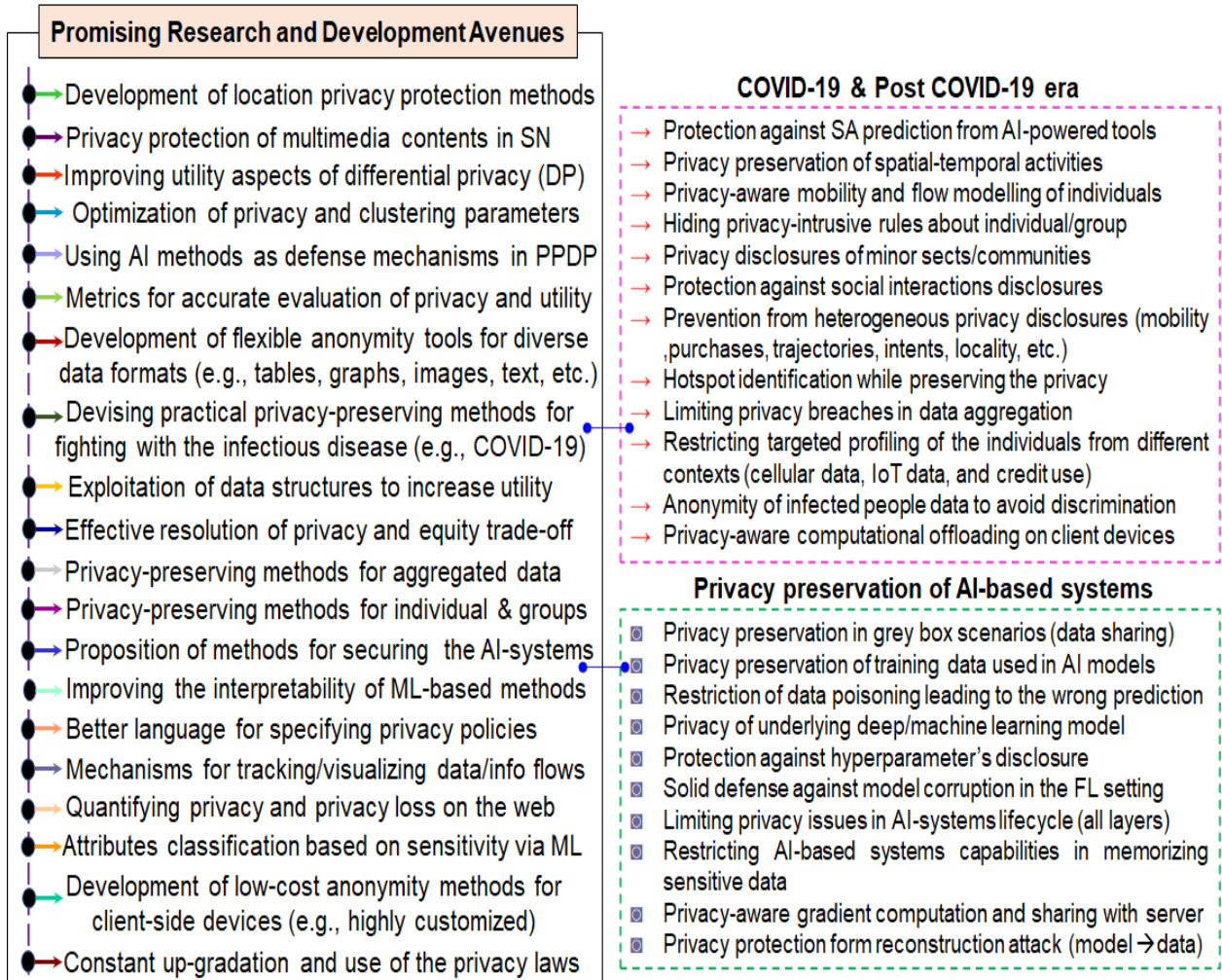


FIGURE 27. Comprehensive overview of promising opportunities for future research and developments in the privacy arena.

Apart from the technical challenges cited above, accurate quantification of privacy and utility levels offered by CAMs, development of low-cost evaluation metrics for CAMs, improving the interpretability of CAMs, resisting multiple AI-powered attacks, and addressing the privacy versus utility trade-off are all challenging tasks.

B. POTENTIAL OPPORTUNITIES FOR FUTURE RESEARCH IN PRIVACY DOMAIN

Owing to rapid digitization in recent years, especially during the COVID-19 pandemic, privacy protection has become one of the most trendy topics. Recently, many privacy protection techniques have been developed to secure personal data against manipulations in different digital infrastructures. Considering the latest research dynamics and emerging technologies, privacy protection will remain a concern [274]–[278]. Based on the thorough analysis of the published literature, the threats/challenges to information privacy in recent times, and considering the existing countermeasures,

we highlight in Figure 27 various potential avenues for future research.

With the advent of COVID-19, location data have been used as one of the potential tools for accomplishing multiple goals (e.g., contact tracing, surveillance, and quarantine monitoring) [279], [280]. Since many apps constantly track trajectories and location data, privacy issues of various kinds can arise over matters such as targeted profiling, spatial-temporal activities, web searches, interests, preferences, and web-search patterns. Furthermore, location data published by many location-based services can lead to privacy leaks due to the availability of huge amounts of auxiliary information about users. Recently, there has been an increasing focus on devising practical anonymization methods to restrict corporate surveillance and ensure responsible use of personal data. In this line of work, devising practical, verifiable, and efficient anonymization mechanisms is a vibrant avenue for future research.

Primarily, most research in the information privacy area has mainly focused on tabular and graph data. Moreover,

due to the increase in sources of data generation, privacy preservation mechanisms for images [281], videos [282], stream data [283], and temporal data [284] have become hot research topics. Despite many developments, this is still an emerging avenue of research. The DP model is regarded as one of the most promising solutions for privacy protection in static and dynamic scenarios. However, due to excessive noise added by the DP model during anonymization, the utility of the anonymized data can be significantly low [285]. Hence, devising new methods that can boost the utility of the DP model in most settings, especially in the healthcare sector, is a vital research direction.

Generally, most anonymization methods have certain parameters to consider (k, ϵ, t, ℓ , etc.), and each parameter has a distinct impact on privacy and utility [286]. Furthermore, these parameters do not yield consistent performance in diverse applications. Similarly, the synergy of anonymization approaches with clustering approaches brings another set of parameters. Hence, optimization of anonymization and clustering parameters by introducing adaptive learning strategies (or exploiting the inherent statistics of the data) is an important research direction. Recently, machine learning techniques have shown potential in securing personal data from adversaries [287]. Hence, employing ML to preserve privacy of data encompassed in diverse formats is a vibrant area of research. In the published literature, most privacy/utility evaluation metrics do not yield consistent performance, and fail to provide sufficient resilience against emerging privacy threats. Their performance differs from application to application, and they mainly capture only minor privacy attacks, or measure utility from fewer aspects. Recently, there has been an increasing focus on developing fine-grained evaluation metrics for PPDP [288]. Considering their necessity and significance, devising accurate evaluation metrics that can accurately measure the privacy and utility levels is an active area of research.

Since the emergence of COVID-19, privacy has become a main concern for most people around the globe due to the rapid proliferation of digital surveillance technologies. In these technologies, intimate details of people's lives are collected in order to control the effects from the pandemic. However, due to data transfer in cyberspace and the invasive use of personal data, privacy issues were reported from different regions. In the early days of the pandemic, due to privacy issues and interference in personal lives, some people even committed suicide in South Korea [289]. Furthermore, a lot of personal data (travel logs, mobility data, facility visits, generic personal information, etc.) have been transferred to cyberspace amid this pandemic. Hence, privacy issues will spark renewed interest in the near future. Considering the circumstances, finding practical privacy-preserving methods from the different perspectives shown in Figure 27 is a hot research area. Furthermore, devising solutions for synthetic data generation that can fulfil the data demands of researchers is also an emerging avenue of research [290].

Retaining sufficient utility in anonymized data without compromising privacy is a very hot research area because, in most cases, high-quality data are usually preferred for data mining tasks [291], [292]. Restricting extensive changes during data anonymization is imperative to yielding high-quality anonymized data, but this can only be possible by exploiting hidden characteristics of the underlying data to be anonymized. Considering the significance of high-utility datasets, anonymity methods that can restrict heavier changes in data conversion are required to improve the performance of knowledge-based systems/applications. Recently, it has been suggested that there exist various groups (major, minor, super minor) in data, leading to a new trade-off: privacy versus equity [246]. We demonstrate this trade-off in Figure 28, and an effective resolution of this trade-off is imperative in decision-making. Hence, there is an emerging need to develop privacy-preserving methods for this important research direction. Recently, due to pervasive technologies such as the IoT, SN, and fog/edge computing, a huge amount of distributed data (a.k.a. aggregated data) is available about individuals [293]. The data anonymized in one domain can be de-anonymized by linking them with another domain. To avoid these issues, finding practical anonymization methods that can provide resilience in aggregated data is a vibrant area of research. Most anonymity methods published so far have mainly focused on individual privacy preservation, which can still lead to group-privacy disclosures. With the advent of big data technologies, more practical methods that can simultaneously guarantee individual privacy, as well as group privacy, are needed in the near future.

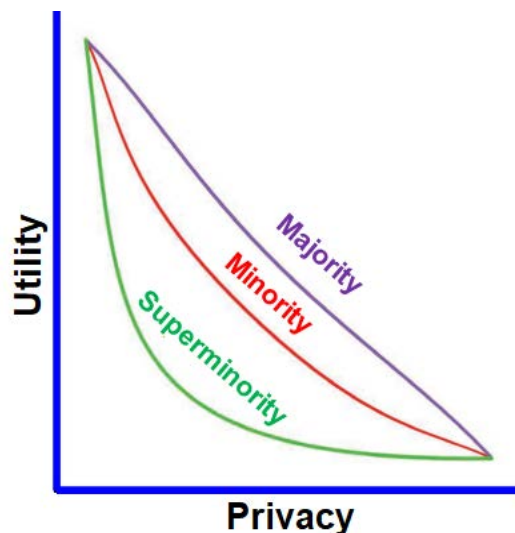


FIGURE 28. Overview of the privacy-versus-equity trade-off in PPDP.

In recent years, privacy preservation in AI-based systems has become one of the famous research areas that require robust mitigating strategies in order to lower potential privacy risks [294]. There is a pressing need to devise privacy-preserving solutions for all critical components of AI-based systems, such as data input (client-devices/sensors),

data pre-processing, ML models, and output [295]. With the advent of federated learning [296], privacy in AI has become a trendy topic, because FL requires privacy preservation from different perspectives. The conceptual overview of FL is shown in Figure 29. The privacy landscape of FL is relatively extensive, compared to centralized learning, due to its distributed nature [297]. The difference between central learning (CL) and FL is demonstrated in Eq. 3:

$$\text{Case}(CL||FL) = \begin{cases} \text{data} \rightarrow \text{algorithms}, & \text{CL} \\ \text{algorithms} \rightarrow \text{data}, & \text{FL} \end{cases} \quad (3)$$

As shown in Figure 29 and Equation 3, FL brings algorithm close to data that is why it is one of the famous privacy preserving paradigms in recent times.

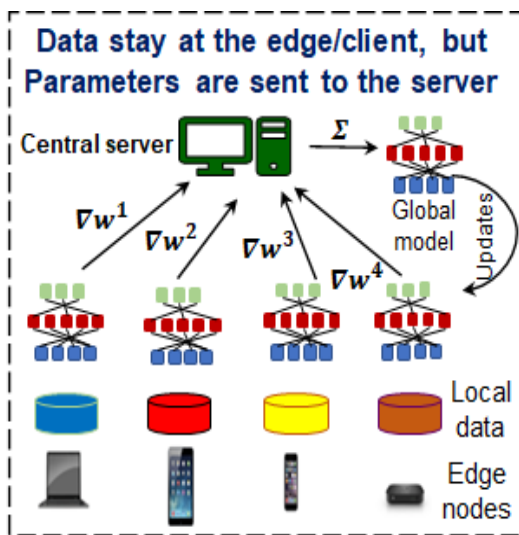


FIGURE 29. Technical overview of federated learning paradigm.

Recently, Ferrag *et al.* [298] comprehensively discussed various methods for mitigating cybersecurity issues in IoT environments using federated deep learning approaches. Through experimental analysis, the authors proved that federated deep learning approaches are superior compared to non-FL approaches in many ways (i.e., privacy preservation of IoT devices' data, and attacks detection accuracy). Treleven *et al.* [299] discussed the data ecosystem, and highlighted the relationship between FL and other data science technologies. The authors highlighted various engineering issues in the FL ecosystem. Bouacida *et al.* [300] discussed many vulnerabilities of the FL paradigm from user/participants, server, and aggregation protocol perspectives. The authors suggested a technology stack and valuable directions to mitigate those vulnerabilities. Benmalek *et al.* [301] provided a holistic view of the security concerns in the FL paradigm. The authors have discussed various attacks and vulnerabilities in FL and recently developed promising defense mechanisms against them. Li *et al.* [302] discussed the challenges and characteristics of the FL ecosystem from technical perspectives. The authors provided a broad survey about the technical problems of the

FL ecosystem, especially regarding privacy preservation. The authors pointed out that significant interdisciplinary efforts are needed in order to solve most technical problems of the FL paradigm. Shyu *et al.* [303] discussed data-related challenges concerning the FL paradigm in the healthcare industry, and suggested valuable directions to solve those challenges. In recent years, FL has been thoroughly investigated from applications as well as threats point of view. We refer interested readers to learn more about the FL ecosystem from recently published previous surveys [304]–[306].

In FL, multiple adversarial attacks can be launched, such as model inversion, data poisoning, model poisoning, and data re-construction. Furthermore, the privacy of participating clients and their associated personal data needs to be preserved in an effective way. A large number of studies have been published on defending against adversarial attacks on FL, however, there is still a lot of room to improve the privacy in such systems. Considering the need for privacy-preserving mechanisms in the FL context, provable privacy-preserving methods to safeguard FL systems from adversarial attacks, as shown in Figure 27, is an emerging avenue of research.

The last six directions listed in Figure 27 are related to development. To this end, devising privacy policies, visualizing and monitoring data flows in digital systems, quantifying privacy and privacy loss in web data are needed, as well as integrating ML-based methods for analyzing the sensitivity of data in multi-party computing environments, developing anonymization methods that can work in client devices, constant updating of security patches, and integration of legal measures with other robust methods, such as privacy by design (PbD) to secure personal data in third-party applications. In this line of work, answering data analysts' queries by ensuring sufficient privacy is an emerging avenue of research. Furthermore, developing low-cost solutions to generate synthetic data from real data in order to fulfill the demands of researchers is a main focus of research these days. In addition, there is a pressing need to develop privacy-preserving methods to ensure privacy for data originating from different computing environments, such as SN, sensors, actuators, and wearables.

With the evolution of FL and FL-based systems, federated analytics (FA) has emerged as a new collaborative analytics paradigm that solves the data-mining-related tasks without centralizing data from edge devices [307]. In line with the trends, it is imperative to develop prototypes and full-scale systems to realize FA on large and high-dimensional datasets. Furthermore, the integration of FA with systems that are used to fight the COVID-19 pandemic is a vibrant area of research. In recent years, there has been an increasing focus on the responsible use of personal data. In this line of work, some anonymization mechanisms have been recently developed for data dissemination [308]–[310]. However, this area still requires practical anonymity solutions to ensure confidentiality and transparency amid continuous data generation from different sectors. Lastly, improving the efficiency and efficacy of anonymization methods leveraging

soft computing techniques is also an emerging avenue for research [311]. Considering the ever-changing landscape of privacy threats, developing computationally efficient and robust anonymization techniques that encompass fewer parameters and steps to increase defenses against adversarial attacks without degrading data utility is a hot research area for the near future.

In recent years, many real-time applications have emerged to facilitate decision-making by utilizing data produced by IoT/wearable devices. Although these applications assist in robust decision-making, privacy issues can also occur due to personal data involved in such applications. Therefore, data protection regulations, as well as fair information principles (FIPs), are being developed/adopted across the globe for the privacy-preservation of personal data. We demonstrate emerging real-time computing applications in Figure 30. All these applications listed in Figure 30 mostly work with real-time data. Recently, Shen *et al.* [312] discussed a real-time pricing method for big data environments based on DP. The proposed method produces aggregated query answers with minimal noise to facilitate data owners and data buyers in a privacy-preserved way. Sanchez *et al.* [313] presented a cyber-security platform that restricts privacy issues in the healthcare ecosystem in an automated way. The developed platform helps in developing many real-time innovative applications in the healthcare sector. In this line of this work, Awotunde *et al.* [314] developed a real-time framework based on an IoT-based cloud system to monitor the patients' condition. The proposed framework works with real-time data obtained from IoT sensors and alerts medical staff to advise patients when their health conditions change in hospitals. Recently, searching for the desired data item (or querying) from encrypted data has been extensively investigated to preserve the privacy of underlying original data in dynamic setting [315]. This technology has been extensively used in real-time applications for data mining-related tasks without compromising users' privacy.

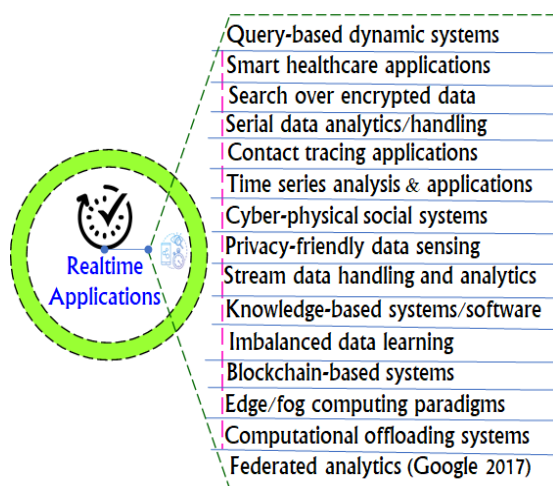


FIGURE 30. Overview of real-time computing (RTC) applications.

Due to the proliferation of IoT and cloud-based smart devices, medical jobs have been taken up by AI systems.

In this regard, a real-time system that utilizes IoT devices to identify/detect patients suffering from respiratory disease was recently proposed by Akram *et al.* [316]. Similarly, a real-time platform, named OnTimeEvidence was proposed by Alarcon *et al.* [317] to find multiple data sources related to healthcare in order to facilitate healthcare data consumers. In the ongoing COVID-19 pandemic, many real-time contact tracing applications have been developed that utilize IoT devices data in order to curb the spread of COVID-19 by identifying potentially suspected COVID-19 cases [318], [319]. In these applications, proximity and nature of contacts were analyzed in real-time to identify the contacts of confirmed cases as quickly as possible to lower the spread. In many smart city applications, real-time data was extensively used to make intelligent decisions, route suggestions, and product recommendations, to name a few. Zhang *et al.* [320] recently proposed a privacy-preserved real-time system for streaming traffic using the DP model. Tang *et al.* [321] proposed a real-time resource management scheme for cyber-physical-social systems (CPSSs). The proposed scheme maximizes the profit of the CPSS operator, and incentives users. Recently, due to an increase in avenues of data generation, many real-time applications that can gather and process stream data have emerged [322]–[325]. These stream-based applications can assist in performing basic data mining tasks (i.e., frequency analysis, alerting, monitoring, association rules, etc.) as well as advanced analytics such as video-based analytics, rules-based co-relations, event detection, pattern recognition, etc.

Knowledge-based systems can assist in solving complex problems using a knowledge base (i.e., a large and complex database). Chen *et al.* [326] proposed a real-time diagnosis method for surveillance videos based on deep learning combined with multiscale feature fusion in order to detect multiple types of anomalies. The proposed method can distinguish between normal and anomalous images with 98.52% accuracy. Recently, training/learning high-quality AI models from imbalanced data in real-time applications has become a popular research topic [327]. Vu *et al.* [328] developed a novel collaborative data model for semi-fully distributed settings for real-time medical applications. The proposed model employs the Naive Bayes classification to provide both privacy and accuracy in many real-life applications. In this line of work, researchers have explored the multi-class imbalanced problems in order to improve classifiers' performance from multiple perspectives [329]. Blockchain technology has revolutionized the privacy domain by removing the heavy reliance on a central server. Li *et al.* [330] discussed a provably secure method for privacy preservation in real-time IoT applications. Wen *et al.* [331] discussed a mobile medical, a system that ensures privacy and security of sensitive data while users can enjoy multiple medical services. Mobile medical makes use of identity authentication and blockchain technology in order to restrict information sharing with the server, and only minimal information is shared with the server.

Recently, many fog/edge-based applications have been developed to avoid any delay in real-time monitoring applications that collect and process real-time data. Sarrab et al. [331] discussed a fog computing method for preserving data privacy in IoT-based healthcare. The proposed healthcare internet of things (H-IoT)-based framework classifies data based on criticality, and only some data is moved to the cloud environments. The h-IoT framework can restrict privacy breaches and can avoid delays in time-critical real-time applications. Alzoubi et al. [332] suggested blockchain as a promising privacy-preserving mechanism for fog computing. Recently, fog and edge computing applications have been recognized as a promising tool for the prognosis and diagnosis of many critical diseases in the healthcare industry [333]. Due to the resource limitations and lack of technical knowledge, many companies outsource computations to external 3rd party servers. Computational offloading has become a very popular trend in recent times due to the huge proliferation of IoT-based applications across the globe. Xu et al. [334] discussed a promising solution for computational offloading in cloud-enabled IoT via federated deep reinforcement learning. The proposed method separates the high context-aware data from low context-aware data, and some parts of low context-aware data are sent to edge devices for processing. Wang Jin [335] discussed a computational offloading method for computation-intensive services without sacrificing guarantees on users' privacy. The proposed method's effectiveness was tested through various workflow parameters.

effects from COVID-19 based on demographics, for identification of COVID-19 risk factors, prediction of vulnerability indexes [339], and mortality/case predictions, to name a few. Although all real-time technologies cited above have helped societies/communities in multiple ways, their investigation regarding privacy protection and real-world deployment is yet to be made. From the extensive analysis of published literature, we suggest devising practical privacy-preserving solutions for real-time technologies that can ensure privacy preservation of personal data enclosed in diverse formats (i.e., logs, tables, graphs, streams, images, etc.) along with service requirements (i.e., scalability, low latency, transparency, trustworthiness, easy to use, availability, etc.).

According to the recent report of stonebranch¹ on IT automation state across the globe, 88 % of companies have a plan to invest in IT orchestration and automation in the year 2022. However, stonebranch identifies many challenges that hinder the adoption of IT solutions (e.g., cloud computing) through an in-depth survey. The respondents of this survey were IT professionals working in different IT-related enterprises. Among many other challenges, security and privacy concerns were regarded as one of the main barriers to IT adoption. In Table 8, we present a list of top reasons that are currently hindering the job's placement in public clouds. As shown in Table 8, 58% respondents think privacy and security as the main barrier when placing (or deciding to place) computing jobs in public clouds. Most of the existing privacy solutions are scenario-specific, and cannot ensure strong privacy and security of data in cloud environments. Considering the expected boom in IT orchestration and automation, robust solutions that can ensure strong privacy and security are required in the coming years. Apart from privacy protection at the data distribution stage, approaches are needed that can secure the complete lifecycle of data handling (i.e., collection, storage, pre-processing, analytics, distribution, use, and archival), especially in cloud computing environments.

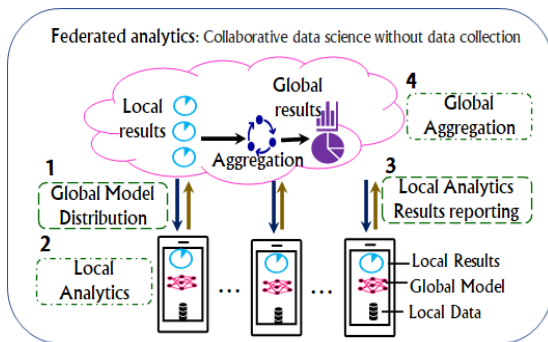


FIGURE 31. Workflow diagram of the federated analytics (FA).

In 2017, Google coined the term federated analytics (FA), performing analytics on local devices with data in way analogous to FL as shown in Figure 31. FA is another real-time technology resulting from FL and is based on the distributed computing paradigm [336]. It has been used in diverse fields such as medical, supply chain, finance, and energy sector for online analytics [337]. In the near future, FA will be one of the mainstream real-time technologies in the collaborative learning domain. In the context of the ongoing COVID-19 pandemic, FA has been widely used from multiple perspectives such as vaccine efficacy analysis for different subgroups, contact tracing [338], for analytics of the impact of COVID-19 and other diseases, in the categorization of

TABLE 8. Top reasons that are hindering jobs placement in public clouds.

Top reason (s)	% age of respondents
Lack of integration	62%
Security/Privacy concerns	58%
No experience with cloud	54%
Difficulty in transferring data to cloud	54%
Performance concerns	41%
Cost management	33%

From the extensive analysis of published literature and existing developments, we found that there are still a lot of opportunities to develop practical privacy-preserving mechanisms or tools that can ensure privacy preservation in static and dynamic scenarios [340]. Devising practical methods to preserve privacy while integrating multiple technologies is another attractive area of research [341]. Furthermore, we suggest devising technical privacy-preserving solutions

¹<https://www.stonebranch.com/>

for real-time technologies, heterogeneous types of personal data (i.e., logs, streams, graphs, time series data, videos, images, etc.), multiple computing paradigms (SN, IoT, CC, AI-based systems, IoT, etc.), and integrated technologies (i.e., FL and blockchain, IoT and cloud computing, anonymization and encryption, etc.). Most importantly, the need for privacy-preserved solutions has been greatly felt during the ongoing COVID-19 pandemic across the globe. Hence, proposing socio-technical and practical solutions to fight infectious diseases without sacrificing the guarantees of individual privacy is also a vibrant area of research in the coming years.

VIII. CONCLUSION

In this article, we described the findings of the latest SOTA research that proposed ways to overcome privacy issues in data sharing by leveraging clustering concepts. Recently, there has been an increasing focus on developing clustering-based anonymization mechanisms (CAMs) for responsible data science,² and this research area is gaining researchers' interest dramatically. CAMs have demonstrated their effectiveness in improving various technical aspects of traditional anonymization methods (e.g., k -anonymity, l -diversity, and t -closeness) regarding better privacy-preservation and utility enhancements in privacy-preserving data publishing (PPDP). Hence, it is of paramount importance to deliver good perspectives on information privacy involving heterogeneous data styles along with recent CAMs. In this work, we presented detailed and systematic coverage of CAMs used for securely publishing personal data enclosed in heterogeneous formats. Specifically, we mapped the existing CAMs to ten different data styles (tables, graphs, matrixes, logs, streams, traces, multimedia, text, documents, and hybrids), and we summarized and analyzed key features, including strengths, weaknesses, and clustering algorithms used in each study. Furthermore, we discussed the significance of CAMs in the emerging computing paradigms (e.g., social networks, cloud computing, location-based services, IoT-based applications/services, and AI-based services) that will assist in understanding research dynamics in these paradigms as well as in developing more practical anonymization solutions for them. In addition, we discussed the dark side of CAMs, exploited by malevolent adversaries to breach individual privacy by leveraging clustering algorithms and their respective data items. We discussed the substantial number of open challenges faced by the anonymization approaches that employ clustering concepts. Finally, we discussed various promising opportunities for future research considering the ever-changing landscape of privacy threats in recent times amid continuous technological developments. Based on the analysis of recent developments in CAMs, we examined that no single CAM could allay all types of privacy threats emanating from personal data handling

in digital environments. However, CAMs that have used low-cost clustering methods and that have shown better performance against major privacy threats on benchmark datasets are believed to be most effective for preserving privacy and utility in data analysis. Moreover, considering the recent research trends, the efficacy of these CAMs against AI-powered attacks in dynamic scenarios needs rigorous verification from both theoretical and experimental perspectives. The contents of this article can pave the way to improving existing CAMs as well as developing new CAMs to safeguard against emerging privacy threats in future endeavours.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

- [1] M. Al-Sartawi, *Big Data-Driven Digital Economy: Artificial and Computational Intelligence*. Cham, Switzerland: Springer, 2021.
- [2] J. Wieringa, P. K. Kannan, X. Ma, T. Reutterer, H. Risselada, and B. Skiera, "Data analytics in a privacy-concerned world," *J. Bus. Res.*, vol. 122, pp. 915–925, Jan. 2021.
- [3] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Boca Raton, FL, USA: CRC Press, 2010.
- [4] L. Sweeney, "Simple demographics often identify people uniquely," *Health, San Francisco*, vol. 671, no. 2000, pp. 1–34, 2000.
- [5] S. K. Kroes, M. P. Janssen, R. H. Groenwold, and M. van Leeuwen, "Evaluating privacy of individuals in medical data," *Health Informat. J.*, vol. 27, no. 2, 2021, Art. no. 1460458220983398.
- [6] F. Yağar, "Growing concern during the COVID-19 pandemic: Data privacy," *Turkiye Klinikleri J. Health Sci.*, vol. 6, no. 2, pp. 387–392, 2021.
- [7] L. Sweeney, " k -anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.
- [8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, " l -diversity: Privacy beyond k -anonymity," *ACM Trans. Knowl. Discovery from Data*, vol. 1, no. 1, p. 3–es, 2007.
- [9] N. Li, T. Li, and S. Venkatasubramanian, " t -closeness: Privacy beyond k -anonymity and l -diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.
- [10] J. Han, H. Yu, and J. Yu, "An improved l -diversity model for numerical sensitive attributes," in *Proc. 3rd Int. Conf. Commun. Netw. China*, 2008, pp. 938–943.
- [11] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 7, pp. 943–956, Jul. 2010.
- [12] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.* Berlin, Germany: Springer, 2008, pp. 1–19.
- [13] Y. Hao, H. Cao, C. Hu, K. Bhattarai, and S. Misra, " K -anonymity for social networks containing rich structural and textual information," *Social Netw. Anal. Mining*, vol. 4, no. 1, pp. 1–40, Dec. 2014.
- [14] X. Li, J. Yang, Z. Sun, and J. Zhang, "Differential privacy for edge weights in social networks," *Secur. Commun. Netw.*, vol. 2017, pp. 1–10, Mar. 2017.
- [15] J. Casas-Roma, J. Salas, F. D. Malliaros, and M. Vazirgiannis, " k -degree anonymity on directed networks," *Knowl. Inf. Syst.*, vol. 61, no. 3, pp. 1743–1768, Dec. 2019.
- [16] K. Rajendran, M. Jayabalan, and M. E. Rana, "A study on k -anonymity, l -diversity, and t -closeness techniques," *IJCSNS*, vol. 17, no. 12, p. 172, 2017.
- [17] H.-Y. Tran and J. Hu, "Privacy-preserving big data analytics a comprehensive survey," *J. Parallel Distrib. Comput.*, vol. 134, pp. 207–218, Dec. 2019.
- [18] M. Siddula, L. Li, and Y. Li, "An empirical study on the privacy preservation of online social networks," *IEEE Access*, vol. 6, pp. 19912–19922, 2018.

²<https://redasci.org/>

- [19] Y. Mengmeng, Z. Tianqing, Z. Wanlei, and X. Yang, "Attacks and countermeasures in social network data publishing," *ZTE Commun.*, vol. 14, pp. 2–9, Nov. 2019.
- [20] A. Sharma, G. Singh, and S. Rehman, "A review of big data challenges and preserving privacy in big data," in *Advances in Data and Information Sciences*. Singapore: Springer, 2020, pp. 57–65.
- [21] A. Majeed and S. Lee, "Anonymization techniques for privacy preserving data publishing: A comprehensive survey," *IEEE Access*, vol. 9, pp. 8512–8545, 2021.
- [22] M. Cunha, R. Mendes, and J. P. Vilela, "A survey of privacy-preserving mechanisms for heterogeneous data types," *Comput. Sci. Rev.*, vol. 41, Aug. 2021, Art. no. 100403.
- [23] R. Gangarde, A. Pawar, and A. Sharma, "Comparisons of different clustering algorithms for privacy of online social media network," in *Proc. IEEE Pune Sect. Int. Conf. (PuneCon)*, Dec. 2021, pp. 1–5.
- [24] M. Becker, "Privacy in the digital age: Comparing and contrasting individual versus social approaches towards privacy," *Ethics Inf. Technol.*, vol. 21, no. 4, pp. 307–317, Dec. 2019.
- [25] S. de Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Data privacy: Definitions and techniques," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 20, no. 6, pp. 793–817, 2012.
- [26] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2016.
- [27] J.-J. Yang, J.-Q. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment," *Future Gener. Comput. Syst.*, vols. 43–44, pp. 74–86, Feb. 2015.
- [28] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–36, Mar. 2022.
- [29] A. Majeed, F. Ullah, and S. Lee, "Vulnerability- and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data," *Sensors*, vol. 17, no. 5, p. 1059, May 2017.
- [30] J. Park and H. Lim, "Privacy-preserving federated learning using homomorphic encryption," *Appl. Sci.*, vol. 12, no. 2, p. 734, Jan. 2022.
- [31] A. Somasundaram and U. S. Reddy, "Data imbalance: Effects and solutions for classification of large and highly imbalanced data," in *Proc. 1st Int. Conf. Res. Eng., Comput. Technol.*, 2016, pp. 1–16.
- [32] S. Ni, M. Xie, and Q. Qian, "Clustering based K-anonymity algorithm for privacy preservation," *Int. J. Netw. Secur.*, vol. 19, no. 6, pp. 1062–1071, 2017.
- [33] S. Kumar and P. Kumar, "Upper approximation based privacy preserving in online social networks," *Expert Syst. Appl.*, vol. 88, pp. 276–289, Dec. 2017.
- [34] L.-E. Wang and X. Li, "A hybrid optimization approach for anonymizing transactional data," in *Proc. Int. Conf. Algorithms Archit. Parallel Process*. Cham, Switzerland: Springer, 2015, pp. 120–132.
- [35] A. Gkoulalas-Divanis and G. Loukides, "Utility-guided clustering-based transaction data anonymization," *Trans. Data Priv.*, vol. 5, no. 1, pp. 223–251, 2012.
- [36] A. Aleroud, F. Yang, S. C. Pallaprolu, Z. Chen, and G. Karabatis, "Anonymization of network traces data through condensation-based differential privacy," *Digit. Threats, Res. Pract.*, vol. 2, no. 4, pp. 1–23, Dec. 2021.
- [37] N. Mamede, J. Baptista, and F. Dias, "Automated anonymization of text documents," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2016, pp. 1287–1294.
- [38] D. Garat and D. Wonsever, "Automatic curation of court documents: Anonymizing personal data," *Information*, vol. 13, no. 1, p. 27, Jan. 2022.
- [39] Y. Saygin, D. Hakkini-Tur, and G. Tur, "Sanitization and anonymization of document repositories," in *Web and Information Security*. Pennsylvania, PA, USA: IGI Global, 2006, pp. 133–148.
- [40] C. Iwendi, S. A. Moqurrab, A. Anjum, S. Khan, S. Mohan, and G. Srivastava, "N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets," *Comput. Commun.*, vol. 161, pp. 160–171, Sep. 2020.
- [41] M. Hintze and K. E. Emam, "Comparing the benefits of pseudonymisation and anonymisation under the GDPR," *J. Data Protection Privacy*, vol. 2, no. 2, pp. 145–158, Aug. 2018.
- [42] S. Zhang, G. H. Yang, L. Singh, and L. Xiong, "Safelog: Supporting web search and mining by differentially-private query logs," in *Proc. AAAI Fall Symp. Ser.*, 2016, pp. 1–10.
- [43] S. Ghiasvand and F. M. Ciorba, "Anonymization of system logs for privacy and storage benefits," 2017, *arXiv:1706.04337*.
- [44] R. Patil, P. D. Patil, S. Kanase, N. Bhegade, V. Chavan, and S. Kasetwar, "System for analyzing crime news by mining live data streams with preserving data privacy," in *Sentimental Analysis and Deep Learning*. Singapore: Springer, 2022, pp. 799–811.
- [45] U. Sopaoglu and O. Abul, "Classification utility aware data stream anonymization," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107743.
- [46] A. R. S. Nasab and H. Ghaffarian, "A new fast framework for anonymizing IoT stream data," in *Proc. 5th Int. Conf. Internet Things Appl. (IoT)*, May 2021, pp. 1–5.
- [47] J. Kumar, "Slide window method adapted for privacy-preserving: Transactional data streams," *Eur. J. Mol. Clin. Med.*, vol. 8, no. 2, pp. 2528–2538, 2021.
- [48] L. Yang, X. Chen, Y. Luo, X. Lan, and W. Wang, "IDEA: A utility-enhanced approach to incomplete data stream anonymization," *Tsinghua Sci. Technol.*, vol. 27, no. 1, pp. 127–140, Feb. 2022.
- [49] F. Dufaux and T. Ebrahimi, "A framework for the validation of privacy protection solutions in video surveillance," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2010, pp. 66–71.
- [50] R. E. Thomas, S. K. Banu, and B. Tripathy, "Image anonymization using clustering with pixelization," *Int. J. Eng. Technol.*, vol. 7, pp. 990–993, Jan. 2018.
- [51] W. Zheng, Z. Wang, T. Lv, Y. Ma, and C. Jia, "K-anonymity algorithm based on improved clustering," in *Proc. Int. Conf. Algorithms Archit. Parallel Process*. Cham, Switzerland: Springer, 2018, pp. 462–476.
- [52] K. Mohammed, A. Ayesah, and E. Boiten, "Utility promises of self-organising maps in privacy preserving data mining," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*. Cham, Switzerland: Springer, 2020, pp. 55–72.
- [53] S. Khan, K. Iqbal, S. Faizullah, M. Fahad, J. Ali, and W. Ahmed, "Clustering based privacy preserving of big data using fuzzification and anonymization operation," 2020, *arXiv:2001.01491*.
- [54] P. Khan, Y. Khan, and S. Kumar, "Single identity clustering-based data anonymization in healthcare," in *Computationally Intelligent Systems and Their Applications*. Singapore: Springer, 2021, pp. 1–9.
- [55] S. Zouinina, N. Grozavu, Y. Bennani, A. Lyhyaoui, and N. Rogovschi, "Efficient k-anonymization through constrained collaborative clustering," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 405–411.
- [56] W. Zheng, Y. Ma, Z. Wang, C. Jia, and P. Li, "Effective L-diversity anonymization algorithm based on improved clustering," in *Proc. Int. Symp. Cyberspace Saf. Secur.* Cham, Switzerland: Springer, 2019, pp. 318–329.
- [57] F. Ashkouti, K. Khamforoosh, A. Sheikhhahmadi, and H. Khamfroush, "DHkmeans- ℓ diversity: Distributed hierarchical K-means for satisfaction of the ℓ -diversity privacy model using Apache Spark," *J. Supercomput.*, vol. 78, no. 2, pp. 2616–2650, Feb. 2022.
- [58] A. Abbasi and B. Mohammadi, "A clustering-based anonymization approach for privacy-preserving in the healthcare cloud," *Concurrency Comput., Pract. Exper.*, vol. 34, no. 1, p. e6487, Jan. 2022.
- [59] J. A. Onesimu, J. Karthikeyan, and Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services," *Peer-Peer Netw. Appl.*, vol. 14, no. 3, pp. 1629–1649, May 2021.
- [60] Y. Yan, E. A. Herman, A. Mahmood, T. Feng, and P. Xie, "A weighted K-member clustering algorithm for K-anonymization," *Computing*, vol. 103, pp. 2251–2273, Feb. 2021.
- [61] R. Gangarde, A. Sharma, A. Pawar, R. Joshi, and S. Gonge, "Privacy preservation in online social networks using multiple-graph-properties-based clustering to ensure k-anonymity, l-diversity, and t-closeness," *Electronics*, vol. 10, no. 22, p. 2877, Nov. 2021.
- [62] H. Zhang, L. Lin, L. Xu, and X. Wang, "Graph partition based privacy-preserving scheme in social networks," *J. Netw. Comput. Appl.*, vol. 195, Dec. 2021, Art. no. 103214.
- [63] S. Shakeel, A. Anjum, A. Asheralieva, and M. Alam, "k-NDDP: An efficient anonymization model for social network data release," *Electronics*, vol. 10, no. 19, p. 2440, Oct. 2021.
- [64] Z.-G. Chen, H.-S. Kang, S.-N. Yin, and S.-R. Kim, "An efficient privacy protection in mobility social network services with novel clustering-based anonymization," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, pp. 1–9, Dec. 2016.
- [65] D. Mohapatra and M. R. Patra, "Graph anonymization using hierarchical clustering," in *Computational Intelligence in Data Mining*. Singapore: Springer, 2019, pp. 145–154.

- [66] M. E. Skarkala, M. Maragoudakis, S. Gritzalis, L. Mitrou, H. Toivonen, and P. Moen, "Privacy preservation by k -anonymization of weighted social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 423–428.
- [67] R. K. Langari, S. Sardar, S. A. A. Mousavi, and R. Radfar, "Combined fuzzy clustering and firefly algorithm for privacy preserving in social networks," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112968.
- [68] R. Mortazavi and S. H. Erfani, "GRAM: An efficient (k, l) graph anonymization method," *Expert Syst. Appl.*, vol. 153, Sep. 2020, Art. no. 113454.
- [69] S. Kumar and P. Kumar, "Privacy preserving in online social networks using fuzzy rewiring," *IEEE Trans. Eng. Manag.*, early access, Apr. 30, 2021, doi: 10.1109/TEM.2021.3072812.
- [70] F. Heidari, A. Ghorbannia, and F. Rashidi, "SACK: Anonymization of social networks by clustering of k -edge-connected subgraphs," *Int. J. Comput. Appl.*, vol. 77, pp. 1–7, Sep. 2013.
- [71] L.-E. Wang, S. Lin, Y. Bai, S.-Y. Chang, X. Li, and P. Liu, "A privacy preserving method for publishing set-valued data and its correlative social network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–7.
- [72] S.-L. Wang, Y.-C. Tsai, H.-Y. Kao, and T.-P. Hong, "On anonymizing transactions with sensitive items," *Appl. Intell.*, vol. 41, no. 4, pp. 1043–1058, 2014.
- [73] A. Gkoulalas-Divanis and G. Loukides, "PCTA: Privacy-constrained clustering-based transaction data anonymization," in *Proc. 4th Int. Workshop Privacy Anonymity Inf. Soc. (PAIS)*, 2011, pp. 1–10.
- [74] N. Awad, J.-F. Couchot, B. A. Bouna, and L. Philippe, "Ant-driven clustering for utility-aware disassociation of set-valued datasets," in *Proc. 23rd Int. Database Appl. Eng. Symp. (IDEAS)*, 2019, pp. 1–9.
- [75] N. Awad, J.-F. Couchot, B. A. Bouna, and L. Philippe, "Publishing anonymized set-valued data via disassociation towards analysis," *Future Internet*, vol. 12, no. 4, p. 71, Apr. 2020.
- [76] S. Barakat, B. A. Bouna, M. Nassar, and C. Guyeux, "On the evaluation of the privacy breach in disassociated set-valued datasets," 2016, *arXiv:1611.08417*.
- [77] W. Pingshui, "Personalized anonymity algorithm using clustering techniques," *J. Comput. Inf. Syst.*, vol. 7, no. 3, pp. 924–931, 2011.
- [78] O. Can, "Personalised anonymity for microdata release," *IET Inf. Secur.*, vol. 12, no. 4, pp. 341–347, Jul. 2018.
- [79] M. Mohammady, L. Wang, Y. Hong, H. Louafi, M. Pourzandi, and M. Debbabi, "Preserving both privacy and utility in network trace anonymization," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 459–474.
- [80] L. Fan and A. Pokkunuru, "DPNeT: Differentially private network traffic synthesis with generative adversarial networks," in *Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy*. Cham, Switzerland: Springer, 2021, pp. 3–21.
- [81] M. Mohammady, M. Oqaily, L. Wang, Y. Hong, H. Louafi, M. Pourzandi, and M. Debbabi, "A multi-view approach to preserve privacy and utility in network trace anonymization," *ACM Trans. Privacy Secur.*, vol. 24, no. 3, pp. 1–36, Aug. 2021.
- [82] A. Aleroud, Z. Chen, and G. Karabatis, "Network trace anonymization using a prefix-preserving condensation-based technique (short paper)," in *Proc. OTM Federated Int. Conf. 'Move Meaningful Internet Syst.'*. Cham, Switzerland: Springer, 2016, pp. 934–942.
- [83] P. Velarde-Alvarado, C. Vargas-Rosales, R. Martinez-Pelaez, H. Toral-Cruz, and A. F. Martinez-Herrera, "An unsupervised approach for traffic trace sanitization based on the entropy spaces," *Telecommun. Syst.*, vol. 61, no. 3, pp. 609–626, 2016.
- [84] S. Shaham, M. Ding, B. Liu, S. Dang, Z. Lin, and J. Li, "Privacy preserving location data publishing: A machine learning approach," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 9, pp. 3270–3283, Sep. 2021.
- [85] N. Mahajan and V. Barkade, "Clustering based efficient privacy preserving multi keyword search over encrypted data," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBE)*, Aug. 2018, pp. 1–6.
- [86] X.-B. Li and J. Qin, "A framework for privacy-preserving medical document sharing," in *Proc. Int. Conf. Inf. Syst. (ICIS)*. Milano, Italy: Association for Information Systems, Dec. 2013.
- [87] X. Li and J. Qin, "Protecting privacy when releasing search results from medical document data," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, 2018, pp. 1–9.
- [88] W. Kong, M. Bendersky, M. Najork, B. Vargo, and M. Colagrosso, "Learning to cluster documents into workspaces using large scale activity logs," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2416–2424.
- [89] D. Garat and D. Wonsever, "Automatic curation of court documents: Anonymizing personal data," *Information*, vol. 13, no. 1, p. 27, Jan. 2022.
- [90] X.-B. Li and J. Qin, "Anonymizing and sharing medical text records," *Inf. Syst. Res.*, vol. 28, no. 2, pp. 332–352, Jun. 2017.
- [91] S. Lima-López, N. Perez, L. García-Sardiña, and M. Cuadros, "HitzalMed: Anonymisation of clinical text in Spanish," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 7038–7043.
- [92] J. Liu and K. Wang, "Anonymizing bag-valued sparse data by semantic similarity-based clustering," *Knowl. Inf. Syst.*, vol. 35, no. 2, pp. 435–461, May 2013.
- [93] W. Liu, M. Pei, C. Cheng, W. She, and C. Q. Wu, "An improved data anonymization algorithm for incomplete medical dataset publishing," in *Proc. Int. Conf. Healthcare Sci. Eng.* Singapore: Springer, 2018, pp. 115–128.
- [94] L. Ghemri and R. Kannah, "Lexical entailment for privacy protection in medical records," in *Proc. Int. Conf. Inform. Appl.*, Kuala Terengganu, Malaysia, pp. 228–232.
- [95] B. Chen, C. Wu, and S. Xie, "Generalization via hierarchical clustering for anonymizing set-valued user search logs," Center Brain-like Comput. Mach. Intell. (BCMI), Shanghai, China, Tech. Rep., 2013.
- [96] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "Utility-preserving privacy protection of nominal data sets via semantic rank swapping," *Inf. Fusion*, vol. 45, pp. 282–295, Jan. 2019.
- [97] N. Yuvaraj, K. Praghash, and T. Karthikeyan, "Data privacy preservation and trade-off balance between privacy and utility using deep adaptive clustering and elliptic curve digital signature algorithm," *Wireless Pers. Commun.*, vol. 124, pp. 655–670, May 2021.
- [98] X. Meng, Z. Xu, B. Chen, and Y. Zhang, "Privacy-preserving query log sharing based on prior n -word aggregation," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, Aug. 2016, pp. 722–729.
- [99] D. Pàmies-Estrens, J. Castellà-Roca, and J. Garcia-Alfaro, "A real-time query log protection method for web search engines," *IEEE Access*, vol. 8, pp. 87393–87413, 2020.
- [100] M. Ullah, "Obsecure logging: A framework to protect and evaluate the web search privacy," Ph.D. dissertation, Dept. Comput. Sci., Capital Univ., Telaiya, India, 2020.
- [101] V. Stephanie, M. A. P. Chamikara, I. Khalil, and M. Atiquzzaman, "Privacy-preserving location data stream clustering on mobile edge computing and cloud," *Inf. Syst.*, vol. 107, Jul. 2022, Art. no. 101728.
- [102] X. Yang, "Towards utility-aware privacy-preserving sensor data anonymization in distributed IoT," in *Proc. 8th ACM Int. Conf. Syst. Energy-Efficient Buildings, Cities, Transp.*, Nov. 2021, pp. 248–249.
- [103] J. Tekli, B. A. Bouna, Y. B. Issa, M. Kamradt, and R. Haraty, " (k, l) -clustering for transactional data streams anonymization," in *Proc. Int. Conf. Inf. Secur. Pract. Exper.* Cham, Switzerland: Springer, 2018, pp. 544–556.
- [104] K. Honda, M. Omori, S. Ubukata, and A. Notsu, "A study on fuzzy clustering-based k -anonymization for privacy preserving crowd movement analysis with face recognition," in *Proc. 7th Int. Conf. Soft Comput. Pattern Recognit. (SoCPaR)*, Nov. 2015, pp. 37–41.
- [105] X. Yang, H. Zhu, R. Lu, X. Liu, and H. Li, "Efficient and privacy-preserving online face recognition over encrypted outsourced data," in *Proc. IEEE Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jul. 2018, pp. 366–373.
- [106] Z. Zhang, T. Cilloni, C. Walter, and C. Fleming, "Multi-scale, class-generic, privacy-preserving video," *Electronics*, vol. 10, no. 10, p. 1172, May 2021.
- [107] M.-H. Le, M. S. N. Khan, G. Tsaloli, N. Carlsson, and S. Buchegger, "AnonFACES: Anonymizing faces adjusted to constraints on efficacy and security," in *Proc. 19th Workshop Privacy Electron. Soc.*, Nov. 2020, pp. 87–100.
- [108] A.-K. Grosselfinger, D. Münch, and M. Arens, "An architecture for automatic multimodal video data anonymization to ensure data protection," *Proc. SPIE*, vol. 11166, Oct. 2019, Art. no. 111660Q.
- [109] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 620–636.

- [110] P. Deivanai, J. J. V. Nayahi, and V. Kavitha, "A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data," in *Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT)*, Jun. 2011, pp. 732–736.
- [111] S. U. Bazai, J. Jang-Jaccard, and H. Alavizadeh, "A novel hybrid approach for multi-dimensional data anonymization for Apache Spark," *ACM Trans. Privacy Secur.*, vol. 25, no. 1, pp. 1–25, Feb. 2022.
- [112] D. Mohapatra and M. R. Patra, "Anonymization of attributed social graph using anatomy based clustering," *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 25455–25486, Sep. 2019.
- [113] S. B. Basapur and B. S. Shylaja, "Attribute assailability and sensitive attribute frequency based data generalization algorithm for privacy preservation," in *Proc. Int. Conf. Forensics, Anal., Big Data, Secur. (FABS)*, Dec. 2021, pp. 1–14.
- [114] K. Stokes, "Cover-up: A probabilistic privacy-preserving graph database model," *J. Ambient Intell. Humanized Comput.*, pp. 1–8, Oct. 2019, doi: 10.1007/s12652-019-01515-8.
- [115] P. Li, F. Zhou, Z. Xu, Y. Li, and J. Xu, "Privacy-preserving graph operations for social network analysis," in *Proc. Int. Symp. Secur. Privacy Social Netw. Big Data*, Singapore: Springer, 2020, pp. 303–317.
- [116] Y. Zhao and I. Wagner, "Using metrics suites to improve the measurement of privacy in graphs," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 1, pp. 259–274, Jan. 2022.
- [117] I. Wagner and I. Yevseyeva, "Designing strong privacy metrics suites using evolutionary optimization," *ACM Trans. Privacy Secur.*, vol. 24, no. 2, pp. 1–35, May 2021.
- [118] T. Ma, Y. Zhang, J. Cao, J. Shen, M. Tang, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, "KDVE: A k -degree anonymity with vertex and edge modification algorithm," *Computing*, vol. 97, no. 12, pp. 1165–1184, 2015.
- [119] H. Xia and W. Yang, "Information entropy models and privacy metrics methods for privacy protection," *Int. J. Netw. Secur.*, vol. 24, no. 1, pp. 1–10, 2022.
- [120] S. Ji, P. Mittal, and R. Beyah, "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 1305–1326, 2nd Quart., 2016.
- [121] W. Haoxiang and S. Smys, "Big data analysis and perturbation using data mining algorithm," *J. Soft Comput. Paradigm*, vol. 3, no. 1, pp. 19–28, Apr. 2021.
- [122] A. Cuzzocrea and E. Damiani, "Privacy-preserving big data exchange: Models, issues, future research directions," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 5081–5084.
- [123] S. Madan, K. Bhardwaj, and S. Gupta, "Critical analysis of big data privacy preservation techniques and challenges," in *Proc. Int. Conf. Innov. Comput. Commun.*, Singapore: Springer, 2022, pp. 267–278.
- [124] D. Xiang and W. Cai, "Privacy protection and secondary use of health data: Strategies and methods," *BioMed Res. Int.*, vol. 2021, pp. 1–11, Oct. 2021.
- [125] K. Guo and Q. Zhang, "Fast clustering-based anonymization approaches with time constraints for data streams," *Knowl.-Based Syst.*, vol. 46, pp. 95–108, Jul. 2013.
- [126] X. Qian, X. Li, and Z. Zhou, "An efficient privacy-preserving approach for data publishing," *J. Ambient Intell. Humanized Comput.*, pp. 1–17, Aug. 2021, doi: 10.1007/s12652-021-03417-0.
- [127] U. Sopaoglu and O. Abul, "Classification utility aware data stream anonymization," *Appl. Soft Comput.*, vol. 110, Oct. 2021, Art. no. 107743.
- [128] M. Milani, Y. Huang, and F. Chiang, "Data anonymization with diversity constraints," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 30, 2021, doi: 10.1109/TKDE.2021.3131528.
- [129] P. Parameshwarappa, "Clustering approaches for anonymizing high-dimensional sequential activity data," Ph.D. dissertation, Dept. Inf. Syst., Univ. Maryland, College Park, MA, USA, 2020.
- [130] U. Ahmed, G. Srivastava, and J. C.-W. Lin, "A machine learning model for data sanitization," *Comput. Netw.*, vol. 189, Apr. 2021, Art. no. 107914.
- [131] U. Ahmed, J. C.-W. Lin, P. Fournier-Viger, and C.-F. Cheng, "Privacy-preserving periodic frequent pattern model in AIoT applications," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Nov. 2021, pp. 1–2.
- [132] U. Ahmed, J. C.-W. Lin, and P. Fournier-Viger, "Privacy preservation of periodic frequent patterns using sensitive inverse frequency," in *Periodic Pattern Mining*, Singapore: Springer, 2021, pp. 215–227.
- [133] T. Henderson, "Short paper: Integrating the data protection impact assessment into the software development lifecycle," in *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, vol. 12484, Guildford, U.K.: Springer, Sep. 2020, pp. 219–228.
- [134] G. Logeswari, D. Sangeetha, and V. Vaidehi, "A cost effective clustering based anonymization approach for storing PHR's in cloud," in *Proc. Int. Conf. Recent Trends Inf. Technol.*, Apr. 2014, pp. 1–5.
- [135] J. U. Lawrance and J. V. N. Jesudhasan, "Privacy preserving parallel clustering based anonymization for big data using MapReduce framework," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 1–34, 2021.
- [136] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, and J. Chen, "Proximity-aware local-recoding anonymization with MapReduce for scalable big data privacy preservation in cloud," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2293–2307, Aug. 2015.
- [137] J. J. V. Nayahi and V. Kavitha, "Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop," *Future Gener. Comput. Syst.*, vol. 74, pp. 393–408, Sep. 2017.
- [138] N. Singh and A. K. Singh, "Data privacy protection mechanisms in cloud," *Data Sci. Eng.*, vol. 3, no. 1, pp. 24–39, Mar. 2018.
- [139] P. L. Lekshmy and M. A. Rahiman, "A sanitization approach for privacy preserving data mining on social distributed environment," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 7, pp. 2761–2777, Jul. 2020.
- [140] I. Jayaraman and A. S. Panneerselvam, "A novel privacy preserving digital forensic readiness provable data possession technique for health care data in cloud," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 5, pp. 4911–4924, May 2021.
- [141] S. Madan and P. Goswami, "Hybrid privacy preservation model for big data publishing on cloud," *Int. J. Adv. Intell. Paradigms*, vol. 20, nos. 3–4, pp. 343–355, 2021.
- [142] S. Madan and P. Goswami, "Adaptive privacy preservation approach for big data publishing in cloud using k -anonymization," *Recent Adv. Comput. Sci. Commun.*, vol. 14, no. 8, pp. 2678–2688, Oct. 2021.
- [143] E. Shanmugapriya and R. Kavitha, "Efficient and secure privacy analysis for medical big data using TDES and MKSVM with access control in cloud," *J. Med. Syst.*, vol. 43, no. 8, pp. 1–12, Aug. 2019.
- [144] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Inf. Syst.*, vol. 35, no. 8, pp. 884–910, Dec. 2010.
- [145] F. Fei, S. Li, H. Dai, C. Hu, W. Dou, and Q. Ni, "A k -anonymity based schema for location privacy preservation," *IEEE Trans. Sustain. Comput.*, vol. 4, no. 2, pp. 156–167, Jun. 2017.
- [146] J. Lee, S. Kim, Y. Cho, Y. Chung, and Y. Park, "A hierarchical clustering-based spatial cloaking algorithm for location-based services," *J. Internet Technol.*, vol. 13, no. 4, pp. 645–653, 2012.
- [147] K. Niu, C. Peng, Y. Tian, and W. Tan, "k-implicit tracking data publishing scheme against geo-matching attacks," *J. Inf. Sci. Eng.*, vol. 38, no. 1, pp. 1–16, 2022.
- [148] L. Yao, G. Wu, J. Wang, F. Xia, C. Lin, and G. Wang, "A clustering k -anonymity scheme for location privacy preservation," *IEICE Trans. Inf. Syst.*, vol. E95, no. 1, pp. 134–142, 2012.
- [149] C. Lin, G. Wu, and C. W. Yu, "Protecting location privacy and query privacy: A combined clustering approach," *Concurrency Comput., Pract. Exper.*, vol. 27, no. 12, pp. 3021–3043, Aug. 2015.
- [150] X. Zhang, G.-B. Kim, and H.-Y. Bae, "An adaptive spatial cloaking method for privacy protection in location-based service," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2014, pp. 480–485.
- [151] T. Altuwaiyan, X. Liang, and M. Hadian, "Towards efficient and privacy-preserving location-based comment sharing," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2016, pp. 1–6.
- [152] S. Mahdaviyar, M. Abadi, M. Kahani, and H. Mahdikhani, "A clustering-based approach for personalized privacy preserving publication of moving object trajectory data," in *Proc. Int. Conf. Netw. Syst. Secur.*, Berlin, Germany: Springer, 2012, pp. 149–165.
- [153] E. Dritsas, M. Trigka, P. Gerolymatos, and S. Sioutas, "Trajectory clustering and k -NN for robust privacy preserving spatiotemporal databases," *Algorithms*, vol. 11, no. 12, p. 207, Dec. 2018.
- [154] C. Chen, W. Lin, S. Zhang, Z. Ye, Q. Yu, and Y. Luo, "Personalized trajectory privacy-preserving method based on sensitive attribute generalization and location perturbation," *Intell. Data Anal.*, vol. 25, no. 5, pp. 1247–1271, Sep. 2021.
- [155] M. Ros-Martín, J. Salas, and J. Casas-Roma, "Scalable non-deterministic clustering-based k -anonymization for rich networks," *Int. J. Inf. Secur.*, vol. 18, no. 2, pp. 219–238, Apr. 2019.

- [156] Y. Gu and J. Lin, "Clustering-based dynamic privacy preserving method for social networks," *J. Commun.*, vol. 36, no. Z1, p. 126, 2015.
- [157] T. Truta, A. Campan, and A. Ralescu, "Preservation of structural properties in anonymized social networks," in *Proc. 8th IEEE Int. Conf. Collaborative Comput., Netw., Appl. Worksharing*, 2012, pp. 619–627.
- [158] A. Campan, Y. Alufaisan, T. M. Truta, and T. Richardson, "Preserving communities in anonymized social networks," *Trans. Data Privacy*, vol. 8, no. 1, pp. 55–87, 2015.
- [159] X. Chen, Z. Jiang, H. Li, J. Ma, and P. S. Yu, "Community hiding by link perturbation in social networks," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 3, pp. 704–715, Jun. 2021.
- [160] E. Ghosh, S. Kamara, and R. Tamassia, "Efficient graph encryption scheme for shortest path queries," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, May 2021, pp. 516–525.
- [161] L. Yu, T. Yang, Z. Wu, J. Zhu, J. Hu, and Z. Chen, "Sensitive edges protection in social networks," in *Proc. Int. Conf. Web-Age Inf. Manage.* Berlin, Germany: Springer, 2013, pp. 564–569.
- [162] P. Gazalian and S. M. Safi, "Presentation of a method for privacy preserving of people in social networks according to the clustering and SFLA," *Int. J. Comput. Sci. Netw. Solutions*, vol. 6, no. 1, pp. 1–9, 2018.
- [163] H. Zhang, X. Li, J. Xu, and L. Xu, "Graph matching based privacy-preserving scheme in social networks," in *Proc. Int. Symp. Secur. Privacy Social Netw. Big Data*. Singapore: Springer, 2021, pp. 110–118.
- [164] X. Liu and X. Yang, "A generalization based approach for anonymizing weighted social network graphs," in *Proc. Int. Conf. Web-Age Inf. Manage.* Berlin, Germany: Springer, 2011, pp. 118–130.
- [165] M. Siddula, Y. Li, X. Cheng, Z. Tian, and Z. Cai, "Anonymization in online social networks based on enhanced equi-cardinal clustering," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 4, pp. 809–820, Aug. 2019.
- [166] A. M. V. V. Sai, K. Zhang, and Y. Li, "User motivation based privacy preservation in location based social networks," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI)*, Oct. 2021, pp. 471–478.
- [167] T. Gao and F. Li, "Differential private social network publication and persistent homology preservation," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 4, pp. 3152–3166, Oct. 2021.
- [168] M. Yuan, L. Chen, P. S. Yu, and T. Yu, "Protecting sensitive labels in social network data anonymization," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 633–647, Mar. 2013.
- [169] F. Zhao, Y. Huang, A. M. V. V. Sai, and Y. Wu, "A cluster-based solution to achieve fairness in federated learning," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Dec. 2020, pp. 875–882.
- [170] Y. Liu, J. J. Q. Yu, J. Kang, D. Niyato, and S. Zhang, "Privacy-preserving traffic flow prediction: A federated learning approach," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7751–7763, Aug. 2020.
- [171] M. Badra and S. Zeadally, "Design and performance analysis of a virtual ring architecture for smart grid privacy," *IEEE Trans. Parallel Distrib. Syst.*, vol. 9, no. 2, pp. 321–329, Feb. 2014.
- [172] Y. Wang, M. Jia, N. Gao, L. Von Krannichfeldt, M. Sun, and G. Hug, "Federated clustering for electricity consumption pattern extraction," *IEEE Trans. Smart Grid*, vol. 13, no. 3, pp. 2425–2439, May 2022.
- [173] M. Ghahramani, A. O'Hagan, M. Zhou, and J. Sweeney, "Intelligent geodemographic clustering based on neural network and particle swarm optimization," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, May 3, 2021, doi: [10.1109/TSMC.2021.3072357](https://doi.org/10.1109/TSMC.2021.3072357).
- [174] K. Mohammed, A. Ayes, and E. Boiten, "Complementing privacy and utility trade-off with self-organising maps," *Cryptography*, vol. 5, no. 3, p. 20, Aug. 2021.
- [175] H. H. Kumar, V. R. Karthik, and M. K. Nair, "Federated k-means clustering: A novel edge AI based approach for privacy preservation," in *Proc. IEEE Int. Conf. Cloud Comput. Emerg. Markets (CEEM)*, Nov. 2020, pp. 52–56.
- [176] M. Stallmann and A. Wilbik, "Towards federated clustering: A federated fuzzy c-means algorithm (FFCM)," 2022, *arXiv:2201.07316*.
- [177] N. Rajesh and A. A. L. Selvakumar, "Association rules and deep learning for cryptographic algorithm in privacy preserving data mining," *Cluster Comput.*, vol. 22, no. S1, pp. 119–131, Jan. 2019.
- [178] S. Virupaksha and V. Dondeti, "Subspace based noise addition for privacy preserved data mining on high dimensional continuous data," *J. Ambient Intell. Humanized Comput.*, pp. 1–17, Mar. 2020, doi: [10.1007/s12652-020-01881-8](https://doi.org/10.1007/s12652-020-01881-8).
- [179] S. Bollaa, "An efficient probabilistic multi labeled big data clustering model for privacy preservation using linked weight optimization model," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 11, pp. 5510–5517, 2021.
- [180] L. U. Khan, M. Alsenwi, Z. Han, and C. S. Hong, "Self organizing federated learning over wireless networks: A socially aware clustering approach," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Jan. 2020, pp. 453–458.
- [181] S. Virupaksha and D. Venkatesulu, "Subspace-based aggregation for enhancing utility, information measures, and cluster identification in privacy preserved data mining on high-dimensional continuous data," *Int. J. Comput. Appl.*, pp. 1–10, Nov. 2019, doi: [10.1080/1206212X.2019.1686211](https://doi.org/10.1080/1206212X.2019.1686211).
- [182] X. Guo, H. Lin, Y. Wu, and M. Peng, "A new data clustering strategy for enhancing mutual privacy in healthcare IoT systems," *Future Gener. Comput. Syst.*, vol. 113, pp. 407–417, Dec. 2020.
- [183] J. Zhu, L. Huo, M. D. Ansari, and M. A. Ikbali, "Research on data security detection algorithm in IoT based on k-means," *Scalable Comput., Pract. Exper.*, vol. 22, no. 2, pp. 149–159, Oct. 2021.
- [184] N. Almusallam, A. Alabdulatif, and F. Alarfaj, "Analysis of privacy-preserving edge computing and Internet of Things models in healthcare domain," *Comput. Math. Methods Med.*, vol. 2021, pp. 1–6, Dec. 2021.
- [185] D. Huang, X. Yao, S. An, and S. Ren, "Private distributed K-means clustering on interval data," in *Proc. IEEE Int. Perform., Comput., Commun. Conf. (IPCCC)*, Oct. 2021, pp. 1–9.
- [186] M. Elhoseny, K. Haseeb, A. A. Shah, I. Ahmad, Z. Jan, and M. I. Alghamdi, "IoT solution for AI-enabled privacy-preserving with big data transferring: An application for healthcare using blockchain," *Energies*, vol. 14, no. 17, p. 5364, Aug. 2021.
- [187] J. S. Kumar and M. A. Zaveri, "Clustering for collaborative processing in IoT network," in *Proc. 2nd Int. Conf. IoT Urban Space*, May 2016, pp. 95–97.
- [188] J. Shuja, M. A. Humayun, W. Alasmay, H. Sinky, E. Alanazi, and M. K. Khan, "Resource efficient geo-textual hierarchical clustering framework for persistent IoT applications," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25114–25122, Nov. 2021.
- [189] A. Otgonbayar, Z. Pervez, and K. Dahal, "Toward anonymizing IoT data streams via partitioning," in *Proc. IEEE 13th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, Oct. 2016, pp. 331–336.
- [190] S. Patil, S. Joshi, and D. Patil, "Enhanced privacy preservation using anonymization in IoT-enabled smart homes," in *Smart Intelligent Computing and Applications*. Singapore: Springer, 2020, pp. 439–454.
- [191] F. Ullah, I. Ullah, A. Khan, M. I. Uddin, H. Alyami, and W. Alosaimi, "Enabling clustering for privacy-aware data dissemination based on medical healthcare-IoTs (MH-IoTs) for wireless body area network," *J. Healthcare Eng.*, vol. 2020, pp. 1–10, Nov. 2020.
- [192] Y. Li, X. Tao, X. Zhang, M. Wang, and S. Wang, "Break the data barriers while keeping privacy: A graph differential privacy method," *IEEE Internet Things J.*, early access, Feb. 15, 2022, doi: [10.1109/JIOT.2022.3151348](https://doi.org/10.1109/JIOT.2022.3151348).
- [193] J. Liu, C. Zhang, K. Xue, and Y. Fang, "Privacy preservation in multi-cloud secure data fusion for infectious-disease analysis," *IEEE Trans. Mobile Comput.*, early access, Jan. 25, 2022, doi: [10.1109/TMC.2022.3145745](https://doi.org/10.1109/TMC.2022.3145745).
- [194] C. Zhang, H. Jiang, Y. Wang, Q. Hu, J. Yu, and X. Cheng, "User identity de-anonymization based on attributes," in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl.* Cham, Switzerland: Springer, 2019, pp. 458–469.
- [195] C. Zhang, S. Wu, H. Jiang, Y. Wang, J. Yu, and X. Cheng, "Attribute-enhanced de-anonymization of online social networks," in *Proc. Int. Conf. Comput. Data Social Netw.* Springer, 2019, pp. 256–267.
- [196] H. Jiang, J. Yu, X. Cheng, C. Zhang, B. Gong, and H. Yu, "Structure-attribute-based social network de-anonymization with spectral graph partitioning," *IEEE Trans. Computat. Social Syst.*, early access, May 31, 2021, doi: [10.1109/TCSS.2021.3082901](https://doi.org/10.1109/TCSS.2021.3082901).
- [197] Y. Shao, J. Liu, S. Shi, Y. Zhang, and B. Cui, "Fast de-anonymization of social networks with structural information," *Data Sci. Eng.*, vol. 4, no. 1, pp. 76–92, Mar. 2019.
- [198] S. Gams, M.-O. Killijian, and M. N. D. P. Cortez, "De-anonymization attack on geolocated data," *J. Comput. Syst. Sci.*, vol. 80, no. 8, pp. 1597–1614, Dec. 2014.

- [199] C.-F. Chiasserini, M. Garetto, and E. Leonardi, "Impact of clustering on the performance of network de-anonymization," in *Proc. ACM Conf. Online Social Netw.*, Nov. 2015, pp. 83–94.
- [200] C.-F. Chiasserini, M. Garetto, and E. Leonardi, "Social network de-anonymization under scale-free user relations," *IEEE/ACM Trans. Netw.*, vol. 24, no. 6, pp. 3756–3769, Dec. 2016.
- [201] C.-F. Chiasserini, M. Garetto, and E. Leonardi, "De-anonymizing clustered social networks by percolation graph matching," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 2, pp. 1–39, Mar. 2018.
- [202] L. Fu, X. Fu, Z. Hu, Z. Xu, and X. Wang, "De-anonymization of social networks with communities: When quantifications meet algorithms," 2017, *arXiv:1703.09028*.
- [203] L. Fu, J. Zhang, S. Wang, X. Wu, X. Wang, and G. Chen, "De-anonymizing social networks with overlapping community structure," *IEEE/ACM Trans. Netw.*, vol. 28, no. 1, pp. 360–375, Feb. 2020.
- [204] M. Francia, E. Gallinucci, M. Golfarelli, and N. Santolini, "DART: De-anonymization of personal gazetteers through social trajectories," *J. Inf. Secur. Appl.*, vol. 55, Dec. 2020, Art. no. 102634.
- [205] T. Orekondy, S. J. Oh, Y. Zhang, B. Schiele, and M. Fritz, "Gradient-leaks: Understanding and controlling deanonymization in federated learning," 2018, *arXiv:1805.05838*.
- [206] Z. Chen, Y. Fu, M. Zhang, Z. Zhang, and H. Li, "The de-anonymization method based on user spatio-temporal mobility trace," in *Proc. Int. Conf. Inf. Commun. Secur. Cham, Switzerland: Springer*, 2017, pp. 459–471.
- [207] T. Murakami, A. Kanemura, and H. Hino, "Group sparsity tensor factorization for de-anonymization of mobility traces," in *Proc. IEEE Trustcom/BigDataSE/ISPA*, vol. 1, Aug. 2015, pp. 621–629.
- [208] H. Li, Q. Chen, H. Zhu, D. Ma, H. Wen, and X. Shen, "Privacy leakage via de-anonymization and aggregation in heterogeneous social networks," *IEEE Trans. Depend. Sec. Comput.*, vol. 17, no. 2, pp. 350–362, Mar. 2020.
- [209] C. Zhen-Yu, Z. Min, F. Yan-Yan, Z. Zhen-Feng, and L. Hao, "A user de-anonymization attack method for trajectory data publishing," *J. Inf. Secur. Reserach*, vol. 3, no. 10, pp. 1–11, 2017.
- [210] H. Wang, C. Gao, Y. Li, Z.-L. Zhang, and D. Jin, "Revealing physical world privacy leakage by cyberspace cookie logs," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 4, pp. 2550–2566, Dec. 2020.
- [211] Z. Zhang, Q. Gu, T. Yue, and S. Su, "Identifying the same person across two similar social networks in a unified way: Globally and locally," *Inf. Sci.*, vol. 394, pp. 53–67, Jul. 2017.
- [212] J. Chen, X. Lin, Z. Shi, and Y. Liu, "Link prediction adversarial attack via iterative gradient attack," *IEEE Trans. Computat. Social Syst.*, vol. 7, no. 4, pp. 1081–1094, Aug. 2020.
- [213] C. Ma, C. Chang, T. Ma, J. Huang, and Z. Niu, "User identity matching for multisource location data," in *Proc. IEEE 21st Int. Conf. Commun. Technol. (ICCT)*, Oct. 2021, pp. 686–694.
- [214] S. Nilizadeh, A. Kapadia, and Y.-Y. Ahn, "Community-enhanced de-anonymization of online social networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 537–548.
- [215] N. Takbiri, M. Chen, D. L. Goeckel, A. Houmansadr, and H. Pishro-Nik, "Asymptotic privacy loss due to time series matching of dependent users," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1079–1083, Apr. 2021.
- [216] F. Shirani, S. Garg, and E. Erkip, "An information theoretic framework for active de-anonymization in social networks based on group memberships," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 470–477.
- [217] J. Aliakbari, M. Delavar, J. Mohajeri, and M. Salmisazadeh, "A technique to improve de-anonymization attacks on graph data," in *Proc. Electr. Eng. (ICEE), Iranian Conf.*, May 2018, pp. 704–709.
- [218] X. Xueshuo, W. Jiming, Y. Junyi, F. Yaozheng, L. Ye, L. Tao, and W. Guiling, "AWAP: Adaptive weighted attribute propagation enhanced community detection model for bitcoin de-anonymization," *Appl. Soft Comput.*, vol. 109, Sep. 2021, Art. no. 107507.
- [219] H. Li, Q. Chen, H. Zhu, and D. Ma, "Hybrid de-anonymization across real-world heterogeneous social networks," in *Proc. ACM Turing 50th Celebration Conf.-China*, 2017, pp. 1–7.
- [220] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.
- [221] Z. Tu, F. Xu, Y. Li, P. Zhang, and D. Jin, "A new privacy breach: User trajectory recovery from aggregated mobility data," *IEEE/ACM Trans. Netw.*, vol. 26, no. 3, pp. 1446–1459, Jun. 2018.
- [222] Z. Yang, R. Wang, D. Luo, and Y. Xiong, "Rapid re-identification risk assessment for anonymous data set in mobile multimedia scene," *IEEE Access*, vol. 8, pp. 41557–41565, 2020.
- [223] M. Miculan, G. L. Foresti, and C. Picciarelli, "Towards user recognition by shallow web traffic inspection," in *Proc. ITASEC*, 2019, pp. 1–11.
- [224] A. De Nardin, M. Miculan, C. Picciarelli, and G. L. Foresti, "A time-series classification approach to shallow web traffic de-anonymization," in *Proc. Italian Conf. Cyber Secur.—ITASEC—CEUR-WS* Genoa, Italy: All Digital Event, Apr. 2021.
- [225] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 358–372, Feb. 2016.
- [226] W. Tian, J. Mao, J. Jiang, Z. He, Z. Zhou, and J. Liu, "Deeply understanding structure-based social network de-anonymization," *Proc. Comput. Sci.*, vol. 129, pp. 52–58, Jan. 2018.
- [227] T. Iwata, J. R. Lloyd, and Z. Ghahramani, "Unsupervised many-to-many object matching for relational data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 607–617, Mar. 2016.
- [228] A. Cecaj, M. Mamei, and N. Biccocchi, "Re-identification of anonymized CDR datasets using social network data," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PERCOM Workshops)*, Mar. 2014, pp. 237–242.
- [229] A. Cecaj, M. Mamei, and F. Zambonelli, "Re-identification and information fusion between anonymized CDR and social network data," *J. Ambient Intell. Humanized Comput.*, vol. 7, no. 1, pp. 83–96, Feb. 2016.
- [230] R. S. Hirschprung and O. Leshman, "Privacy disclosure by de-anonymization using music preferences and selections," *Telematics Inform.*, vol. 59, Jun. 2021, Art. no. 101564.
- [231] S. Linoy, N. Stakhanova, and S. Ray, "De-anonymizing Ethereum blockchain smart contracts through code attribution," *Int. J. Netw. Manage.*, vol. 31, no. 1, p. e2130, Jan. 2021.
- [232] K. Sharad and G. Danezis, "De-anonymizing D4D datasets," in *Proc. Workshop Hot Topics Privacy Enhancing Technol.*, Bloomington, IN, USA, 2013, p. 10.
- [233] A. Acquisti, R. Gross, and F. Stutzman, "Face recognition and privacy in the age of augmented reality," *J. Privacy Confidentiality*, vol. 6, no. 2, pp. 1–20, Dec. 2014.
- [234] H. H. Huang, J. W. Lin, and C. H. Lin, "Data re-identification—A case of retrieving masked data from electronic toll collection," *Symmetry*, vol. 11, no. 4, p. 550, Apr. 2019.
- [235] Q. Chen, Z. Wang, M. Zhang, H. Zhu, and S. Feng, "A de-anonymization attack for social network graph based on structural and node feature similarity," *DEStech Trans. Comput. Sci. Eng.*, Feb. 2018.
- [236] H. Ong and J. Shao, "De-anonymising set-generalised transactions based on semantic relationships," in *Proc. Int. Conf. Future Data Secur. Eng. Cham, Switzerland: Springer*, 2014, pp. 107–121.
- [237] M. Lin, H. Cao, V. Zheng, K. C. Chang, and S. Krishnaswamy, "Mobile user verification/identification using statistical mobility profile," in *Proc. Int. Conf. Big Data Smart Comput. (BIGCOMP)*, Feb. 2015, pp. 15–18.
- [238] G. M. de Castro Silva and J. S. Sichman, "Using social group trajectories for potential impersonation detection on smart buildings access control," in *Proc. 8th Brazilian Conf. Intell. Syst. (BRACIS)*, Oct. 2019, pp. 389–394.
- [239] A. Majeed and S. Lee, "Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data," *Appl. Intell.*, vol. 50, no. 8, pp. 2555–2574, Aug. 2020.
- [240] P. Mavriki and M. Karyda, "Big data analytics in healthcare applications: Privacy implications for individuals and groups and mitigation strategies," in *Proc. Eur., Medit., Middle Eastern Conf. Inf. Syst. Cham, Switzerland: Springer*, 2020, pp. 526–540.
- [241] Y. Hu, Y. Zhang, D. Gong, and X. Sun, "Multi-participant federated feature selection algorithm with particle swarm optimization for imbalanced data under privacy protection," *IEEE Trans. Artif. Intell.*, early access, Jan. 25, 2022, doi: 10.1109/TAI.2022.3145333.
- [242] A. Majeed and S. O. Hwang, "A practical anonymization approach for imbalanced datasets," *IT Prof.*, vol. 24, no. 1, pp. 63–69, Jan. 2022.
- [243] E. Erdemir, P. L. Dragotti, and D. Gunduz, "Active privacy-utility trade-off against a hypothesis testing adversary," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2660–2664.
- [244] D. Deutch, A. Frankenthal, A. Gilad, and Y. Moskovitch, "On optimizing the trade-off between privacy and utility in data provenance," in *Proc. Int. Conf. Manage. Data*, Jun. 2021, pp. 379–391.

- [245] N. Yuvaraj, R. A. Raja, and N. Kousik, "Privacy preservation between privacy and utility using ECC-based PSO algorithm," in *Intelligent Computing and Applications*. Singapore: Springer, 2021, pp. 567–573.
- [246] D. Pujol and A. Machanavajjhala, "Equity and privacy: More than just a tradeoff," *IEEE Secur. Privacy*, vol. 19, no. 6, pp. 93–97, Nov. 2021.
- [247] M. E. Kabir, H. Wang, and E. Bertino, "Efficient systematic clustering method for k -anonymization," *Acta Inf.*, vol. 48, no. 1, pp. 51–66, Feb. 2011.
- [248] P. Parameshwarappa, Z. Chen, and G. Koru, "An effective and computationally efficient approach for anonymizing large-scale physical activity data: Multi-level clustering-based anonymization," *Int. J. Inf. Secur. Privacy*, vol. 14, no. 3, pp. 72–94, 2020.
- [249] P. Parameshwarappa, Z. Chen, and G. Koru, "Anonymization of daily activity data by using ℓ -diversity privacy model," *ACM Trans. Manage. Inf. Syst.*, vol. 12, no. 3, pp. 1–21, May 2021.
- [250] S. Dilmaghani, M. R. Brust, G. Danoy, N. Cassagnes, J. Pecero, and P. Bouvry, "Privacy and security of big data in AI systems: A research and standards perspective," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 5737–5743.
- [251] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When machine learning meets privacy in 6G: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2694–2724, 4th Quart., 2020.
- [252] R. C. Wong, A. W. Fu, K. Wang, P. S. Yu, and J. Pei, "Can the utility of anonymized data be used for privacy breaches?" *ACM Trans. Knowl. Discovery Data (TKDD)*, vol. 5, no. 3, pp. 1–24, Aug. 2011.
- [253] J. Pastor-Galindo, F. G. Marmol, and G. M. Perez, "Nothing to hide? On the security and privacy threats beyond open data," *IEEE Internet Comput.*, vol. 25, no. 4, pp. 58–66, Jul. 2021.
- [254] J. Lahann, M. Scheid, and P. Fettke, "Utilizing machine learning techniques to reveal VAT compliance violations in accounting data," in *Proc. IEEE 21st Conf. Bus. Informat. (CBI)*, vol. 1, Jul. 2019, pp. 1–10.
- [255] Q. Gong, M. Yang, Z. Chen, W. Wu, and J. Luo, "A framework for utility enhanced incomplete microdata anonymization," *Cluster Comput.*, vol. 20, no. 2, pp. 1749–1764, Jun. 2017.
- [256] K. V. Ramana, V. V. Kumari, and K. Raju, "Impact of outliers on anonymized categorical data," in *Proc. Int. Conf. Digit. Image Process. Inf. Technol.* Berlin, Germany: Springer, 2011, pp. 326–335.
- [257] Y. Canbay, Y. Vural, and C. S. Sağıroğlu, "OAN: Outlier record-oriented utility-based privacy preserving model," *J. Fac. Eng. Archit. Gazi Univ.*, vol. 35, no. 1, pp. 1–14, 2020.
- [258] İ. Civelek and M. A. Aydın, "Mondrian based real time anonymization model," *Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi*, vol. 8, no. 1, pp. 472–483, May 2021.
- [259] J. J. V. Nayahi and V. Kavitha, "An efficient clustering for anonymizing data and protecting sensitive labels," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 23, no. 5, pp. 685–714, Oct. 2015.
- [260] J. Yuan, Y. Ou, and G. Gu, "An improved privacy protection method based on k -degree anonymity in social network," in *Proc. IEEE Int. Conf. Artif. Intell. Comput. Appl. (ICAICA)*, Mar. 2019, pp. 416–420.
- [261] D. Ruta, L. Cen, and E. Damiani, "Fast summarization and anonymization of multivariate big time series," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 1901–1904.
- [262] M. A. Mohamed, M. H. Nagi, and S. M. Ghanem, "A clustering approach for anonymizing distributed data streams," in *Proc. 11th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2016, pp. 9–16.
- [263] M. Mohamed, S. Ghanem, and M. Nagi, "Privacy-preserving for distributed data streams: Towards l -diversity," *Int. Arab J. Inf. Technol.*, vol. 17, no. 1, pp. 52–64, Jan. 2020.
- [264] M. Asad, A. Moustafa, and T. Ito, "FedOpt: Towards communication efficiency and privacy preservation in federated learning," *Appl. Sci.*, vol. 10, no. 8, p. 2864, Apr. 2020.
- [265] L. Fan, K. W. Ng, C. Ju, T. Zhang, C. Liu, C. S. Chan, and Q. Yang, "Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks," in *Federated Learning*. Cham, Switzerland: Springer, 2020, pp. 32–50.
- [266] L. T. Phong and T. T. Phuong, "Privacy-preserving deep learning via weight transmission," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 11, pp. 3003–3015, Nov. 2019.
- [267] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Mach. Intell.*, vol. 3, no. 6, pp. 473–484, Jun. 2021.
- [268] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security Privacy*, vol. 17, no. 2, pp. 49–58, Mar./Apr. 2019.
- [269] X. Wu, Y. Zhang, M. Shi, P. Li, R. Li, and N. N. Xiong, "An adaptive federated learning scheme with differential privacy preserving," *Future Gener. Comput. Syst.*, vol. 127, pp. 362–372, Feb. 2022.
- [270] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, Q. S. T. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 3454–3469, 2020.
- [271] 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.
- [272] H. Cao, S. Liu, R. Zhao, and X. Xiong, "IFed: A novel federated learning framework for local differential privacy in power Internet of Things," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 5, 2020, Art. no. 1550147720919698.
- [273] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, Dec. 2022.
- [274] L. Silva, N. Magaia, B. Sousa, A. Kobusinska, A. Casimiro, C. X. Mavromoustakis, G. Mastorakis, and V. H. C. de Albuquerque, "Computing paradigms in emerging vehicular environments: A review," *IEEE/CAA J. Autom. Sinica*, vol. 8, no. 3, pp. 491–511, Mar. 2021.
- [275] A. Alwarafy, K. A. Al-Thelaya, M. Abdallah, J. Schneider, and M. Hamdi, "A survey on security and privacy issues in edge-computing-assisted Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4004–4022, Mar. 2021.
- [276] W. Kim and J. Seok, "Privacy-preserving collaborative machine learning in biomedical applications," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIC)*, Feb. 2022, pp. 179–183.
- [277] B. D. Deebak, F. H. Memon, X. Cheng, K. Dev, J. Hu, S. A. Khowaja, N. M. F. Qureshi, and K. H. Choi, "Seamless privacy-preservation and authentication framework for IoT-enabled smart eHealth systems," *Sustain. Cities Soc.*, vol. 80, May 2022, Art. no. 103661.
- [278] J. Li and X. Liao, "Security and privacy in new computing environments (SPNCE 2016)," *Mobile Netw. Appl.*, vol. 26, pp. 2488–2489, Dec. 2022.
- [279] N. Ahmad and P. Chauhan, "State of data privacy during COVID-19," *IEEE Ann. Hist. Comput.*, vol. 53, no. 10, pp. 119–122, Oct. 2020.
- [280] G. Newlands, C. Lutz, A. Tamò-Larrieux, E. F. Villaronga, R. Harasgama, and G. Scheitlin, "Innovation under pressure: Implications for data privacy during the COVID-19 pandemic," *Big Data Soc.*, vol. 7, no. 2, 2020, Art. no. 2053951720976680.
- [281] J. Park and K. Kim, "Image perturbation-based deep learning for face recognition utilizing discrete cosine transform," *Electronics*, vol. 11, no. 1, p. 25, Dec. 2021.
- [282] Z. Sun, L. Yin, C. Li, W. Zhang, A. Li, and Z. Tian, "The QoS and privacy trade-off of adversarial deep learning: An evolutionary game approach," *Comput. Secur.*, vol. 96, Sep. 2020, Art. no. 101876.
- [283] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Digital infrastructure policies for data security and privacy in smart cities," in *Smart Cities Policies and Financing—Approaches and Solutions*. Amsterdam, The Netherlands: Elsevier, pp. 249–261.
- [284] A. Cuzzocrea, C. K. Leung, A. M. Olawoyin, and E. Fadda, "Supporting privacy-preserving big data analytics on temporal open big data," *Proc. Comput. Sci.*, vol. 198, pp. 112–121, Jan. 2022.
- [285] W. Xue, Y. Shen, C. Luo, W. Xu, W. Hu, and A. Seneviratne, "A differential privacy-based classification system for edge computing in IoT," *Comput. Commun.*, vol. 182, pp. 117–128, Jan. 2022.
- [286] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x -vector singular value," *Comput. Speech Lang.*, vol. 73, May 2022, Art. no. 101326.
- [287] I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization," 2022, *arXiv:2202.00443*.
- [288] D. Sadhya and B. Chakraborty, "Quantifying the effects of anonymization techniques over micro-databases," *IEEE Trans. Emerg. Topics Comput.*, early access, Jan. 17, 2022, doi: 10.1109/TETC.2022.3141754.
- [289] J. G. Grisafi, "A marginal religion and COVID-19 in South Korea: Shincheonji, public discourse, and the shaping of religion," *Nova Religio, J. Alternative Emergent Religions*, vol. 25, no. 1, pp. 40–63, 2021.

- [290] J. Sheng, J. Amankwah-Amoah, Z. Khan, and X. Wang, "COVID-19 pandemic in the new era of big data analytics: Methodological innovations and future research directions," *Brit. J. Manage.*, vol. 32, no. 4, pp. 1164–1183, Oct. 2021.
- [291] A. R. Shree, P. Kiran, and S. Chhibber, "Sensitivity context-aware privacy-preserving sentiment analysis," in *Intelligent Manufacturing and Energy Sustainability*. Singapore: Springer, 2021, pp. 407–416.
- [292] X. Xian, T. Wu, Y. Liu, W. Wang, C. Wang, G. Xu, and Y. Xiao, "Towards link inference attack against network structure perturbation," *Knowl.-Based Syst.*, vol. 218, Apr. 2021, Art. no. 106674.
- [293] J. Liu, C. Zhang, K. Xue, and Y. Fang, "Privacy preservation in multi-cloud secure data fusion for infectious-disease analysis," *IEEE Trans. Mobile Comput.*, early access, Jan. 25, 2022, doi: [10.1109/TMC.2022.3145745](https://doi.org/10.1109/TMC.2022.3145745).
- [294] J. Curzon, T. A. Kosa, R. Akalu, and K. El-Khatib, "Privacy and artificial intelligence," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 96–108, Apr. 2021.
- [295] C. Yufei, S. Chao, W. Qian, L. Qi, W. Cong, J. Shouling, L. Kang, and G. Xiaohong, "Security and privacy risks in artificial intelligence systems," *J. Comput. Res. Develop.*, vol. 56, no. 10, p. 2135, 2019.
- [296] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 13, no. 3, pp. 1–207, 2019.
- [297] M. S. Jere, T. Farnan, and F. Koushanfar, "A taxonomy of attacks on federated learning," *IEEE Secur. Privacy*, vol. 19, no. 2, pp. 20–28, Dec. 2021.
- [298] M. A. Ferrag, O. Friha, L. Maglaras, H. Janicke, and L. Shu, "Federated deep learning for cyber security in the Internet of Things: Concepts, applications, and experimental analysis," *IEEE Access*, vol. 9, pp. 138509–138542, 2021.
- [299] P. Treleaven, M. Smietanka, and H. Pithadia, "Federated learning: The pioneering distributed machine learning and privacy-preserving data technology," *Computer*, vol. 55, no. 4, pp. 20–29, Apr. 2022.
- [300] N. Bouacida and P. Mohapatra, "Vulnerabilities in federated learning," *IEEE Access*, vol. 9, pp. 63229–63249, 2021.
- [301] M. Benmalek, M. A. Benrekia, and Y. Challal, "Security of federated learning: Attacks, defensive mechanisms, and challenges," *Revue d'Intell. Artificielle*, vol. 36, no. 1, pp. 49–59, Feb. 2022.
- [302] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
- [303] C.-R. Shyu, K. T. Putra, H.-C. Chen, Y.-Y. Tsai, K. S. M. Hossain, W. Jiang, and Z.-Y. Shae, "A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications," *Appl. Sci.*, vol. 11, no. 23, p. 11191, Nov. 2021.
- [304] N. Rodríguez-Barroso, D. Jiménez López, M. Victoria Luzón, F. Herrera, and E. Martínez-Cámara, "Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges," 2022, [arXiv:2201.08135](https://arxiv.org/abs/2201.08135).
- [305] A. Choudhury, C. Sun, A. Dekker, M. Dumontier, and J. V. Soest, "Privacy-preserving federated data analysis: Data sharing, protection, and bioethics in healthcare," in *Machine and Deep Learning in Oncology, Medical Physics and Radiology*. Cham, Switzerland: Springer, 2022, pp. 135–172.
- [306] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 13, 2021, doi: [10.1109/TKDE.2021.3124599](https://doi.org/10.1109/TKDE.2021.3124599).
- [307] S. R. Pandey, M. N. H. Nguyen, T. N. Dang, N. H. Tran, K. Thar, Z. Han, and C. S. Hong, "Edge-assisted democratized learning toward federated analytics," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 572–588, Jan. 2022.
- [308] M. Milani, Y. Huang, and F. Chiang, "Data anonymization with diversity constraints," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 30, 2021, doi: [10.1109/TKDE.2021.3131528](https://doi.org/10.1109/TKDE.2021.3131528).
- [309] M. D. Jena, S. S. Singhar, B. K. Mohanta, and S. Ramasubbarreddy, "Ensuring data privacy using machine learning for responsible data science," in *Intelligent Data Engineering and Analytics*. Singapore: Springer, 2021, pp. 507–514.
- [310] L. Arbuckle and K. E. Emam, *Building an Anonymization Pipeline: Creating Safe Data*. Sebastopol, CA, USA: O'Reilly Media, 2020.
- [311] S. Aanjanakumar and S. Poonkuntran, "An efficient soft computing approach for securing information over GAMEOVER Zeus botnets with modified CPA algorithm," *Soft Comput.*, vol. 24, no. 21, pp. 16499–16507, Nov. 2020.
- [312] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, H. Zhang, C. Zhang, and Y. Jiang, "Personal big data pricing method based on differential privacy," *Comput. Secur.*, vol. 113, Feb. 2022, Art. no. 102529.
- [313] R. Sanchez-Iborra and A. Skarmeta, "Securing the hyperconnected healthcare ecosystem," in *AI and IoT for Sustainable Development in Emerging Countries*. Cham, Switzerland: Springer, 2022, pp. 455–471.
- [314] J. B. Awotunde, R. G. Jimoh, R. O. Ogundokun, S. Misra, and O. C. Abikoye, "Big data analytics of IoT-based cloud system framework: Smart healthcare monitoring systems," in *Artificial Intelligence for Cloud and Edge Computing*. Cham, Switzerland: Springer, 2022, pp. 181–208.
- [315] Y. Miao, W. Zheng, X. Jia, X. Liu, K.-K.-R. Choo, and R. Deng, "Ranked keyword search over encrypted cloud data through machine learning method," *IEEE Trans. Services Comput.*, early access, Jan. 4, 2022, doi: [10.1109/TSC.2021.3140098](https://doi.org/10.1109/TSC.2021.3140098).
- [316] M. A. Khelili, S. Slatnia, O. Kazar, and S. Harous, "IoMT-fog-cloud based architecture for COVID-19 detection," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103715.
- [317] M. L. Alarcon, R. Oruche, A. Pandey, and P. Callyam, "Cloud-based data pipeline orchestration platform for COVID-19 evidence-based analytics," in *Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 159–180.
- [318] P. K. Rao and D. Rawtani, "Modern digital techniques for monitoring and analysis," in *COVID-19 in the Environment*. Amsterdam, The Netherlands: Elsevier, 2022, pp. 115–130.
- [319] M. Li, N. Varble, B. Turkbey, S. Xu, and B. J. Wood, "Camera-based distance detection and contact tracing to monitor potential spread of COVID-19," *Proc. SPIE*, vol. 12035, pp. 329–335, Apr. 2022.
- [320] X. Zhang, J. Hamm, M. K. Reiter, and Y. Zhang, "Defeating traffic analysis via differential privacy: A case study on streaming traffic," *Int. J. Inf. Secur.*, pp. 1–18, Jan. 2022, doi: [10.1007/s10207-021-00574-3](https://doi.org/10.1007/s10207-021-00574-3).
- [321] H. Tang, J. Chen, Y. Zhou, and L. Chen, "A novel resource management scheme for virtualized cyber-physical-social system," *Phys. Commun.*, vol. 50, Feb. 2022, Art. no. 101513.
- [322] M. Snehi and A. Bhandari, "A novel distributed stack ensemble meta-learning-based optimized classification framework for real-time prolific IoT traffic streams," *Arabian J. Sci. Eng.*, pp. 1–24, Jan. 2022, doi: [10.1007/s13369-021-06472-z](https://doi.org/10.1007/s13369-021-06472-z).
- [323] S. Xiong, A. D. Sarwate, and N. B. Mandayam, "Network traffic shaping for enhancing privacy in IoT systems," *IEEE/ACM Trans. Netw.*, early access, Jan. 12, 2022, doi: [10.1109/TNET.2021.3140174](https://doi.org/10.1109/TNET.2021.3140174).
- [324] S.-H. Liao, R. Widowati, and P. Puttong, "Data mining analytics investigate Facebook live stream users' behaviors and business models: The evidence from Thailand," *Entertainment Comput.*, vol. 41, Mar. 2022, Art. no. 100478.
- [325] C. Karras, A. Karras, and S. Sioutas, "Pattern recognition and event detection on IoT data-streams," 2022, [arXiv:2203.01114](https://arxiv.org/abs/2203.01114).
- [326] F. Chen, W. Wang, H. Yang, W. Pei, and G. Lu, "Multiscale feature fusion for surveillance video diagnosis," *Knowl.-Based Syst.*, vol. 240, Mar. 2022, Art. no. 108103.
- [327] G. Aguiar, B. Krawczyk, and A. Cano, "A survey on learning from imbalanced data streams: Taxonomy, challenges, empirical study, and reproducible experimental framework," 2022, [arXiv:2204.03719](https://arxiv.org/abs/2204.03719).
- [328] D.-H. Vu, "Privacy-preserving naive Bayes classification in semi-distributed data model," *Comput. Secur.*, vol. 115, Apr. 2022, Art. no. 102630.
- [329] M. Lango and J. Stefanowski, "What makes multi-class imbalanced problems difficult? An experimental study," *Expert Syst. Appl.*, vol. 199, Aug. 2022, Art. no. 116962.
- [330] T. Li, H. Wang, D. He, and J. Yu, "Blockchain-based privacy-preserving and rewarding private data sharing for IoT," *IEEE Internet Things J.*, early access, Jan. 31, 2022, doi: [10.1109/JIOT.2022.3147925](https://doi.org/10.1109/JIOT.2022.3147925).
- [331] H. Wen, M. Wei, D. Du, and X. Yin, "A blockchain-based privacy preservation scheme in mobile medical," *Secur. Commun. Netw.*, vol. 2022, pp. 1–11, Mar. 2022.
- [332] M. Sarrab and F. Alshohoumi, "Assisted fog computing approach for data privacy preservation in IoT-based healthcare," in *Security and Privacy Preserving for IoT and 5G Networks*. Cham, Switzerland: Springer, 2022, pp. 191–201.
- [333] D. Peng, L. Sun, R. Zhou, and Y. Wang, "Study QoS-aware fog computing for disease diagnosis and prognosis," *Mobile Netw. Appl.*, pp. 1–8, Apr. 2022, doi: [10.1007/s11036-022-01957-z](https://doi.org/10.1007/s11036-022-01957-z).

[334] Y. Xu, M. Z. A. Bhuiyan, T. Wang, X. Zhou, and A. Singh, "C-fDRL: Context-aware privacy-preserving offloading through federated deep reinforcement learning in cloud-enabled IoT," *IEEE Trans. Ind. Informat.*, early access, Feb. 8, 2022, doi: 10.1109/TII.2022.3149335.

[335] J. Wang, "Workflow offloading with privacy preservation in a cloud-edge environment," *Concurrency Comput., Pract. Exper.*, p. e7002, Apr. 2022, doi: 10.1002/cpe.7002.

[336] D. Wang, S. Shi, Y. Zhu, and Z. Han, "Federated analytics: Opportunities and challenges," *IEEE Netw.*, vol. 36, no. 1, pp. 151–158, Jan. 2022.

[337] S. Flowerday and C. Xenakis, "Security and privacy in distributed healthcare environments," *Methods Inf. Med.*, pp. 1–4, Feb. 2022, doi: 10.1055/a-1768-2966.

[338] L. Sankar, "Lalitha Sankar, Arizona state university: Federated analytics based contact tracing for COVID-19," Arizona State Univ., USA, Tech. Rep. 2031799, Apr. 2021.

[339] A. Kallel, M. Rekik, and M. Khemakhem, "Hybrid-based framework for COVID-19 prediction via federated machine learning models," *J. Supercomput.*, vol. 78, no. 5, pp. 7078–7105, 2021.

[340] F. Kserawi, S. Al-Marri, and Q. Malluhi, "Privacy-preserving fog aggregation of smart grid data using dynamic differentially-private data perturbation," *IEEE Access*, vol. 10, pp. 43159–43174, 2022.

[341] C. Stergiou, K. E. Psannis, B.-G. Kim, and B. Gupta, "Secure integration of IoT and cloud computing," *Future Gener. Comput. Syst.*, vol. 78, pp. 964–975, Jan. 2018.



ABDUL MAJEED received the B.S. degree in information technology from UIIT, PMAS-UAAR, Rawalpindi, Pakistan, in 2013, the M.S. degree in information security from COMSATS University, Islamabad, Pakistan, in 2016, and the Ph.D. degree in computer information systems and networks from Korea Aerospace University, South Korea, in 2021. He worked as a Security Analyst with Trillium Information Security Systems (TISS), Rawalpindi, from 2015 to 2016. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, South Korea. His research interests include privacy preserving data publishing, statistical disclosure control, privacy-aware analytics, federated learning, and machine learning.



SAFIULLAH KHAN received the B.Sc. degree in electronic engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2013, and the M.Sc. degree in electrical engineering from COMSATS University Islamabad, Abbottabad Campus, Pakistan, in 2017. He is currently pursuing the Ph.D. degree in computer engineering with Gachon University, Seongnam, South Korea. He worked with the Research and Development Department, National Radio and Telecommunication Corporation, Haripur, Pakistan, for two years. He is also the Chair of IEEE Student Branch at Gachon University. His research interests include efficient hardware implementations of cryptographic protocols, blockchain, and network security.



SEONG OUN HWANG (Senior Member, IEEE) received the B.S. degree in mathematics from Seoul National University, in 1993, the M.S. degree in information and communications engineering from the Pohang University of Science and Technology, in 1998, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology, South Korea, in 2004. He worked as a Software Engineer with LG-CNS Systems Inc., from 1994 to 1996. He also worked as a Senior Researcher with the Electronics and Telecommunications Research Institute (ETRI), from 1998 to 2007. He worked as a Professor with the Department of Software and Communications Engineering, Hongik University, from 2008 to 2019. He is currently working as a Full Professor with the Department of Computer Engineering, Gachon University, South Korea. His research interests include cryptography, cybersecurity, and artificial intelligence.

...