Towards Privacy Preservation using Clustering based Anonymization: Recent Advances and Future Research Outlook

ABDUL MAJEED, SAFIULLAH KHAN, and SEONG OUN HWANG, (Senior Member, IEEE) 
1Department of Computer Engineering, Gachon University, Seongnam 13120, Republic of Korea 
2Department of IT Convergence Engineering, Gachon University, Seongnam 13120, Republic of Korea

Corresponding author: Seong Oun Hwang (e-mail: sohwang@gachon.ac.kr) and Abdul Majeed (e-mail: 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 under Grant 2021-0-01532 (50%), and in part by the National Research Foundation of Korea (NRF) under Grant 2020R1A2B5B01002145 (50%), all funded by the Korean Government through Ministry of Science and ICT (MSIT).

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. 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

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, 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
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 \( \rightarrow 87\% \)
- place of residence, date of birth, gender \( \rightarrow 50\% \)
- country of origin, date of birth, gender \( \rightarrow 18\% \)

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 solu-
tions 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 op-
erations, 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 ap-
proaches applied to tabular data are \( k \)-anonymity [7], \( \ell \) -
diversity [8], and \( \ell \) -closeness [9]. These models showed
remarkable results in terms of privacy preservation in the
eye. However, they proved unsuccessful against cer-
tain 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 origi-
inal. For example, in 2006, differential privacy (DP) [12]
was proposed for dynamic scenarios (e.g., query-answer).

Afterwards, researchers extended the anonymization con-
cepts 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 fo-
cus 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 opera-
tions, data anonymity frameworks, privacy-protection tools,
and evaluation metrics used by the PPDP mechanisms. Ra-
jendran et al. [16] discussed the strengths and weaknesses
of three famous anonymity models: \( k \)-anonymity, \( \ell \) -diversity,
and \( \ell \) -closeness. Tran and Hu [17] provided a systematic
review of big data analytics that preserves privacy. Other
authors have discussed many generic privacy-preserving ap-
proaches for data querying, data publishing, and data min-
ing. A few surveys have been published on outsourcing SN
users’ data while preserving privacy [18], [19]. Some au-
thors 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 cov-
ered 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 review to the PPDP field
are summarized as follows. (i) We summarize the key find-
ings of state-of-the-art (SOTA) clustering-based anonymiza-
tion mechanisms that have been proposed for effective reso-
lution of privacy and utility in PPDP. (ii) 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 of the latest anonymization techniques and
metrics employed for evaluations. (iii) We describe the role of
CAMs regarding privacy preservation in different computing
paradigms, such as cloud computing, the IoT, location-based
services, and application-specific SN scenarios (community
clustering, collaborative filtering, privacy-aware recommen-
dations, graph mining, etc.) that remained unexplored in the
recent literature. (iv) This paper highlights various represen-
tative 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). (v) We present the technical
challenges in protecting user privacy by leveraging CAMs,
and list potential avenues for future research to address contemporary privacy threats. (vi) This is the first work centering 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 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](image_url)

**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 matrixes. 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 continuous fashion called streams. Figure 3 presents a generic overview of 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].

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 explo-
sure, 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 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., 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.

In recent years, AI approaches have opened up new chal-
lenges 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, preprocessing 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.

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, $\ell$-diversity, $\ell$-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).

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 between two users (or between a user and the cluster center) is given in Eq. 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:

\[
T_{\text{users,attributes}} = \left( \begin{array}{cccccc}
  u_1 & Q_{i1} & Q_{i2} & \cdots & Q_{ip} & S \\
  u_2 & v_{q1} & v_{q2} & \cdots & v_{qj} & v_1 \\
  u_3 & v_{q1} & v_{q2} & \cdots & v_{qj} & v_2 \\
  \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  N & v_{q1} & v_{q2} & \cdots & v_{qj} & v_n \\
\end{array} \right)
\]

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

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

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

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.

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 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.

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.

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].

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 widowering concept during conversion of raw data into anonymized data.

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.

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 [116] 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 [117]–[119]. 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.
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 |
| Khosravi et al. [52] | Table | Improving 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 |
| Jafari et al. [54] | Table | Ensured better privacy in healthcare sector using clustering concepts and by improving k-closeness | Single-linkage clustering |
| Zouining 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 anonymization | K means algorithm |
| Ahmad et al. [57] | Table | k-diversity based model for better privacy protection in big data platforms (e.g., Apache Spark) | K-means++ algorithm |
| Abbas et al. [58] | Table | Addressed scalability and utility issues in PPDP involving high dimensional data | K-means algorithm |
| Gaspar et al. [59] | Table | Ensured privacy protection against five attacks by using k-anonymity and clustering | K-means algorithm |
| Yang et al. [60] | Table | Lowered the influence of outliers in the clustering process and improved data availability in PPDP | Bottom-up clustering |
| Rapetti et al. [61] | Graph | Ensured high-level of privacy compared to SOTA methods using clustering concepts on SN data | Weighted K-means algorithm |
| Zhu et al. [62] | Graph | Minimized the information loss in clustering data using k-degree and density clustering | K means algorithm |
| Shaked et al. [63] | Graph | Provided strong privacy guarantees against active attacks while sustaining the usefulness of data | K-means algorithm |
| Chen et al. [64] | Graph | Maintained the balance between utility of mobile SN data and performance of authentication | K-means algorithm |
| Debiasi et al. [65] | Graph | Provided better protection against identity disclosure and maintained higher utility | Node clustering |
| Shokarlo et al. [66] | Graph | Ensured strong privacy protection in weighted graphs using k-anonymity and clustering | Hierarchical clustering |
| Luangani et al. [67] | Graph | Provided strong privacy protection in SN data such as Google+, Facebook, YouTube, and Twitter | Node and edge clustering |
| Reza et al. [68] | Graph | Provided superior results in information loss, privacy, and time complexity in GNN anonymization | K-member fuzzy clustering |
| Kumar et al. [69] | Graph | Resolved the privacy and utility trade-off in anonymizing SN data enclosed in a G form | Edge clustering |
| Heidari et al. [70] | Graph | Analyzed the distribution of nodes in clustering process to improve running time and utility of G | Fuzzy clustering |

TABLE 2. 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 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.) [130]–[134]. Furthermore, some studies also used application-specific metrics to...
rather than relying on one or two metrics, while evaluating
metric may not be monotonic. Hence, some approaches have
Table 2. Detailed analysis (i.e., strengths and weaknesses) of the SOTA studies presented in Table 1.

| Ref. | Study nature | Strengths | Weaknesses |
|------|--------------|-----------|------------|
| Zeng et al. [32] | Practical | Encourage the distribution of QFs 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 values |
| Kubrin et al. [31] | Practical | Ensures higher utility for data mining by increasing processing time, i.e., fewer attributes | The privacy breaches can be higher due to the linkage attacks with auxiliary data (i.e., voter list) |
| Khan et al. [33] | Pervasive | Ensures minimal changes in data anonymization and better data to construction | Promote to background knowledge and other practical attacks (i.e., election, table linkages, etc.) |
| Potam et al. [34] | Conceptual | Better privacy preservation using enhanced k-clustering concept | Poor utility due to suppressing operation and more diversity in SA values |
| Zhang et al. [35] | Theoretical | Effective re-organization of privacy-utility trade-off in off-line data publishing approaches | Promote to the disclosure of personal information and other detection attacks |
| Alahoud et al. [36] | Practical | High level of privacy-utility trade-off in big data environments | Fails to provide privacy and utility when data is highly distributed (distribution is uneven) |
| Abhau et al. [37] | Practical | A low-cost anonymization with a 5.5% reduction in IL and 3.5× reduction in time | Promote to skewness and loss applicability to highly distributed databases |
| Ouamen et al. [38] | Practical | Guarantees user privacy at the collection time in healthcare sectors using clustering | Disclose less frequent data items that may hinder knowledge discovery process |
| Yan et al. [39] | Practical | Efficiently anonymizes data in the presence of outliers and lowers the IL, as well as clustering effects | Data to precise and robust privacy disclosure as well as in-thesis linkages with the data available at auxiliary sources |

TABLE 2. 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 privacy, and privacy preservation from active attacks, etc. |
| Utility | Information loss, distortion, accuracy, F-measure, coverage of generalized values, discriminability metrics, information theoretic metrics, size of equivalence classes, degree preservation, precision, normalized mutual information, recall, clustering, K-L-divergence, queries’ accuracy, network resilience, centralities, amount of knowledge, authority score, effective diameter, and data mining tasks, etc. |

quantify the level of privacy and utility [135]–[137]. In many real-world cases, the privacy measured by one evaluation metric may not be monotonic. Hence, some approaches have suggested employing a metrics suit (i.e., multiple metrics), rather than relying on one or two metrics, while evaluating the performance of anonymization methods [138], [139]. With the rapid increase in the diversity 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 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](image)

**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.

![Figure 21](image)

**Figure 21.** Overview of emerging computing paradigms.

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 healthcare, 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.

![Figure 22](image)

**Figure 22.** Quantitative analysis of CAMs proposed for different computing paradigms.

From the analysis, it can be observed that CAMs can improve both the privacy and utility results significantly.
### 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             | Privacy-preserving cluster analysis and data sharing                                | K-Means clustering                        | Data mining                                   | Logeswaran et al. [140] |
|                             | Privacy-preserving healthcare data sharing                                          | Parallel Clustering                       | Big data handling                             | Usha et al. [144] |
|                             | Prevention of privacy breaches                                                      | k-annealing clustering                    | Medical applications                          | Zhang et al. [142] |
|                             | Protection from similarity and inference attacks                                    | (G, S) clustering                         | Data mining and analysis                      | Naughty et al. [143] |
|                             | Ensured the higher confidentiality of data                                         | KNN(G, S) clustering                     | Privacy-preserving keyword search            | Singh et al. [144] |
|                             | Privacy-preserving healthcare data sharing                                          | k-means clustering                         | Healthcare data analysis                      | Li et al. [145] |
|                             | Maintained privacy of data in cloud environment                                      | ECDOSA algorithm                          | Healthcare industry                           | Jayaraman et al. [146] |
|                             | Data protection from 3rd parties                                                    | Dragonfly (DP) algorithm                  | Data analytics                                | Madian et al. [147] |
|                             | Privacy-preserved data storage and processing                                       | Adaptive Dragon-PQO                      | Data analytics                                | Madian et al. [148] |
|                             |                                                                                 | Fuzzy C means                             | Medical applications                          | Shammea et al. [149] |
| Location-based Services     | Privacy protection of location information                                         | Halo clustering                           | Utility analytics                             | Fri et al. [151] |
|                             | Privacy mechanism to control locations tracking                                    | Hierarchical clustering                   | Query answering                               | Lee et al. [152] |
|                             | Strong privacy protection in geo-matching attacks                                    | k-implicit clustering                      | Location-based services                       | Nie et al. [153] |
| Social Networks             | Protection of location privacy                                                      | Clustering                               | Mobile applications                          | Yao et al. [154] |
|                             | Preservation of location anonymity                                                  | Enhanced cloaking                         | Location-based services                       | Liu et al. [155] |
|                             | Strongly protected users’ location privacy                                         | Spatial cloaking                          | Location-based services                       | Zhang et al. [156] |
|                             | Maintain privacy of location, choices, & comments                                   | Spatial clustering                        | Communication apps                           | Alhajwani et al. [157] |
|                             | Privacy protection of trajectory data                                              | Groovy clustering                         | Location-based services                       | Mahdavifar et al. [158] |
|                             | Privacy protection of spatiotemporal locations data                                 | K-means clustering                        | Location-based services                       | Chen et al. [160] |
|                             | Personalized privacy protection in trajectories                                      |                                                       |                                               | Anonymous et al. [161] |

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.

### 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. [200] described an identity-revelation method based on attributes from the anonymized $G$. The authors achieved accuracy of up to 80% in de-anonymization of identities. Similarly, some de-anonymization approaches have used the community-clustering concept to group users, and users’ de-anonymization was subsequently successful [201]. Some approaches have jointly used clustering and attribute
TABLE 5. Detailed analysis (i.e., strengths and weaknesses) of the SOTA studies presented in Table 4.

| Author | Year | Study | Research Gap | Strengths | Weaknesses | Study Nature |
|--------|------|-------|--------------|-----------|------------|--------------|
| Gao et al. [188] | 2018 | Strong privacy protection at user-level and cluster centers-level by using encryption | Failed to provide resilience against probabilistic, decision, similarity, and anomalous attacks | Practical |
| Khan et al. [186] | 2018 | Strong privacy protection in moving databases using | Potential for use in changing environments and databases | Practical |
| Liu et al. [170] | 2018 | Strong defense against identity disclosure problem using | Prone to community privacy disclosure and SA disclosure if the random number is not secure | Practical |
| Liu et al. [179] | 2018 | 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 |
| Mahdavifar et al. [158] | 2018 | Personalized privacy protection of location data considering moving objects requirements | Prone to identity and SA disclosure when certain users specify lower privacy preferences | Practical |
| Mallick et al. [164] | 2018 | 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 |
| Maudgal et al. [163] | 2018 | Fostering social mining and graph analytics by making less changes in the graph structure | Prone to community privacy disclosure and SA disclosure in case of high-dimensional data | Practical |
| Nath et al. [162] | 2018 | Strong privacy protection of users from high dimensional social networks data | Prone to the disclosure of group privacy as well as data in the form of graph patterns | Practical |
| Otgonbayar et al. [195] | 2018 | 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 |
| Pathak et al. [197] | 2018 | Fosters data re-usability by sharing it at a large scale without compromising user’s privacy | Prone to SA disclosure and community privacy disclosure | Practical |
| Qiu et al. [182] | 2018 | Privacy-preserving perturbation based on user-level and cluster centers-level | Potential for use in changing environments and databases | Practical |
| Rajesh et al. [183] | 2018 | Ensures privacy protection in association rule mining cases via perturbation-based approach | Less resilience against SA reconstruction, derivation, and prediction in medical environments | Practical |
| Ray et al. [181] | 2018 | Privacy protection in data mining using federated learning | Failure to prevent privacy breaches as data is relatively small and belong to an identical sector/domain | Practical |
| Ros et al. [180] | 2018 | Strong privacy protection in multi-dimension social networking data | Prone to the disclosure of group privacy as well as data in the form of graph patterns | Practical |
| Ullah et al. [197] | 2018 | Fosters data re-usability by sharing it at a large scale without compromising user’s privacy | Prone to SA disclosure and community privacy disclosure | Practical |
| Truta et al. [163] | 2018 | Fostering social mining and graph analytics by making less changes in the graph structure | Prone to community privacy disclosure and SA disclosure in case of high-dimensional data | Practical |
| Zhang et al. [142] | 2018 | Highly efficient and scalable solution towards big data anonymization using | User-and group level privacy breaches cannot be effectively prevented from the segmented area | Practical |

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 clustering methods, more secure and resilient anonymity methods can be developed to protect 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 last column of the table included pertinent studies from which information to correctly identify individuals from privacy-preserved published graphs [202]. Shao et al. [203] 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.
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          | Localization-based services                                                                | Gambis et al. [204]    |
| SN data (i.e., graph)         | Network de-anonymization with higher matches                                                      | Node clustering                  | Social network services                                                                    | Chiasserini et al. [205]|
| SN data (i.e., graph)         | Network de-anonymization via user relations                                                       | Community clustering             | Social network services                                                                    | Chiasserini et al. [206]|
| SN data (i.e., graph)         | Re-identifying users in anonymized G by mapping                                                 | Nodal Clustering                 | Social network services                                                                    | Carnielli et al. [207]  |
| SN data (i.e., graph)         | De-anonymizing users/users焱by mapping                                                           | Partitional clustering           | Social network services                                                                    | Fu et al. [208]         |
| SN data (i.e., graph)         |                                                                                                  | Community clustering             | Social network services                                                                    | Fu et al. [209]         |
| SN data (i.e., graph)         |                                                                                                  | l-means clustering               | Social network services                                                                    | Cellular networks       |
| SN data (i.e., graph)         |                                                                                                  | m-means clustering               | Social network services                                                                    | Francis et al. [210]    |
| SN data (i.e., graph)         |                                                                                                  | Structural clustering            | Collaborative learning                                                                    | Orkendy et al. [211]    |
| SN data (i.e., graph)         |                                                                                                  |                      | Location-based services                                                                    | Chen et al. [212]       |
| SN data (i.e., graph)         |                                                                                                  | Statistical matching            | Location-based services                                                                    | Murakami et al. [213]   |
| SN data (i.e., graph)         |                                                                                                  |                                 | Location-based services                                                                    | Li et al. [214]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Location-based services                                                                    | Zheng et al. [215]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Location-based services                                                                    | Wang et al. [216]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Location-based services                                                                    | Zheng et al. [217]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Chen et al. [218]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Murakami et al. [219]   |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Ma et al. [220]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Internet of things                                                                       | Tabbib et al. [221]     |
| SN data (i.e., graph)         |                                                                                                  |                                 | Knowledge-based systems                                                                   | Shirali et al. [222]    |
| SN data (i.e., graph)         |                                                                                                  |                                 | Blockchain-based services                                                                | Alahbati et al. [223]   |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Li et al. [224]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Xu et al. [225]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Li et al. [226]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Xu et al. [227]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Multimodal applications                                                                  | Yang et al. [228]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Web applications                                                                         | Mediani et al. [229]    |
| SN data (i.e., graph)         |                                                                                                  |                                 | Location-based services                                                                    | Nardini et al. [230]    |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Nandi et al. [231]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Zhang et al. [232]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Chen et al. [233]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network security                                                                   | Murakami et al. [234]   |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Tico et al. [235]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Li et al. [236]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Ashok et al. [237]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Bragg et al. [238]      |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Li et al. [239]         |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Chen et al. [240]       |
| SN data (i.e., graph)         |                                                                                                  |                                 | Social network services                                                                    | Ong et al. [241]        |
| SN data (i.e., graph)         |                                                                                                  |                                 | Cyber-physical systems                                                                   | Castro et al. [242]     |
| SN data (i.e., graph)         |                                                                                                  |                                 | Cyber-physical systems                                                                   | Castro et al. [243]     |
| SN data (i.e., graph)         |                                                                                                  |                                 |kehrmechanism and lower availability of auxiliary data can restrict the re-identification rate. |                        |

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

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 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.

![FIGURE 26. Overview of open challenges in information privacy domain.](image)

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 [245]. 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 [246]. 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 [247], [248]. 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 [249]–[251]. 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 [252]. 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 [253]. Although some efforts have been devoted to lowering the computing complexity in CAMs [254, 255], 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 [256]–[259]. AI-powered attacks can be launched to disclose identities, SAs, and memberships from large and complex datasets with the help of hyperparameter tuning [260]. 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 [261]. 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., $\ell$-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, $\ell$-diversity, and $t$-closeness, cannot guarantee consistent performance when the original data encompass outliers [262]. 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 [263], [264]. 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 [141]. Recently, some parallel clustering algorithms have been devised to address data diversity and heterogeneity issues during anonymization [265]–[267]. 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 [268], [269]. 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 [270]–[274]. 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 [275]–[279]. 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 op-
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 [280]–[284]. 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) [285], [286]. 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 [287], videos [288], stream data [289], and temporal data [290] 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 [291]. 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 \((k, \epsilon, t, \ell, \text{etc.})\), and each parameter has a distinct impact on privacy and utility [292]. 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 [293]. 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 [294]. 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 [295]. 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 [296].

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 [297], [298]. 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 [252]. 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 [299]. 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.

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 [300]. 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 preprocessing, ML models, and output [301]. With the advent of federated learning [302], 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 [303].

![Figure 29. Overview of federated learning paradigm.](image)

Recently, Ferrag et al. [304] 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). Treleaven et al. [305] 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. [306] 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. [307] 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. [308] 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. [309] 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 [310]–[312].

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 [313]. 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 [314]–[316]. 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 [317]. 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. [318] 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. [319] 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. [320] 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 [321]. This technology has been extensively used in real-time applications for data mining-related tasks without compromising users’ privacy.

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. [322]. Similarly, a real-time platform, named OnTimeEvidence was proposed by Alarcon et al. [323] 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 [324], [325]. 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. [326] recently proposed a privacy-preserved real-time system for streaming traffic using the DP model. Tang et al. [327] 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 [328]–[331]. 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. [332] 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 [333]. Vu et al. [334] 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 [335]. Blockchain technology has revolutionized the privacy domain by removing the heavy reliance on a central server. Li et al. [336] discussed a provably secure method for privacy preservation in real-time IoT applications. Wen et al. [337] 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. [338] discussed a fog computing method for preserving data privacy in IoT-based healthcare. The proposed healthcare internet of things (H-IoT)-based framework classifies

![Relevance Figure](Image)

**Figure 30.** Overview of real-time computing (RTC) applications.
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. [338] 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 [339]. 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. [340] 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 [341] 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.

In 2017, Google coined the term federated analytics (FA), performing analytics on local devices with data in way analogous to FL. See Figure 31.

FA is another real-time technology resulting from FL and is based on the distributed computing paradigm [342]. It has been used in diverse fields such as medical, supply chain, finance, and energy sector for online analytics [343]. 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 [344], for analytics of the impact of COVID-19 and other diseases, in the categorization of effects from COVID-19 based on demographics, for identification of COVID-19 risk factors, prediction of vulnerability indexes [345], 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.

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 [346]. Furthermore, we suggest devising technical privacy-preserving solutions 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-

1https://www.stonebranch.com/
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’ interests dramatically. CAMs have demonstrated their effectiveness in improving various technical aspects of traditional anonymization methods (e.g., \(k\)-anonymity, \(\ell\)-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, matrices, logs, streams, traces, multimedia, text, documents, and hybrids), and we summarized and analyzed key features, including 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 threat 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. Springer, 2021.
[2] J. Wieringa, P. Kannan, X. Ma, T. Reuterer, H. Risselada, and B. Skiera, “Data analytics in a privacy-concerned world,” Journal of Business Research, vol. 122, pp. 915–925, 2021.
[3] B. C. Fung, K. Wang, A. W.-C. Fu, and S. Y. Philip, Introduction to privacy-preserving data publishing: Concepts and techniques. Chapman and Hall/CRC, 2017.
[4] L. Sweeney, “Simple demographics often identify people uniquely,” Health (San Francisco), vol. 671, no. 2000, pp. 1–34, 2000.
[5] S. K. Kros, M. P. Janssen, R. H. Groenwold, and M. van Leeuwen, “Evaluating privacy of individuals in medical data,” Health Informatics Journal, vol. 27, no. 2, p. 146045822093398, 2021.
[6] F. YAĞAR, “Growing concern during the covid-19 pandemic: Data privacy,” Türkije Klinikeyi Health Sci, vol. 6, no. 2, pp. 387–92, 2021.
[7] L. Sweeney, “\(k\)-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.
[8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “\(t\)-diversity: Privacy beyond \(k\)-anonymity,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 1, no. 1, pp. 3–es, 2007.
[9] N. Li, T. Li, and S. Venkitasubramaniam, “\(t\)-closeness: Privacy beyond \(k\)-anonymity and \(l\)-diversity,” in 2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007, pp. 106–115.
[10] J. Han, H. Yu, and J. Yu, “An improved \(l\)-diversity model for numerical sensitive attributes,” in 2008 Third International Conference on Communications and Networking in China. IEEE, 2008, pp. 938–943.
[11] N. Li, T. Li, and S. Venkitasubramaniam, “Closeness: A new privacy measure for data publishing,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 7, pp. 943–956, 2009.
[12] C. Dwork, “Differential privacy: A survey of results,” in International conference on theory and applications of models of computation. 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 Network Analysis and Mining, vol. 4, no. 1, pp. 1–40, 2014.
[14] X. Li, J. Yang, Z. Sun, and J. Zhang, “Differential privacy for edge weights in social networks,” Security and Communication Networks, vol. 2017, 2017.
[15] J. Casas-Roma, J. Salas, F. D. Malliaros, and S. Vazirgiannis, “\(k\)-degree anonymity on directed networks,” Knowledge and Information Systems, vol. 61, no. 3, pp. 1743–1768, 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,” Journal of Parallel and Distributed Computing, vol. 134, pp. 207–218, 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. 19 912–19 922, 2018.
[19] B. YANG, T. ZHU, W. ZHOU, and Y. XIANG, “Attacks and counters in online social network data publishing,” ZTE Communications, vol. 14, no. 00, pp. 2–9, 2019.
[20] A. Sharma, G. Singh, and S. Rehman, “A review of big data challenges and preserving privacy in big data,” Advances in Data and Information Science and Applications, pp. 57–65, 2020.
[21] A. Majeed and S. Lee, “Anonymization techniques for privacy preserving data publishing: A comprehensive survey,” IEEE Access, vol. 9, pp. 8 512–8 545, 2020.
[22] M. Cunha, R. Mendes, and J. P. Vilela, “A survey of privacy-preserving mechanisms for heterogeneous data types,” Computer Science Review, vol. 41, p. 100403, 2021.

\(^2\)https://redasci.org/
et al. S. Ghiasvand and F. M. Ciorba, “Anonymization of system logs for privacy protection in IoT stream data,” in 2021 5th International Conference on Internet of Things and Applications (IoT). IEEE, 2021, pp. 1–5.

J. Kumar, “Slide window method adapted for privacy-preserving: Transactional data streams,” European Journal of Medical & Clinical Medicine, vol. 8, no. 2, pp. 2528–2538, 2021.

L. Yang, X. Chen, Y. Luo, X. Lan, and W. Wang, “Idea: A utility-enhanced approach to incomplete data stream anonymization,” Tsinghua Science and Technology, vol. 27, no. 1, pp. 127–140, 2021.

F. Dufaux and T. Ebrahimi, “A framework for the validation of protection solutions in video surveillance,” in 2010 IEEE International Conference on Multimedia and Expo. IEEE, 2010, pp. 66–71.

R. E. Thomas, S. K. Banu, and B. Tripathy, “Image anonymization using clustering with pixelization,” Int. J. Eng. Technol, vol. 7, pp. 990–993, 2018.

W. Zheng, Z. Wang, T. Lv, Y. Ma, and C. Jia, “K-anonymity algorithm based on improved clustering,” in International conference on algorithms and architectures for parallel processing. Springer, 2018, pp. 462–476.

K. Mohammed, A. Ayesh, and E. Boiten, “Utility promises of self-organising maps in privacy preserving data mining,” in Data Privacy Management, Cryptocurrencies and Blockchain Technology. Springer, 2020, pp. 55–72.

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,” arXiv preprint arXiv:2001.01491, 2020.

P. Khan, Y. Khan, and S. Kumar, “Single identity clustering-based data anonymization in healthcare,” in Computationally Intelligent Systems and their Applications. Springer, 2021, pp. 1–9.

S. Zoumina, N. Grozavu, Y. Bennani, A. Lyhyaoui, and N. Rogovschi, “Efficient k-anonymization through constrained collaborative clustering,” in 2018 IEEE IPSN Series on Computational Intelligence (SSCI). IEEE, 2018, pp. 405–411.

W. Zheng, Y. Ma, Z. Wang, C. Jia, and P. Li, “Effective l-diversity anonymization algorithm based on improved clustering,” in International Symposium on Cyberspace Safety and Security. Springer, 2019, pp. 318–329.

F. Ashkouti, K. Khamforoosh, A. Sheikhhahmadi, and H. Khamfroush, “Diksmeans-l-diversity: distributed hierarchical k-means for satisfaction of the l-diversity privacy model using apache spark,” The Journal of Supercomputing, vol. 78, no. 2, pp. 2616–2650, 2022.

A. Abbasi and B. Mohammadi, “A clustering-based anonymization approach for privacy-preserving in the healthcare cloud,” Concurrency and Computation: Practice and Experience, vol. 34, no. 1, p. e6487, 2022.

J. A. Onesim, J. Karthikeyan, and Y. Sei, “An efficient clustering-based anonymization scheme for privacy-preserving data collection in iot based healthcare services,” Peer-to-Peer Networking and Applications, vol. 14, no. 3, pp. 1629–1649, 2021.

Y. Yan, E. A. Herman, A. Mahmood, T. Feng, and P. Xie, “A weighted k-member clustering algorithm for k-anonymization,” Computing, vol. 103, no. 10, pp. 2251–2273, 2021.

R. Gangarde, A. Sharma, A. Pawar, R. Joshi, and S. Gonga, “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, 2021.

H. Zhang, L. Lin, L. Xu, and X. Wang, “Graph partition based privacy-preserving scheme in social networks,” Journal of Network and Computer Applications, vol. 195, p. 103214, 2021.

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, 2021.

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 Journal on Wireless communications and networking, vol. 2016, no. 1, pp. 1–9, 2016.

D. Mohapatra and M. R. Patra, “Graph anonymization using hierarchical clustering,” in Computational Intelligence in Data Mining. Springer, 2019, pp. 145–154.

M. E. Skarkala, M. Maragoudakis, S. Gritzalis, L. Mitrou, H. Toivonen, and P. Moen, “Privacy preservation by k-anonymization of weighted social networks,” in 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2012, pp. 423–428.

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 Systems with Applications, vol. 141, p. 112968, 2020.
[86] X.-B. Li and J. Qin, “Anonymizing and sharing medical text records,” Information Systems Research, vol. 28, no. 2, pp. 332–352, 2017.

[89] L. Kong, M. Bendersky, M. Najork, B. Vargo, and M. Colagrosso, “Learning to cluster documents into workspaces using large scale activity logs,” in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2416–2424.

[90] D. Garat and D. Wonsever, “Automatic curation of court documents: Anonymizing personal data,” Information, vol. 13, no. 1, p. 27, 2022.

[91] X.-B. Li and J. Qin, “Privacy-preserving online face recognition over encrypted outsourced data,” in Proceedings of the 12th International Conference on Information Security Practice and Experience (iSPPR). Springer, 2015, pp. 548–555.

[92] X.-B. Li and J. Qin, “Anonymizing bag-valued sparse data by semantic-similarity-based clustering,” Knowledge and information systems, vol. 35, no. 2, pp. 435–461, 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 The International Conference on Healthcare Science and Engineering, Springer, 2018, pp. 115–128.

[94] L. Ghemri and R. Kannab, “Lexical entailment for privacy protection in medical records.”

[95] B. Chen, C. Wu, and S. Xie, “Generalization via hierarchical clustering for anonymizing set-valued user search logs.”

[96] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, “Utility-preserving privacy protection of nominal data sets via semantic ranking swapping,” Information Fusion, vol. 45, pp. 282–295, 2019.

[97] N. Vuvagar, 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 Personal Communications, pp. 1–16, 2021.

[98] X. Meng, Z. Xu, B. Chen, and Y. Zhang, “Privacy-preserving query log sharing based on prior n-word aggregation,” in 2016 IEEE TrustCom/BigDataSE/SPDA. IEEE, 2016, pp. 722–729.

[99] D. Panies-Estremes, J. Castella-Roca, and J. Garcia-Alfaro, “A real-time query log protection method for web search engines,” IEEE Access, vol. 8, pp. 87 393–87 413, 2020.

[100] M. Ullah, “Obscure logging: A framework to protect and evaluate the web search privacy,” Ph.D. dissertation, CAPITAL UNIVERSITY, 2020.

[101] A. R. S. Nasab and H. Ghaffarian, “A new fast framework for anonymizing iot stream data,” in 2021 5th International Conference on Internet of Things and Applications (IoT). IEEE, 2021, pp. 1–5.

[102] J. Kumar, “Slide window method adapted for privacy-preserving Transactional data streams,” International Journal of Molecular & Cellular Medicine, vol. 8, no. 2, pp. 2528–2538, 2021.

[103] V. Stephanie, M. Chamikara, I. Khalil, and M. Atiquzzaman, “Privacy-preserving location data stream clustering on mobile edge computing and cloud,” Information Systems, p. 101728, 2021.

[104] X. Yang, “Towards utility-aware privacy-preserving sensor data anonymization in distributed iot,” in Proceedings of the 8th ACM International Conference on Internet of Things for Energy-Efficient Buildings, Cities, and Transportation, 2021, pp. 248–249.

[105] R. Patil, P. D. Patil, S. Kanase, N. Bhegade, V. Chavan, and S. Kashetwar, “System for analyzing crime news by mining live data streams with preserving data privacy,” in Sentimental Analysis and Deep Learning. Springer, 2022, pp. 799–811.

[106] J. Tekli, B. A. Bouna, Y. Bou Issa, M. Kamrad, and R. Haraty, “(k,l)-clustering for transactional data streams anonymization,” in International Conference on Information Security Practice and Experience. Springer, 2018, pp. 544–556.

[107] 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 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR). IEEE, 2015, pp. 37–41.

[108] X. Yang, H. Zhu, R. Lu, X. Liu, and H. Li, “Efficient and privacy-preserving online face recognition over encrypted outsourced data,” in 2015 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2018, pp. 366–373.

[109] Z. Zhang, T. Cilloni, C. Walter, and C. Fleming, “Multi-scale, classification, privacy-preserving video,” Electronics, vol. 10, no. 10, p. 1172, 2021.

[110] M.-H. Le, M. S. N. Khan, G. Tsolli, N. Carlsson, and S. Buchegger, “Anonfaces: Anonymizing faces adjusted to constraints on efficacy and
[112] A.-K. Grossfingert, D. Münch, and M. Arens, “An architecture for automatic multimodal video data anonymization to ensure data protection,” in Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III, vol. 11166. International Society for Optics and Photonics, 2019, p. 111660Q.

[113] Z. Ren, Y. J. Lee, and M. S. Ryoo, “Learning to anonymize faces for privacy preserving action detection,” in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 620–636.

[114] P. Deivanai, I. J. V. Nath, and V. Kavitha, “A hybrid data anonymization integrated with suppression for preserving privacy in mining multi party data,” in 2011 International Conference on Recent Trends in Information Technology (ICRITT). IEEE, 2011, pp. 732–736.

[115] S. U. Bazai, J. Jang-Jaccard, and H. Alavizadeh, “A novel hybrid approach for multi-dimensional data anonymization for apache spark,” ACM Transactions on Privacy and Security, vol. 25, no. 1, pp. 1–25, 2021.

[116] D. Mohapatra and M. R. Patra, “Anonymization of attributed social graph using anatomy based clustering,” Multimedia Tools and Applications, vol. 78, no. 18, pp. 25 455–25 486, 2019.

[117] S. B. Basapur, B. Shyala et al., “Attribute assemblability and sensitive attribute frequency based data generalization algorithm for privacy preservation,” in 2021 International Conference on Forensics, Analytics, Big Data Security (FABS), vol. 1. IEEE, 2021, pp. i–xiv.

[118] K. Stokes, “Cover-up: A probabilistic privacy-preserving graph database model,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–8, 2019.

[119] P. Li, F. Zhou, Z. Xu, Y. Li, and J. Xu, “Privacy-preserving graph operations for social network analysis,” in International Symposium on Security and Privacy in Social Networks and Big Data. Springer, 2020, pp. 303–317.

[120] Y. Zhao and I. Wagner, “Using metrics suites to improve the measurement of privacy in graphs,” IEEE Transactions on Dependable and Secure Computing, 2020.

[121] I. Wagner and I. Yevseyeva, “Designing strong privacy metrics suites using evolutionary optimization,” ACM Transactions on Privacy and Security (TOPS), vol. 24, no. 2, pp. 1–35, 2021.

[122] T. Ma, Y. Zhang, J. Cao, J. Shen, M. Tang, Y. Tian, A. Al-Dhelaan, and M. Al-Rodhaan, “Kdvm: a k-degree anonymity with vertex and edge modification algorithm,” Computing, vol. 97, no. 12, pp. 1165–1184, 2015.

[123] Y. Zhao and I. Wagner, “Using metrics suites to improve the measurement of privacy in graphs,” IEEE Transactions on Dependable and Secure Computing, 2020.

[124] H. Xia and W. Yang, “Information entropy models and privacy metrics methods for privacy protection,” International Journal of Network Security & Its Applications, vol. 24, no. 1, pp. 101–112, 2012.

[125] S. Ji, P. Mittal, and R. Beyah, “Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey,” IEEE Communications Surveys & Tutorials, vol. 19, no. 2, pp. 1305–1326, 2016.

[126] W. Huoxiang, S. Smys et al., “Big data analysis and perturbation using data mining algorithm,” Journal of Soft Computing Paradigm (JSCP), vol. 3, no. 01, pp. 19–28, 2021.

[127] A. Cuzzocrea and E. Damiani, “Privacy-preserving big data exchange: Models, issues, future research directions,” in 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021, pp. 5081–5084.

[128] S. Madan, K. Bhardwaj, and S. Gupta, “Critical analysis of big data privacy preservation techniques and challenges,” in International Conference on Innovative Computing and Communications. Springer, 2022, pp. 267–278.

[129] D. Xiang and W. Cai, “Privacy protection and secondary use of health data: Strategies and methods,” BioMed Research International, vol. 2021, 2021.

[130] K. Guo and Q. Zhang, “Fast clustering-based anonymization approaches with time constraints for data streams,” Knowledge-Based Systems, vol. 46, pp. 95–108, 2013.

[131] X. Qian, X. Li, and Z. Zhou, “An efficient privacy-preserving approach for data publishing,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–17, 2021.

[132] U. Sopaoglu and O. Abul, “Classification utility aware data stream anonymization,” Applied Soft Computing, vol. 110, p. 107743, 2021.

[133] M. Milani, Y. Huang, and F. Chiang, “Data anonymization with diversity constraints,” IEEE Transactions on Knowledge and Data Engineering, 2021.

[134] P. Parameswarappa, “Clustering approaches for anonymizing high-dimensional sequential activity data,” Ph.D. dissertation, University of Maryland, Baltimore County, 2020.

[135] U. Ahmed, G. Srivastava, and J. C.-W. Lin, “A machine learning model for data sanitization,” Computer Networks, vol. 189, p. 107914, 2021.

[136] U. Ahmed, J. C.-W. Lin, P. Fournier-Viger, and C.-F. Cheng, “Privacy-preserving periodic frequent pattern model in aiot applications,” in 2021 International Symposium on Wireless and Communication Systems (ISWCS). IEEE, 2021, pp. 1–2.

[137] U. Ahmed, J. C.-W. Lin, and P. Fournier-Viger, “Privacy preservation of periodic frequent patterns using sensitive inverse frequency,” in Periodic Pattern Mining. Springer, 2021, pp. 215–227.

[138] Y. Zhao and I. Wagner, “Using metrics suites to improve the measurement of privacy in graphs,” IEEE Transactions on Dependable and Secure Computing, 2020.

[139] T. Henderson, “Short paper: Integrating the data protection impact assessment into the software development lifecycle,” in Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2020 International Workshops, DPM 2020 and CBT 2020, Guildford, UK, September 17–18, 2020, Revised Selected Papers, vol. 12484. Springer Nature, 2020, p. 219.

[140] G. Logeswari, D. Sangeetha, and V. Vaidhehi, “A cost effective clustering based anonymization approach for storing phone’s in cloud,” in 2014 International conference on recent trends in information technology. IEEE, 2014, pp. 1–5.

[141] J. Usha Lawrance and J. V. Nayai Jesudhasan, “Privacy preserving parallel clustering based anonymization for big data using mapreduce framework,” Applied Artificial Intelligence, pp. 1–34, 2021.

[142] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, and J. Chen, “Privacy-aware local-recoding anonymization for scalable big data privacy preservation in cloud,” IEEE Transactions on computers, vol. 64, no. 8, pp. 2293–2307, 2014.

[143] J. V. Nayai and V. Kavitha, “Privacy and utility preserving data clustering for data anonymization and distribution on hadoop,” Future Generation Computer Systems, vol. 74, pp. 393–408, 2017.

[144] N. Singh and A. K. Singh, “Data privacy protection mechanisms in cloud,” Data Science and Engineering, vol. 3, no. 1, pp. 24–39, 2018.

[145] P. Leksamy and M. A. Rahiman, “A sanitization approach for privacy preserving data mining on social distributed environment,” Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 7, pp. 2761–2777, 2020.

[146] J. Jayaraman and A. Stanislaus Panneerselvam, “A novel privacy preserving digital forensic readiness provable data possession technique for health care data in cloud,” Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 5, p. 1942, 2021.

[147] S. Madan and P. Goswami, “Hybrid privacy preservation model for big data publishing on cloud,” International Journal of Advanced Intelligent Paradigms, vol. 20, no. 3–4, pp. 343–355, 2021.

[148] S. Madan and G. Puneet, “Adaptive privacy preservation approach for big data publishing in cloud using k-anonymization,” Recent Advances in Computer Science and Communications (Formerly; Recent Patents on Computer Science), vol. 14, no. 8, pp. 2678–2688, 2021.

[149] E. Shannumgapiya and R. Kavitha, “Efficient and secure privacy analysis for medical big data using tdes and mksvm with access control cloud,” Journal of Medical Systems, vol. 43, no. 8, pp. 1–12, 2019.

[150] O. Abul, F. Bonchi, and M. Nanni, “Anonymization of moving objects databases by clustering and perturbation,” Information systems, vol. 35, no. 8, pp. 884–910, 2010.

[151] F. Fei, S. Li, H. Dai, C. Hu, W. Dou, and Q. Ni, “A k-anonymity based schema for location privacy preservation,” IEEE Transactions on Sustainable Computing, vol. 4, no. 2, pp. 156–167, 2017.

[152] J. Lee, S. Kim, Y. Cho, Y. Chung, and Y. Park, “A hierarchical clustering-based spatial cloaking algorithm for location-based services,” Journal of Internet Technology, vol. 13, no. 4, pp. 645–653, 2012.

[153] K. Niu, C. Peng, Y. Tian, and W. Tan, “K-implicit tracking data publishing scheme against geo-matching attacks,” Journal of Information Science & Engineering, vol. 38, no. 1, 2022.

[154] L. Yao, G. Wu, J. Wang, F. Xia, C. Lin, and G. Wang, “A clustering k-anonymity scheme for location privacy preservation,” IEICE Transactions on information and systems, vol. 95, no. 1, pp. 134–142, 2012.
S. Patil, S. Joshi, and D. Patil, “Enhanced privacy preservation using anonymization in iot-enabled smart homes,” in Smart Intelligent Computing and Applications. Springer, 2020, pp. 439–454.

F. Ullah, I. Ullah, A. Khan, M. I. Edin, I. Alamy, and W. Alosaimi, “Enabling clustering for privacy-aware data dissemination based on medical-healthcare-iot (mh-iot) for wireless body area network,” Journal of Healthcare Engineering, vol. 2020, 2020.

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 of Things Journal, 2022.

J. Liu, C. Zhang, K. You, and Y. Fan, “Privacy preservation in multi-cloud secure data fusion for infectious-disease analysis,” IEEE Transactions on Mobile Computing, 2022.

C. Zhang, H. Jiang, Y. Wang, Q. Hu, J. Yu, and X. Cheng, “User identity de-anonymization based on attributes,” in International Conference on Wireless Algorithms, Systems, and Applications. Springer, 2019, pp. 458–469.

C. Zhang, S. Wu, H. Jiang, Y. Wang, J. Yu, and X. Cheng, “Attribute-enhanced de-anonymization of online social networks,” in International Conference on Computational Data and Social Networks. Springer, 2019, pp. 256–267.

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 Transactions on Computational Social Systems, 2021.

Y. Shao, J. Liu, S. Shi, Y. Zhang, and B. Cui, “Fast de-anonymization of social networks with structural information,” Data Science and Engineering, vol. 4, no. 1, pp. 76–92, 2019.

S. Gambis, M.-O. Killijian, and M. N. del Prado Cortez, “De-anonymization attack on geolocated data,” Journal of Computer and System Sciences, vol. 80, no. 8, pp. 1597–1614, 2014.

C.-F. Chiasserini, L. Garetto, and E. Michele, “Impact of clustering on the performance of network de-anonymization,” in Proceedings of the 2015 Network on Computational Social Networks, 2015, pp. 83–94.

C.-F. Chiasserini, M. Garetto, and E. Leonardi, “Social network de-anonymization under scale-free user relations,” IEEE/ACM transactions on networking, vol. 24, no. 6, pp. 3756–3769, 2016.

C. Carla-Fabiana, M. Garetto, and E. Leonardi, “De-anonymizing clustered social networks by percolation graph matching,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 12, no. 2, pp. 1–39, 2018.

L. Fu, X. Fu, Z. Hu, Z. Xu, and X. Wang, “De-anonymization of social networks with communities: When quantifications meet algorithms,” arXiv preprint arXiv:1703.09028, 2017.

L. Fu, J. Zhang, S. Wang, X. Wu, X. Wang, and G. Chen, “De-anonymizing social networks with overlapping community structure,” IEEE/ACM Transactions on Networking, vol. 28, no. 1, pp. 360–375, 2020.

M. Francia, E. Gallinucci, M. Gofarelli, and N. Santolini, “Dart: De-anonymization of personal gazetteers through social trajectories,” Journal of Information Security and Applications, vol. 55, p. 102634, 2020.

T. Orekondy, S. J. Oh, Y. Zhang, B. Schiele, and M. Fritz, “Gradient-leaks: Understanding and controlling de-anonymization in federated learning,” arXiv preprint arXiv:1805.05838, 2018.

Z. Chen, Y. Fu, M. Zhang, Z. Zhang, and H. Li, “The de-anonymization method based on user spatio-temporal mobility trace,” in International Conference on Information and Communications Security. Springer, 2017, pp. 459–471.

T. Murakami, A. Kanemura, and H. Hino, “Group sparsity tensor factorization for de-anonymization of mobility traces,” in 2015 IEEE TrustCom/BigDataSE/SP/A, vol. 1. IEEE, 2015, pp. 621–629.

H. Li, Q. Chen, H. Zhu, D. Ma, H. Wen, and X. S. Shen, “Privacy leakage via de-anonymization and aggregation in heterogeneous social networks,” IEEE Transactions on Dependable and Secure Computing, vol. 17, no. 2, pp. 350–362, 2017.

C. Zhen-Yu, Z. Min, F. Yan-Yan, Z. Zhen-Feng, and L. Hao, “A user de-anonymization attack method for trajectory data publishing,” Journal of Information Security Research, vol. 3, no. 10, p. 0, 2017.

H. Wang, C. Gao, Y. Li, Z.-L. Zhang, and D. Jin, “Revealing physical world privacy leakage by cyberspace cookie logs,” IEEE Transactions on Network and Service Management, vol. 17, no. 4, pp. 2550–2566, 2020.

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,” Information Sciences, vol. 394, pp. 53–67, 2017.
S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, Z. Sun, L. Yin, C. Li, W. Zhang, A. Li, and Z. Tian, “The qos and privacy J. Park and K. Kim, “Image perturbation-based deep learning for face G. Newlands, C. Lutz, A. Tantí-Larrieux, E. F. Villaronga, R. Haras- N. Ahmad and P. Chauhan, “State of data privacy during covid-19,” M. S. Jere, T. Farnan, and F. Koushanfar, “A taxonomy of attacks on Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, “Federated J. Liu, C. Zhang, K. Xue, and Y. Fang, “Privacy preservation in multi- X. Xian, T. Wu, Y. Liu, W. Wang, C. Wang, G. Xu, and Y. Xiao, “Towards J. Li and X. Liao, “Security and privacy in new computing environments M. Li, N. Varble, B. Turkbey, S. Xu, and B. J. Wood, “Camera-based dis- P. K. Rao and D. Rawtani, “Modern digital techniques for monitoring M. D. Jena, S. S. Singhar, B. K. Mohanta, and S. Ramasubbareddy, S. Aanjankumar and S. Poonkuntran, “An efficient soft computing ap- M. Milani, Y. Huang, and F. Chiang, “Data anonymization with diversity constraints,” IEEE Transactions on Knowledge and Data Engineering, M. D. Jena, S. S. Singhar, B. K. Mohanta, and S. Ramasubbareddy, Ensuring data privacy using machine learning for responsible data science,” in Intelligent Data Engineering and Analytics. Springer, 2021, pp. L. Arbuckle and K. El Emam, Building an anonymization pipeline: Creating safe data. O’Reilly Media, 2020. S. Aanjankumar and S. Poonskuntran, “An efficient soft computing approach for securing information over gameove zeus botnets with modified cpa algorithm,” Soft Computing, vol. 24, no. 21, pp. 16499–16507, 2020. 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,” Computers & Security, vol. 113, pp. 102529, 2022. R. Sanchez-Borra and A. Skarmeta, “Securing the hyperconnected healthcare ecosystem,” in AI and IoT for Sustainable Development in Emerging Countries. Springer, 2022, pp. 455–471. J. B. Avotunde, R. G. Jimoh, R. O. Ogundokun, S. Misra, and O. C. Ahikuye, “Big data analytics of iot-based cloud system framework: Smart healthcare monitoring systems,” in Artificial Intelligence for Cloud and Edge Computing. Springer, 2022, pp. 181–208. 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 Transactions on Services Computing, 2022. K. M. Akram, S. Sihem, K. Okba, and S. Harous, “IoT-mfg-cloud-based architecture for covid-19 detection,” Biomedical Signal Processing and Control, vol. 103715, 2022. M. L. Alarcon, R. Oruche, A. Pandey, and P. Calyam, “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. Elsevier, 2022, pp. 159–180. P. K. Rao and D. Rawtani, “Modern digital techniques for monitoring and analysis,” in COVID-19 in the Environment. Elsevier, 2022, pp. 130–130. M. Li, N. Varble, B. turkby, S. Xu, and B. J. Wood, “Camera-based dis- tance detection and contact tracing to monitor potential spread of covid-19,” in Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment, vol. 12035. SPIE, 2022, pp. 329–335.
[326] X. Zhang, J. Hamm, M. K. Reiter, and Y. Zhang, “Defeating traffic analysis via differential privacy: a case study on streaming traffic,” International Journal of Information Security, pp. 1–18, 2022.

[327] H. Tang, J. Chen, Y. Zhou, and L. Chen, “A novel resource management scheme for virtualized cyber–physical–social system,” Physical Communication, vol. 50, p. 101513, 2022.

[328] M. Snehi and A. Bhandari, “A novel distributed stack ensemble meta-learning-based optimized classification framework for real-time prolific iot traffic streams,” Arabian Journal for Science and Engineering, pp. 1–24, 2022.

[329] S. Xiong, A. D. Sarwate, and N. B. Mandayam, “Network traffic shaping for enhancing privacy in iot systems,” IEEE/ACM Transactions on Networking, 2022.

[330] 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 Computing, p. 100478, 2022.

[331] C. Karras, A. Karras, and S. Sioutas, “Pattern recognition and event detection on iot data-streams,” arXiv preprint arXiv:2203.01114, 2022.

[332] F. Chen, W. Wang, H. Yang, W. Pei, and G. Lu, “Multiscale feature fusion for surveillance video diagnosis,” Knowledge-Based Systems, p. 108103, 2023.

[333] G. Aguiar, B. Krawczyk, and A. Cano, “A survey on learning from imbalanced data streams: taxonomy, challenges, empirical study, and reproducible experimental framework,” arXiv preprint arXiv:2204.03719, 2022.

[334] D.-H. Vu, “Privacy-preserving naive bayes classification in semi-full distributed data model,” Computers & Security, vol. 115, p. 102630, 2022.

[335] M. Lango and J. Stefanowski, “What makes multi-class imbalanced problems difficult? an experimental study,” Expert Systems with Applications, p. 116962, 2022.

[336] T. Li, H. Wang, D. He, and J. Yu, “Blockchain-based privacy-preserving and rewarding private data sharing for iot,” IEEE Internet of Things Journal, 2022.

[337] H. Wen, M. Wei, D. Du, and X. Yin, “A blockchain-based privacy preservation scheme in mobile medical,” Security and Communication Networks, vol. 2022, 2022.

[338] M. Sarrab and F. Alishohoumi, “Assisted fog computing approach for data privacy preservation in iot-based healthcare,” in Security and Privacy Preserving for IoT and 5G Networks. Springer, 2022, pp. 191–201.

[339] D. Peng, L. Sun, R. Zhou, and Y. Wang, “Study qos-aware fog computing for disease diagnosis and prognosis,” Mobile Networks and Applications, pp. 1–8, 2022.

[340] 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 Transactions on Industrial Informatics, 2022.

[341] J. Wang, “Workflow offloading with privacy preservation in a cloud-edge environment,” Concurrency and Computation: Practice and Experience, p. e7002, 2022.

[342] D. Wang, S. Shi, Y. Zhu, and Z. Han, “Federated analytics: Opportunities and challenges,” IEEE Network, 2021.

[343] S. Flowerday and C. Xenakis, “Security and privacy in distributed healthcare environments,” Methods of Information in Medicine, no. AAM, 2022.

[344] L. Sankar, “Lalitha sankar, arizona state university: Federated analytics based contact tracing for covid-19,” 2020.

[345] A. Kallel, M. Rekik, and M. Khemakhem, “Hybrid-based framework for covid-19 prediction via federated machine learning models,” The Journal of Supercomputing, pp. 1–28, 2021.

[346] F. Kserawi, S. Al-Marri, and Q. Malluhi, “Privacy-preserving fog aggregation of smart grid data using dynamic differentially-private data perturbation,” IEEE Access, 2022.

ABDUL MAJEED received the B.Sc. degree in Information Technology from the UUIT, PMAS-UAAR, Rawalpindi, Pakistan, in 2013, the M.Sc. degree in Information Security from the COMSATS University Islamabad, Pakistan, in 2016, and the Ph.D. degree in Computer Information Systems & Networks from the Korea Aerospace University, Korea, in 2021. He worked as a Security Analyst with Trillum Information Security Systems (TISS), Rawalpindi, Pakistan, from 2015 to 2016. He is currently working as an Assistant Professor with the Department of Computer Engineering, Gachon University, Korea. His research interests include privacy preserving data publishing, statistical disclosure control, privacy-aware analytics, and machine learning. Contact him at ab09@gachon.ac.kr.

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 Research and Development Department, National Radio and Telecommunication Corporation, Haripur, Pakistan, for two years. His research interests include efficient hardware implementations of cryptographic protocols, blockchain, and network security. Contact him at safi@gachon.ac.kr.

SEONG OUN HWANG 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, in 2004, South Korea. He 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, Korea. His research interests include cryptography, cybersecurity, and artificial intelligence. Contact him at sohwang@gachon.ac.kr.

---

***