How to Cite:
Fernandez, T. F., & Kumar, E. D. (2022). A survey of privacy-preserving mechanisms for heterogeneous data types. *International Journal of Health Sciences*, *6*(S4), 5692–5700. https://doi.org/10.53730/ijhs.v6nS4.9409

A survey of privacy-preserving mechanisms for heterogeneous data types

DR. Terrance Frederick Fernandez
Associate Professor, Department of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology

MR. E. Dhilip Kumar
Assistant Professor, Department of Computer Applications, Dhanalakshmi Srinivasan College of Engineering and Technology

***Abstract*---**Due to the pervasiveness of always connected devices, large amounts of heterogeneous data are continuously being collected. Beyond the benefits that accrue for the users, there are private and sensitive information that is exposed. Therefore, Privacy-Preserving Mechanisms (PPMs) are crucial to protect users' privacy. In this paper, we perform a thorough study of the state of the art on the following topics: heterogeneous data types, PPMs, and tools for privacy protection. Building from the achieved knowledge, we propose a privacy taxonomy that establishes a relation between different types of data and suitable PPMs for the characteristics of those data types. Moreover, we perform a systematic analysis of solutions for privacy protection, by presenting and comparing privacy tools. From the performed analysis, we identify open challenges and future directions, namely, in the development of novel PPMs.

***Keywords*---**privacy, privacy taxonomy, privacy-preserving mechanisms, heterogeneous data types, privacy tools.

Introduction

Data is continuously being collected due to the pervasiveness of always connected devices and the iniquitousness of Internet of Things (IoT) technologies in people's lives. IoT provides the interconnection between multiple heterogeneous devices and sensors that are able to monitor and gather all types of data about machines and human social life . Despite the benefits that can come from collecting data, users are exposing sensitive and private information with possibly untrustworthy entities. These entities can process, analyze and mine data in order to extract useful information, but also sell and/or share the collected data with third parties, using it maliciously. With the growing number of misuse of data and data

breaches , privacy has been an emergent topic and serious privacy concerns have been aroused. To address these issues, numerous Privacy-Preserving Mechanisms (PPMs) and tools have been proposed.

Although PPMs aim to preserve users' privacy, this can come at the expense of a degraded utility of data . Therefore, the selection of a PPM should take into account not only the users' objective but also the trade-off between the privacy level and the utility of data, which are many times application-specific. Considering the heterogeneity of the collected data, selecting and configuring the proper PPM is quite challenging. To automates this process and to give a logical and systematic structure of the main components and concepts of privacy, several tools were developed . These tools were proposed to facilitate the configuration of PPMs and the analysis of results. However, selecting the proper PPM according to the characteristics of the data remains as a challenge.

To better understand how to identify PPMs according to the data characteristics, this survey presents an up-to-date and thorough review on heterogeneous data types and applicable PPMs. In recent years, several general surveys have focused on PPMs for data mining and how they can be compared in terms of achieved privacy level, data utility, complexity, and/or application fields. Other more specific surveys discuss PPMs for a specific data type or a restrict group of data types as well application of PPMs for specific domains. Our survey differs from previous literature by proposing a privacy taxonomy for heterogeneous data types that establishes a relation between different data types and PPMs. In this survey, PPMs are classified according to the overall categories of data they can be applied to (structured, semi-structured and unstructured), as well as their suitability for real-time or offline application. The main contribution of this survey is the specification of a taxonomy of data types for each category of data that is amenable for the identification of corresponding PPMs, so as to allow the reader to properly understand the underlying principles of the addressed PPMs and their applicability to the data types in the taxonomy within. This survey further contributes by presenting and comparing existing privacy tools with respect to the data types and PPMs made available, as well the privacy and utility evaluation features of such tools.

The remainder of the survey is structured as follows. Section provides a study and classification of heterogeneous data types. Section presents the state-of-the-art PPMs. Section proposes a privacy taxonomy for heterogeneous data types. Section provides an overview of existing tools for privacy protection. Section presents open challenges and future directions. Finally, Section concludes the survey paper.

Heterogeneous data types

Everyday, various devices and services collect large amounts of heterogeneous data with different purposes. Although the collection purpose may vary, collected data may have similar characteristics. In the domain of IoT, considerable amounts of data are continuously collected by different sensors. According to, the top ten IoT sensors includes: temperature sensors, humidity sensors, pressure sensors, proximity sensors, level sensors, accelerometers, gyroscope, gas sensors,

infrared sensors, and optical sensors. From these sensors, several services are provided and different data types are collected. This section gives an overview of existing types of data.

Commonly, data is classified according to its structure, that is, how the data is organized. From this classification, we have structured data, semi-structured data and unstructured data. Structured data corresponds to data often stored in tables, such as relational databases or spreadsheets. Following the structure imposed by the database, we may have data types such as numbers, strings, Booleans, dates, and others. Structured data is divided in categorical data, that is, data types that can be divided into groups, and numerical data, that corresponds to data types represented by numeric values of specific variables. Categorical data is subdivided in nominal, which represents a set of possible values, and ordinal, which also represents a set of values but with a rank order. In its turn, numerical data is subdivided in interval and ratio, which represent variables that can be measured with an interval scale (e.g. Celsius scale) or a ratio scale (e.g. Kelvin temperature scale), respectively. Unstructured data consists in data that does not have a predefined data model or a specified organization. Examples of unstructured data are images, videos, streaming sensor data, and text documents. Within unstructured data, we may also have dates, numbers or facts. Semi-structured data is a type of structured data that does not have a rigid structure imposed by a data model. For example, emails are constituted by structured information (e.g. sender, recipient) and unstructured data that corresponds to the email message content and/or attachments. Semi-structured data are often represented as graphs, XML and other markup languages.

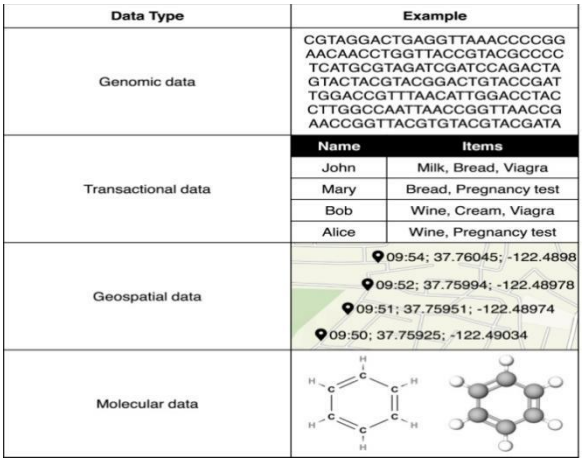


Fig. 1. Examples of data types

Beyond the aforementioned, unstructured data can be further divided in several time series data, streaming data, sequence data, multimedia data, and spatial data. While time series data consists in sequences of values/events repeatedly collected over time (e.g. stock market data), sequence data corresponds to sequences of ordered values/events that are recorded with or without a certain timestamp (e.g. genomic data). Streaming data consists in data continuously arriving (e.g. sensor data). Multimedia data includes data such as images, videos

or audios. The last category is spatial data that corresponds to space-related data, such as maps. Although the terms spatial and geospatial data are often used as equivalents, geospatial data corresponds to a type of spatial data that is related to Earth and that contains geographic components, such as location coordinates. Finally, textual and transactional data can also be unstructured data types, whereby textual data refers to unstructured text (e.g. documents) and transactional data, a canonical example of set-valued data, corresponds to data in which each record contains a set of arbitrary items (e.g. online shopping, see).

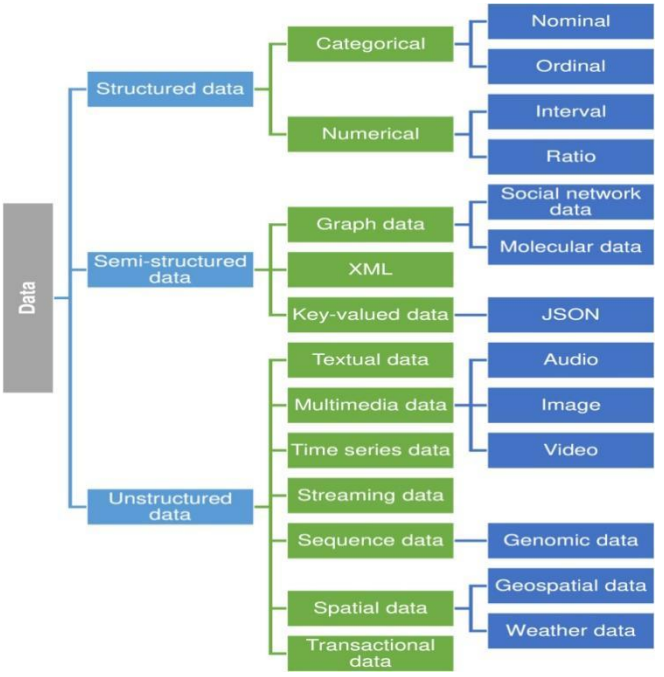
Datasets can also be divided in three categories: record data, graph-based data, and ordered data. Record data is usually stored in relational databases or flat files and each record is described with the same set of attributes. Graph-based data is typically used to represent data objects that can be mapped as nodes of a graph, while their relationship is mapped as a link (e.g. social network data and molecules). The ordered data category pertains data that is ordered in time or space, such as, sequential data, time series data, or spatial data.

A relevant matter for processing heterogeneous data types is the amount of data to be considered. The integration and analysis of heterogeneous data types is quite challenging, specially, due to the increase of data collection, that results in big data issues. Rob Thomas, general manager for IBM Analytics, defined big data as “diverse datasets that include structured, semi-structured and unstructured data, from different sources and in different volumes, from terabytes to zettabytes. It is about datasets so large and diverse that it is difficult, if not impossible, for traditional relational databases to capture, manage, and process them with low-latency”. To deal with the processing, integration, and analysis of heterogeneous data and big data, some methods have been developed and presented in.

The focus of this survey is on heterogeneous data types and corresponding PPMs. Big data aspects have been the subject of other surveys, where, for example, a well-defined taxonomy is presented according to six dimensions: data, compute infrastructure, storage infrastructure, analytics, visualization, and security and privacy. In the dimension of data, the authors divided data according to different characteristics, such as the structure of data, as mentioned before. Similarly, the survey presents a rich taxonomy of big data on the following domains: semantic, compute infrastructure, storage system, big data management, data mining and machine learning, and security and privacy. With respect to the semantic of big data, the authors consider diverse characteristics, such as volume, velocity, variety, and others. Within variety, there is a data classification that also divides data according to its structure (i.e. structured, semi-structured, and unstructured data). In the data taxonomy proposed by , big data was presented as a category that was likewise divided according to the data structure and included streaming data as a subcategory.

To summarize this section, Fig. 2 presents a data taxonomy according to the structure of data, where data is first divided into structured, semi-structured and unstructured. Within each category, structured data is divided into categorical and numerical data, semi-structured data is divided into graph data, XML, and key-valued data, and unstructured data is divided into textual, multimedia, time

series, streaming, sequence, spatial, and transactional data. This data taxonomy will be instrumental so as to identify PPMs suitable to the identified data categories, as we will now address.



Privacy-preserving mechanisms

This section gives an overview of existing PPMs over different domains. Before presenting the PPMs, some concepts are briefly presented as background knowledge. PPMs are applied to protect user's sensitive and private information. In general, we consider a sensitive attribute (SA) when we have user-specific private data that can be shared for research/statistical analysis purposes, but should not be linkable to the individual user. A quasi-identifier (QID) consists in a non-sensitive attribute (or a set of attributes) that can be combined or linked with external/background information to re-identify the individual to whom data refers. Finally, a key attribute consists in a explicit/uniquely identifier (ID) of an individual, or in other words, personally identifiable information (PII).

To preserve users' privacy, PPMs often apply one or a combination of data sanitizing operations, such as generalization, suppression, perturbation, anatomization, permutation and/or slicing. The sanitization goal is to protect sensitive information by removing or modifying attributes of data. Generalization corresponds to the replacement of a value with a broader one. For instance, the replacement of numerical data with intervals and the definition of a hierarchy for categorical attributes. Suppression consists in removing some values of an attribute to prevent the disclosure of information. Typically, this operation is used in tables by removing all values of an attribute in a column or by removing an entry row. Perturbation corresponds to the replacement of the original data with values with identical statistical information. This operation is

commonly achieved with the addition of noise. Anatomization consists in the de-association of quasi-identifiers (QIDs) and sensitive attributes (SAs) in two separated tables in order to prevent the linkage of QIDs to SAs. Permutation corresponds to the rearrangement of values after their partitioning into group of values. Although this operation alone is not suitable for real-world data, it is often combined with slicing. Slicing partitions the data both vertically and horizontally, which makes this technique able to handle high-dimensional data and data without a clear separation between QIDs and SAs. Briefly, vertical partitioning consists in having each attribute or subset of attributes contained in each column and horizontal partitioning consists in randomly permute the values within columns, thus breaking the linkage among different columns. Since the presented list of sanitization operations is not an extensive list, please refer to for a more thorough analysis.

ID	QID			SA
Name	Age	Sex	Zip Code	Disease
Mary	25	F	46909	Hepatitis C
John	25	M	46909	HIV
Bob	35	M	46900	HIV
Ian	54	M	46900	Gastritis
Alice	57	F	46761	HIV
Helen	60	F	46761	Hepatitis C
Cindy	60	F	46760	Hepatitis C
Alex	63	M	46760	Diabetes

a.OriginalTable

ID	QID			SA
Name	Age	Sex	Zip Code	Disease
*	[20-55]	*	4690*	Hepatitis C
*	[20-55]	*	4690*	HIV
*	[20-55]	*	4690*	HIV
*	[20-55]	*	4690*	Gastritis
*	[56-65]	*	4676*	HIV
*	[56-65]	*	4676*	Hepatitis C
*	[56-65]	*	4676*	Hepatitis C
*	[56-65]	*	4676*	Diabetes

b.Anonymized Table

Privacy taxonomy for heterogeneous data types

This section starts by providing a literature review of existing data privacy taxonomies, that is, taxonomies that take into consideration privacy aspects, and ends with the proposal of a novel privacy taxonomy for heterogeneous data types, presenting PPMs that fit to the characteristics of different data types. Moreover, the PPMs are classified according to their application mode in real world, i.e. weather they are suitable for real-time or offline application.

To better understand data privacy, Barker et al. created a taxonomy based on a 3D graph. This graph contains three contributors of data privacy: visibility, granularity, and purpose. Each one of these categories has specific values. For instance, visibility means if the data is visible to all world, third party, house, owner or none. This taxonomy allows us to select the privacy-preserving mechanism according to the values of the categories. Although the authors present a table of the privacy taxonomy with mechanisms from the literature, they only present this analysis for three mechanisms, exclusively according to the axes of the 3D graph, and lacks important PPMs proposed since its publication in 2009.

Sharma et al. presents a comparative study of privacy-preserving techniques. The privacy-preserving techniques are compared according to different characteristics, such as the dataset type, the data type, the information loss, and others. However, the provided comparison considers techniques instead of specific examples of PPMs. Moreover, regarding data types, the authors compare techniques only according to the following three data types: numerical, categorical, and Boolean data. On the other hand, Puri et al. only focus on relational and transaction data. For these data types, the authors present existing techniques that ensure privacy while publishing data and, in particular, they present a case study concerning algorithms to anonymize patient data.

Due to the diversity of privacy techniques, Kanwal et al. presents a comparison between different techniques considering their merits, demerits, and their data applications. The authors present a data taxonomy and possible techniques for structured data, semi-structured data, unstructured data, and big data. However, the presented analysis is mainly focused on privacy techniques that can be applied in e-health. In addition, the analysis is performed according to the privacy techniques (e.g. suppression and generalization) and, then, in which PPMs are those techniques applied.

Data privacy is also considered in the analysis of big data, where the main challenge of applying privacy models is the computational cost. Both present taxonomies of big data according to different domains and include security and privacy as one of the aspects. In the domain of security and privacy, the work discusses some existing issues and possible solutions for the following five types of data: streaming data, graph data, scientific, web, retail and financial data. However, the presented solutions are related to both security and privacy issues and, in some cases, correspond only to recommendations/best practices and not to PPMs. With respect to data privacy, the survey only mentions existing mechanisms to preserve privacy without specifying how those mechanisms work or for what types of data they are suitable.

Since the realm of big data contains structured and unstructured data, finding the suitable PPM remains as an open issue. Although there are mechanisms for structured data, extracting the sensitive information from unstructured data is not trivial. In the domain of big data, this is harder due to the amount of data and the associated computational cost. Victor et al. provide a survey on privacy models for big data. In particular, several privacy models are studied, starting with the traditional mechanisms and, then, presenting mechanisms that can be

extended for big data. In contrast to our focus that is identifying PPMs according to the data characteristics, the goal of the authors of consists in distinguishing which big data issue is addressed by the mechanisms.

While some existing taxonomies focus on privacy aspects, heterogeneous data types might share common aspects. This results in a challenge when choosing an efficient PPM for each specific and heterogeneous data type. In this paper, we propose a privacy taxonomy that maps data types and their common characteristics with appropriate PPMs, thus serving as a guideline to assess which PPMs are available for specific data types and their underlying characteristics.

Privacy tools

This section covers existing privacy tools, namely, their objectives and implementation details. Beyond anonymizing data, some tools allow the assessment of different configurations of PPMs, which in turn enables the evaluation of the achieved privacy and utility level.

ARX Data Anonymization Tool is an open source tool for anonymizing sensitive personal data. This tool enables users to import data, configure, explore, analyze, and export data. In each step, the user is able to define a privacy model, to filter and analyze the solution space, and to evaluate the utility of the data. ARX is a complete tool that imports structured data only. This tool is written in Java and provides an API. Regarding the privacy models, it has already some mechanisms implemented, namely: syntactic privacy models (such as *k*-anonymity, *l*-diversity, *t*-closeness, and many others), statistical privacy models (e.g. population uniqueness), and semantic privacy models (e.g. differential privacy). ARX does not implement any attack/adversary model, but features the implementation of models for assessment of the risk of re-identification.

sdcTools consist in tools to provide Statistical Disclosure Control (SDC). The ARGUS software was developed in order to have a free software solution that guarantees the SDC. *Aircloak* is a privacy-preserving solution that uses a unique and patented data anonymization method. *Aircloak* does not modify the database and supports all data types including unstructured text. *Aircloak*'s anonymization is based on existing techniques such as *k*-anonymity, low-count, suppression, top and bottom coding, differential privacy noise, and other patented open concepts. From these techniques, *Aircloak* provides a dynamic anonymization approach that consists in adding noise. Finally, *Aircloak* has a free-to-use version for universities and a full version for enterprises.

Conclusion

Due to the iniquitousness of smart devices, there are large amounts of data continuously being collected by possibly untrustworthy entities, which raises several privacy concerns. Privacy-Preserving Mechanisms (PPMs) have been proposed to address this challenge and to protect users' privacy. However, due to the heterogeneity of the data and the lack of generic PPMs, selecting the proper mechanism remains a challenge. This survey identifies and classifies existing

heterogeneous data types and presents the state-of-the-art PPMs according to their purpose. With this knowledge, we propose a novel privacy taxonomy that establishes a relation between PPMs and data types. Specifically, the proposed taxonomy differentiates which PPMs are applicable for the characteristics of each data type. Additionally, it distinguishes whether the PPMs are applicable in real-time or offline. Finally, this survey presents and compares tools for privacy protection. The performed analysis allows us to conclude about the need of novel PPMs for heterogeneous data types and a unified tool that implements PPMs for different types of data, as well as techniques for privacy evaluation, including methodologies for re-identification risk assessment, complemented with practical re-identification attacks.

References

1. Yan Z., Zhang P., Vasilakos A.V.A survey on trust management for Internet of Things J. Netw. Comput. Appl., 42 (2014), pp. 120-134.
2. Clement J. Online privacy in the United States - statistics & facts (2020).
3. Aldeen Y.A.A.S., Salleh M., Razzaque MA. A comprehensive review on privacy preserving data mining Springer Plus, 4 (1) (2015), p. 694.
4. Shah A., Gulati R. Privacy preserving data mining: Techniques classification and implications—A survey Int. J. Comput. Appl., 137 (12) (2016), pp. 40-46.
5. Mendes R., Vilela J.P. Privacy-preserving data mining: Methods, metrics, and applications IEEE Access, 5 (2017), pp. 10562-10582.
6. Shokri R., Theodorakopoulos G., Troncoso C., Hubaux J.-P., Le Boudec J.-Y. Protecting location privacy: Optimal strategy against localization attacks Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12, Association for Computing Machinery, New York, NY, USA (2012), pp. 617-627, 10.1145/2382196.2382261.
7. Prasser F., Kohlmayer F. Putting statistical disclosure control into practice: The arx data anonymization tool Medical Data Privacy Handbook, Springer International Publishing, Cham (2015), pp. 111-148, 10.1007/978-3-319-23633-9_6.
8. Susilo, C. B., Jayanto, I., & Kusumawaty, I. (2021). Understanding digital technology trends in healthcare and preventive strategy. *International Journal of Health & Medical Sciences*, *4*(3), 347-354. https://doi.org/10.31295/ijhms.v4n3.1769
9. Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.
10. Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.