Received: 19 April 2020 | Revised: 31 July 2020 | Accepted: 27 August 2020

DOI: 10.1002/widm.1392

ADVANCED REVIEW

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

Data mining privacy preserving: Research agenda

Inda Kreso | **Amra Kapo** | **Lejla Turulja**

School of Economics and Business, University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Correspondence
Amra Kapo, School of Economics and Business, University of Sarajevo, Sarajevo 71000, Bosnia and Herzegovina.
Email: amra.kapo@efsa.unsa.ba

Abstract

In the modern days, the amount of the data and information is increasing along with their accessibility and availability, due to the Internet and social media. To be able to search this vast data set and to discover unknown useful data patterns and predictions, the data mining method is used. Data mining allows for unrelated data to be connected in a meaningful way, to analyze the data, and to represent the results in the form of useful data patterns and predictions that help and predict future behavior. The process of data mining can potentially violate sensitive and personal data. Individual privacy is under attack if some of the information leaks and reveals the identity of a person whose personal data were used in the data mining process. There are many privacy-preserving data mining (PPDM) techniques and methods that have a task to preserve the privacy and sensitive data while providing accurate data mining results at the same time. PPDM techniques and methods incorporate different approaches that protect data in the process of data mining. The methodology that was used in this article is the systematic literature review and bibliometric analysis. This article identifieds the current trends, techniques, and methods that are being used in the privacy-preserving data mining field to make a clear and concise classification of the PPDM methods and techniques with possibly identifying new methods and techniques that were not included in the previous classification, and to emphasize the future research directions.

This article is categorized under:
Commercial, Legal, and Ethical Issues > Security and Privacy

KEYWORDS

bibliometric analysis, data mining, privacy preserving data mining, privacy preserving methods, privacy preserving techniques

1 | INTRODUCTION

Modern age hides many threats that come in every shape and form when it comes to individual privacy. Most of the time, it is not easily noticeable how big of an issue it is. Still, the fact is that the Internet, social media, and every other IT trend tends to threaten the privacy of the individuals using them (Mendes & Vilela, 2017). Since the right to individual privacy is one of the fundamental human rights, it is highly important to address the issues of privacy concerns in the data mining process. Information leak, unwanted identification, misuse of the private data, unauthorized data access, sensitive data disclosure, and many more are examples of how harmful the data mining process can be, and often it can result in an invasion of privacy. Individual privacy is a very sensitive issue in the modern age, and it should

be discussed more often. It is the responsibility of humanity to focus on the problem of individual privacy threats and solve the problems that have existed for quite some time and to make sure that the privacy issues will not be a problem in the future and data mining in general.

Data mining analysis has been evolving throughout the time, and it has started to develop in a new direction named privacy-preserving data mining (PPDM) to properly protect the sensitive data and private data of the individuals that are sharing their data for data mining proposes. According to Vaghashia and Ganatra (2015), it is actually about improving the "classical" data mining and making it more privacy aware using special privacy-preserving techniques that will make the correct data mining predictions, reveal interesting data patterns, but at the same time preserve the privacy and preserve the data utility. Yadav and Yadav states that PPDM is a data mining process that allows knowledge discovering from big data sets along with preserving privacy.

The goal of the PPDM is to avoid revealing sensitive data that are present in the data set that is being analyzed using the data mining process (Bourvil & Levi, 2017). The unwanted information leak can result in the identification of the person whose private data are being analyzed. This type of identity revealing can be possibly harmful and can put a specific person in danger or at least can jeopardize the reputation of the person. According to Garg and Chhinkaniwala (2011), new modern PPDM algorithms are focused on lessening the privacy invasion that could potentially lead to the harmful activities and misuses of the data that was initially intended exclusively for data mining process. The PPDM techniques that are used in the PPDM have a side effect of damaging the data that are being "protected" by specific PPDM technique. This effect causes the data utility to degrade, and it can potentially result in having incorrect predictions, patterns, and data mining results in general.

Data mining is a very important concept that contributes to almost every social sector, especially medicine and finance sector. Its importance is more recognized with the introduction of general protection regulation (GDPR). To make sure that the data mining process can be conducted smoothly, without quality or information loss, it is of utter importance to explore all the methods and techniques of PPDM, together with their performances and effectiveness. There are many PPDM techniques and even more PPDM methods that are based on different PPDM techniques. Also, there are many different approaches and classification of PPDM technics and methods derived from current scientific materials (such as Bellis, 2016; Moher, Liberati, Tetzlaff, Altman, & Grp, 2009; Shah & Gulati, 2016; Verykios et al., 2004; Wahono, 2015). All of them make quite a complex PPDM picture, and it is of utter importance to have a clear and concise classification of the PPDM techniques, methods, or any other data mining concepts. This article is focusing on the update of PPDM classification and offers an optimized PPDM classification model that can serve as a base for making significant progress in the field of PPDM.

In that sense, this research presents findings from a systematic literature review and bibliometric analysis to explore the area of PPDM. Thus, the contribution of this study is threefold. First, this study contributes to the data mining literature by analyzing the privacy preserving issues and current trends in the area of PPDM. Through bibliometric analysis, the study provides a detailed overview of the current number of published papers, the most influential authors, and presents a thematic analysis of the content of papers published in the field. Besides, this article draws on bibliometric data and identifies, analyzes, and interprets citation patterns, as well as the most influential works resulting from this analysis, using co-citation techniques. Furthermore, the paper offers a detailed analysis of two decades of scientific thought on PPDM techniques and methods. The second contribution of the paper is reflected in the systematization of PPDM methods and techniques through a systematic review of the literature. Second, to our best knowledge, the recent studies did not offer classification or upgrade of techniques and methods. Hence, this article seeks to fill the gap by providing an overview of the methods and techniques used for PPDM. In other words, this study aims to address the research trends present in the PPDM field and to direct future research toward providing studies that will result in effective improvement in the field. This study also aims to review, systematize, and present all the techniques and methods used in preserving privacy for data mining. This classification can serve as a direction for other researchers and scientific fields to be able to choose the technique/method aligned with their needs. On this basis, it identifies a comprehensive framework of PPDM techniques and methods, and complements it with recent techniques. Given the dynamic nature of the field, regular revision of PPDM techniques and methods is necessary. Third, by reviewing the scientific material published in the past 3 years, we have conducted a content analysis of the recommendations for future research offered by other authors in this field, and we have presented them in a systematic way. In addition, we presented our proposal of the scientific field that should be the research focus in the future. Accordingly, we have formulated the following research questions:

1. What are the research trends and research agenda for the future?

2. Is it possible to discover current methods and techniques related to privacy preserving in data mining?

The study is structured as follows. The next section outlines the literature review of current scientific discourse related to methods and techniques in the field of PPDM. The methodology section follows, as well as the results section. Finally, we summarize conclusions and discuss the results, as well as future research directions.

2 | LITERATURE REVIEW

2.1 | Privacy-preserving data mining

Recently, as the use of data mining is gaining momentum, vast amounts of data collected contain a large percentage of personal information (Custers, 2013). In the case where sensitive and personal data are analyzed, the question of violating the privacy of the individuals whose data are being analyzed arises by itself (Sun, Strang, & Pambel, 2018). This issue is of a purely ethical nature in the first place. Privacy is informally defined as the right of an individual to control personal information (Sun, Strang, & Pambel, 2018). Another definition of privacy is based on the claim that privacy is the ability for an individual, group, or institution to decide when, how, what, and to what extent to disclose about themselves (Sun, Strang, & Pambel, 2018). Furthermore, according to Herman and Custers (2004), in 1967, privacy was defined as the ability to control information, with the full rights of the person to decide when, how, and to what extent the information about the person will be disclosed.

Hoping that the problem of endangering individual privacy can be circumvented, modern data mining is increasingly leaning towards PPDM. In other words, it is working to improve the "classic" process of data mining by incorporating new methods and techniques that allow finding user information, while protecting the private and personal data of persons whose data is analyzed (Vaghashia & Ganatra, 2015). PPDM is a data mining process that allows you to find knowledge from a bunch of data, but at the same time, allows privacy (Yadav & Yadav,). Also, the methods used by PPDM are designed to guarantee the protection of privacy at a certain level, but also have the task of preserving the usability of data, so that data mining analysis can be successfully conducted (Vaghashia & Ganatra, 2015). The goal of PPDM is to preserve sensitive information hidden in a bunch of data that will be subjected to data mining analysis (Bourvil & Levi, 2017). Unwanted leakage, detection, and disclosure of private information during the data analysis process may compromise privacy at the individual level (Nathiya, Kuyin & Sundari, 2016).

According to Garg and Chhinkaniwala (2011), PPDM algorithms try to reduce the privacy threat caused by malicious parties during the data mining process. Most PPDM techniques perform data transformation to protect their privacy. Unfortunately, these techniques reduce the quality of the data to preserve privacy. What one aspect of the PPDM process implies is the application of data mining algorithms that will provide essential, useful, and relevant information from which knowledge will be constructed later, and at the same time, these algorithms will be able to adequately protect private and sensitive data and thus preserve privacy on an individual level (Bourvil & Levi, 2017). On the other hand, there are opinions that it is through PPDM inquiries that open and direct access to private information is created, which could be obtained based on ordinary inquiries, and that in fact even greater privacy is endangered in this way (Bourvil & Levi, 2017). What is very important to emphasize is that PPDM is based on the fact that data protection methods in some way involve the transformation of data, which would mean that the data may lose its quality and that the previous analysis may give inaccurate results (Vaghashia & Ganatra, 2015). Data quality degradation is one of the side effects that occur when trying to preserve privacy and is called utility (Vaghashia & Ganatra, 2015).

2.2 | Techniques and methods

In the literature so far, to the best of our knowledge, insufficient attention has been paid to the systematization of techniques and methods related to PPDM. Some of the authors have systematized previous research in this area (see Rajesh, Sujatha, & Arul, 2016; Shah & Gulati, 2016; Vaghashia & Ganatra, 2015; Verykios, Bertino, et al., 2004).

According to Vaghashia and Ganatra (2015), PPDM techniques can be classified to: anonimization, perturbation, randomization, condensation, and cryptography, They offered detailed analysis and synthesis of advantages and limitations of these techniques. On the other hand, Rajesh et al. (2016) study is based on a detailed review of algorithms based on PPDM technologies (data anonymization, data perturbation, cryptography, fuzzy logic, and neural networks) and

secure sum methods (based on cryptography) and methods of probability neural networks (based on neural network technique). The paper presents three PPDM techniques (cryptography, fuzzy logic, and neural networks), and presents PPDM methods that support these techniques (cryptography, fuzzy logic, and neural networks), and the identification of algorithms that have satisfactory privacy performance and preserving the usefulness of the data. Another approach is presented by Verykios, Bertino, et al. (2004), where techniques are classified as follows: data distribution, data modification, data mining algorithm, data or rule hiding, privacy preservation. Rajesh et al. (2016) argue that most appropriate methods are random perturbation, k-anonymity, horizontal portioned distribution, vertical portioned distribution, clustering, classification, association rule mining, secured sum computation, and aggregation. Those methods are previously described by Aggarwal and Yu (). The limits of computer calculations and theoretical limits associated with PPDM through huge data sets are also presented in the article. The study also provides examples of fields in which certain PPDM techniques are used.

One of the most detailed analysis of current methods and techniques is provided by Shah and Gulati (2016). According to them, there are two scenarios related to the collection of sensitive data from individuals in PPDM. The first scenario is for a trusted party to collect data, independent of the party that will be mining the data and the party whose data is being collected. This scenario is called the Central Commodity Server Scenario. In the Central Commodity scenario, the third party trusts the game to play a very important role, as each of the remaining two parties entrusts the central commodity server with a task that involves preserving the privacy of the individual (i.e., the party). Before the collected data is published, all participants in this process transfer their data to the server. The data mining process itself is performed independently on the server, and the server must privatize the data before the process itself, that is, protect the data with adequate protection before data mining (Shah & Gulati, 2016). Another scenario called a Distributed Scenario is a scenario where individuals themselves privatize their data before the data mining process is performed at all (Shah & Gulati, 2016). This approach differs from the Central Commodity scenario in that the parties involved in the data mining process must protect the data (privatize the data) before publishing the data. The data mining process can be performed by the party that owns the data, and then the rules of the association can be applied to such aggregated results. Also, there is an option to generate a version of the data with noise (data perturbation), and based on the required level of data protection generates a version of the data set with noise based on the original version of the data set.

Techniques used to protect data and preserve privacy in data mining include neural networks, fuzzy algorithms, anonymization, perturbation, and cryptography. These techniques represent five basic techniques in the field of privacy and data protection in data mining. In addition, Mendes and Vilela (2017) and Shah and Gulati (2016) argue that these techniques are used to develop certain methods that will be used for the purpose of privacy and data protection in data mining. Based on the logic of neural networks, the following methods have been created: probability neural networks, Bayesian network, and Cohen's SOMs (self-organizing maps). Methods that are created based on the techniques of fuzzy logic and fuzzy algorithms are K-means clusters, a priori algorithm, C-regression, and fuzzy clusters. The following methods were created on the basic anonymization techniques: K-anonymization, T-proximity, Condensation, L-diversity, P-sensitive anonymity, and (k, e) anonymity.

It is obvious that many authors confuse the terms of techniques and methods, and this fact opens an additional motivation for a detailed systematic review and bibliometric analysis of scientific material in this field to make a synthesis of papers, and try to classify the methods and techniques used so far into adequate units.

3 | METHODOLOGY

A systematic literature review and bibliometric analysis approaches were chosen for this study. Systematic Literature Review (SLR) is a popular method that involves the process of identifying, evaluating, and interpreting all available research evidence to provide answers to specific research questions (Wahono, 2015). The SLR was used to identify studies relevant to this research and to identify the agenda for the future, as well as to discover current methods and techniques related to privacy preserving in data mining? Besides, a bibliometric analysis technique was used to identify research trends, that is, to provide the mapping of the PPDM research field. Bibliometric analysis was used for the identification of the most influential papers in the PPDM field and to identify the roots of the field. Bibliometric analysis has been a very popular method in the last decade for analyzing content or citations within scientific material, to explore the area of interest more closely. The importance or influence of particular authors or specific scientific work is further explored. Bibliometric citation analysis represents a relatively new form of meta-analytical research or "meta-review" of the literature (Fetscherin, Voss, & Gugler, 2010; Kim & McMillan, 2008). It is a technique that considers the citation as the basic unit of analysis and investigates the

relationship between the articles in the given research area (Kim & McMillan, 2008). Key terms analysis was also used to identify the most relevant segments in the identified framework. Subsequently, qualitative content analysis was used to obtain answers to the research questions posed.

The software tools used to analyze bibliometric data and generate scientific clusters for the field were Bibexcel (Persson, Danell, & Schneider, 2009), VOSviewer (Krauskopf, 2018), and Pajek (Wang et al., 2016).

Table 1 shows the research methods associated with the research objectives of this study and the software used in this research.

The literature review was conducted through the stages: planning, conducting, and reporting literature reviews (Wahono, 2015), as presented in Figure 1.

3.1 | Planning stage

As indicated, the first stage implies systematic review requirements together with research questions and the development of a review protocol. The research questions are presented in the introductory part of this study. The review protocol involves identifying a "search string" as well as a search strategy (the process of selecting studies with inclusion and exclusion criteria) to reduce the possibility of researcher bias (Wahono, 2015).

The search string developed was "TITLE (privacy preserving data mining)". Before starting the search, an appropriate set of databases must be selected to increase the likelihood of finding highly relevant posts. For the purposes of this study, the Web of Science citation database was selected.

3.2 | Conducting stage

The conducting stage starts with the selection process, where the records are identified through the database searching. The Web of Science database was searched on April 10th, 2020. In this phase, a total of 104 articles was identified. The conducting stage followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which can be seen in Figure 2 in detail (Moher et al., 2009).

In total, 104 records were identified by following the developed search protocol. Then, several screenings of the identified studies were conducted:

- First, a screening of the titles was carried out, and no study was excluded. The exclusion/inclusion criteria at this stage were that the paper deals with PPDM.

TABLE 1 RQ, methods, and software tool used

Research question	Method/analysis		Aim	Tool/software
What are the research trends and research agenda for the future?	Number of published papers through years		Dynamics of publishing	Web of science
	The most prolific authors		Dynamics of publishing	Web of science
	Number of citations (in other journals) through years		Scientific growth	Web of science
	Bibliometric analysis	Citation analysis	Scientific growth within	Web of science, VosViewer
			The most influential papers published in	Web of science, VosViewer
	Qualitative content analysis	Systematic literature review	Identification of future research areas	
Is it possible to discover current methods and techniques related to privacy preserving in data mining?	Qualitative content analysis	Systematic literature review	Identification of current methods and techniques in PPDM	

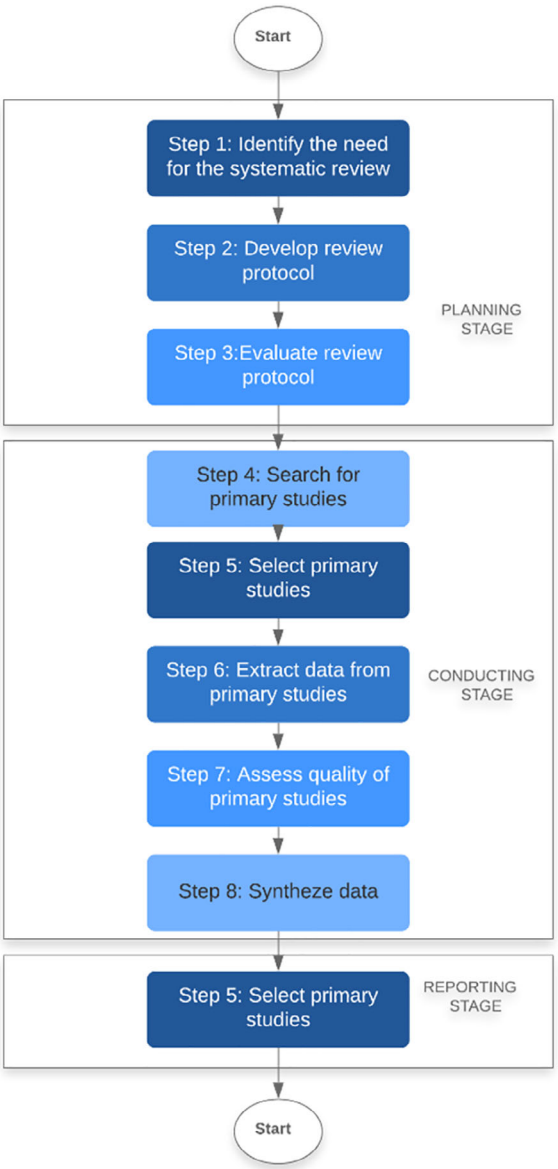

FIGURE 1 Systematic Literature Review Steps (Wahono, 2015)

- Second, screening of the abstracts was performed, whereby exclusion/inclusion criteria were
  - The article addresses the security and privacy issues in the process of data mining.
  - The article addresses the methods and techniques of PPDM.

  Nineteen studies were excluded in this stage.
- Third, content analysis of 85 studies has been conducted whereby exclusion/inclusion criteria were the same as in the previous step. During the content analysis of the identified studies, the need for expanding the pool of papers based on cross-referencing was noted. Hence, 75 additional papers were identified, which were included in further analysis. Finally, out of a total of 160 papers, 50 were identified as primary studies for further narrative synthesis.

4 | RESULTS

4.1 | Research trends of data mining privacy preserving concept

In order to answer the first research question posed by bibliometric analysis (What are the research trends and research agenda for the future?), we used a bibliometric analysis of identified papers in the WoS database. Figure 3 is a graphical representation of the number of publications by year in the field of privacy protection and data protection in data

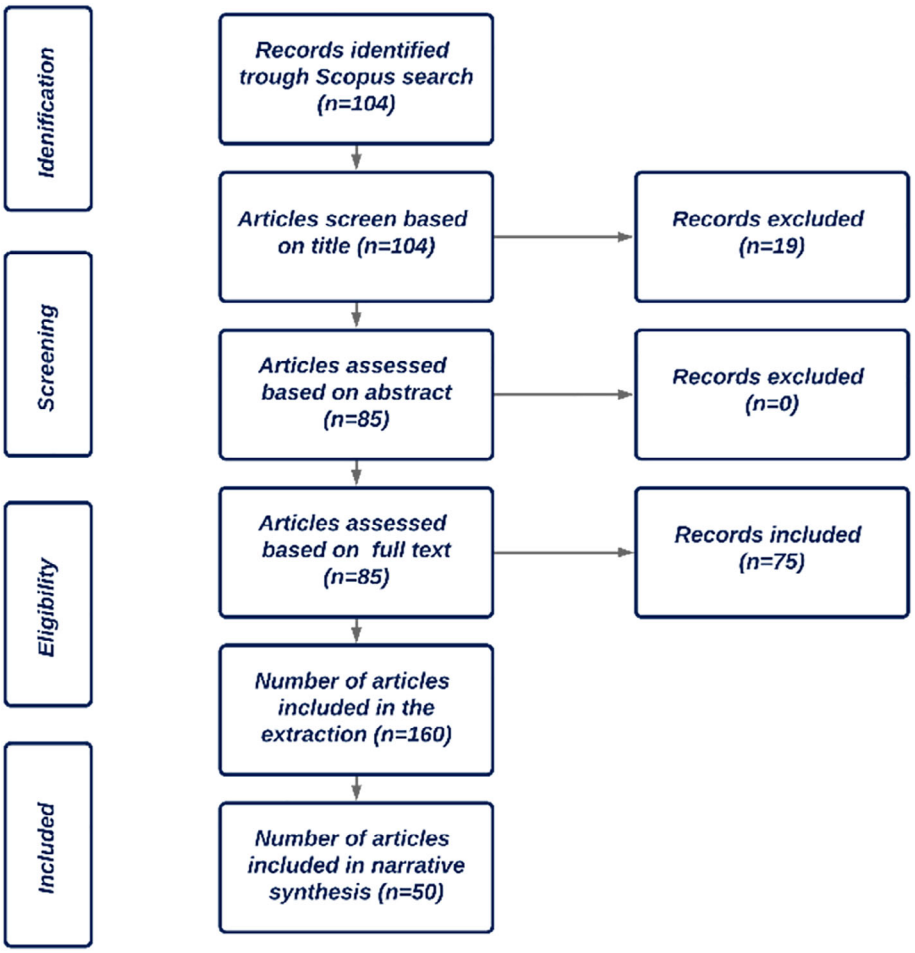

FIGURE 2 PRISMA flow diagram presenting steps in the identification and screening of records (Moher et al., 2009)

mining from 2000 to 2020. As it can be observed, the early beginnings of this field date back to 2000 when there are already perceived privacy issues that data mining could cause. However, interest in this field has been sharply declining throughout the next year, and only since 2004, we see the emergence of studies and papers on a given topic. As we approach the current date, it can be observed that the number of studies over the next 10 years (from 2004 to 2014) is generally increasing (with a slight decrease in 2007, 2010, 2012, and 2013). Since 2014, awareness of the privacy problem caused by the data mining process is gaining importance, and the number of studies is increasing significantly, and in 2018, we have a scientific peak in the number of publications.

The next step in the research is to analyze the citation and co-citation of the 104 papers previously identified, which was conducted using VosViewer software. In recent literature, VOSviewer is a software tool used to create clusters based on data connectivity and to visualize and explore these clusters.

Figure 4 shows the citation network and the citation frequency of the author. Each link has a specific value that is represented by a specific numerical value. This software does not show the value of the link, but together terms and links make up the citation network. Nodes in the image in different colors represent the representation of a particular term in the literature, and the size of the node represents the frequency of the term, the greater the node is, the more significant paper is. In this case, it can be seen from Figure 4 that the author Agrawal is actually the most cited author followed by Lindell and Pinkas (2000) and Kantarcioglu and Clifton (2004).

Author Rakesh Agrawal has been singled out as the most cited author in the field of data privacy and data mining. Agrawal's research field covers data mining, privacy, and data protection in data mining, and is often considered a major pioneer in data mining, as many of Agrawal's papers and studies actually represent basic data mining and privacy and data protection in data mining. Also, Rakesh Agrawal coined the term “privacy preserving data mining”.

We run a new analysis using Bibexcel and Pajek software tools to identify the root of the scientific field. According to the results, we derived the conclusion that the oldest source is dating from 1985 and has used the term individual privacy and security when it comes to the concept of data analysis. In this article, authors suggest three steps

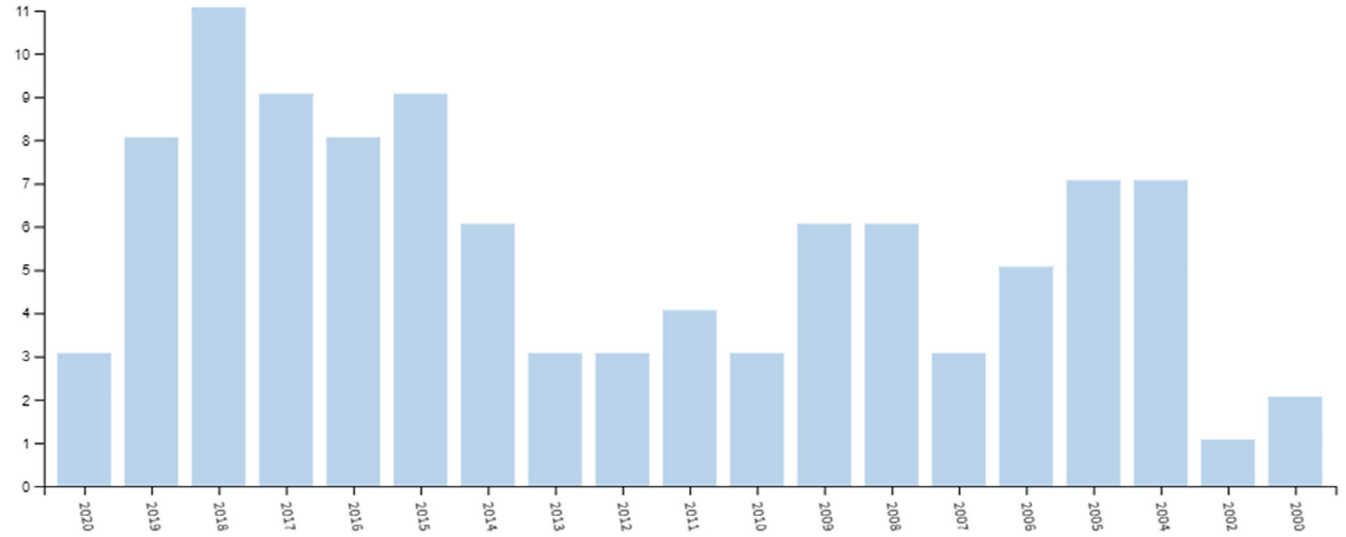

FIGURE 3 Number of publications by year. *Total = 104

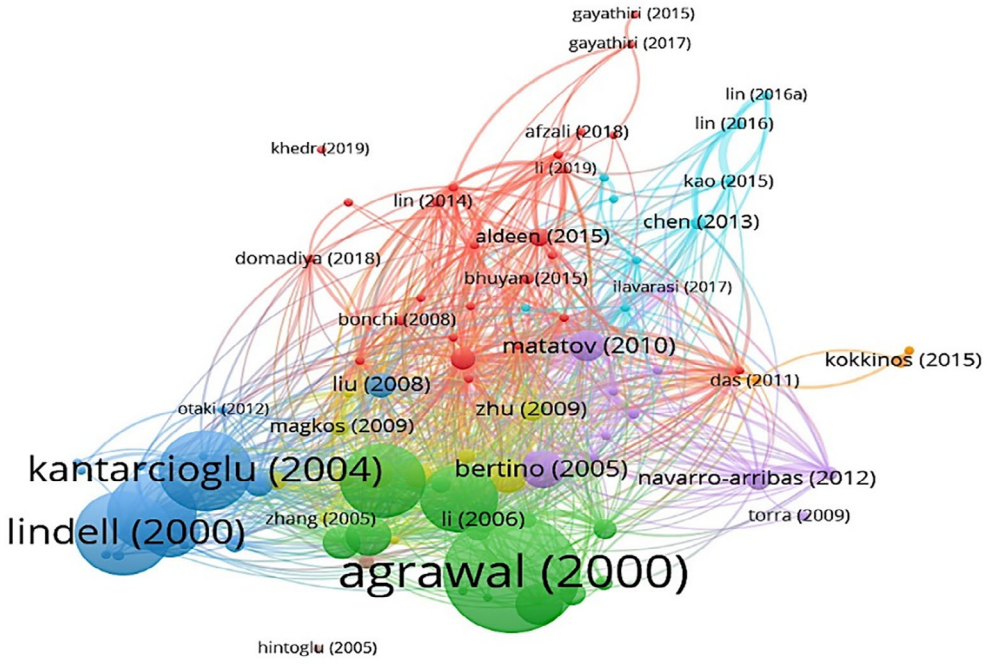

FIGURE 4 Most influential authors separated in clusters

(identification of the underlying density function; generation of a distorted series for each variable and mapping and replacement of the distorted series) when it comes to data distortion by probability distribution taking into account the privacy of an individual belonging to the original data set (Liew, Choi, & Liew, 1985).

4.2 | Document co-citation analysis

To further explore trends and hot topics in the analyzed research field, we conducted co-citation analysis using VOSViewer. Bibliometric analysis of co-citation is a meta-analytical method that shows the interconnections between research articles and topics (Shah, Lei, Ali, Doronin, & Hussain, 2019). This analysis yielded three clusters (Figure 5). Different clusters of closely associated authors are denoted by the same cluster color (green, red, and blue). After a comprehensive analysis of the articles that appear in the cluster, the characteristics of each cluster are presented, as well as the corresponding name defined.

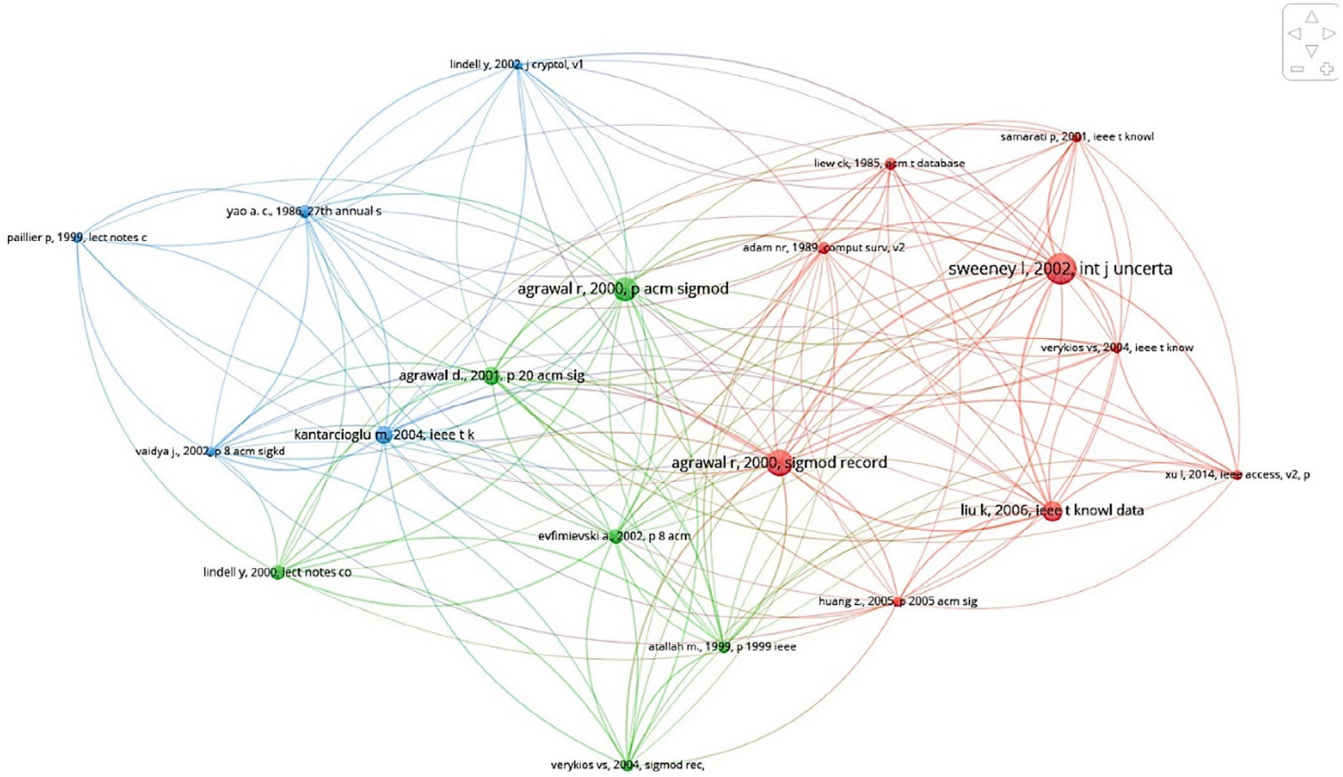

FIGURE 5 Document co-citation analysis

Representative authors and publications in *green cluster* entitled "The most frequently used techniques in PPDM" generally deals with the important need for data protection and privacy. While Agrawal and Srikant (2000) suggest that sensitive values in a user's record needs be perturbed using a randomizing function so that they cannot be estimated with sufficient precision, Evfimievski, Srikant, Agrawal, and Gehrke (2004) postulates that data has to be randomized to preserve the privacy of individual transactions. The rest of the papers in green cluster describe certain data mining privacy preserving techniques used for privacy preserving (Evfimievski et al., 2004; Lindell & Pinkas, 2002; Paillier, 1999; Verykios, Elmagarmid, Bertino, Saygin, & Dasseni, 2004; Yao, 1986). The most prolific technique used in this systematic literature review, specifically, in this cluster is recognized as cryptography (Lindell & Pinkas, 2002; Paillier, 1999; Yao, 1986) Since the discovery of public-key cryptography by Diffie and Hellman (1976), very few convincingly secure asymmetric schemes have been discovered despite considerable research efforts (Goldwasser, 1997; Paillier, 1999). Kantarcioglu and Clifton (2004) states that cryptographic tools can enable data mining that would otherwise be prevented due to security concerns and provide procedures to mine distributed association rules on horizontally partitioned data. In this way, they draw the conclusion that distributed association rule mining can be done efficiently under reasonable security assumptions. Agrawal and Aggarwal (2001), in later work, are proposing metric for quantification and measurement of the effectiveness of PPDM algorithm.

Red cluster entitled "Different methods for PPDM "presents different methods for PPDM that are going to be explained in deal in Section 3 of this article: for example, randomization, oblivious transfer, and k-anonymity. While Huang, Du, and Chen (2005) proposed a modified randomization scheme where random noises are "similar" to the original data, Liu, Kargupta, and Ryan (2006) explored the possibility of using multiplicative random projection matrices for privacy preserving distributed data mining which ultimately proved very successful. In addition, the randomization method is also explored by Agrawal and Srikant (2000) by keeping sensitive values in user records in privacy. Another important method derived in this cluster, as it can be seen in Figure 5, is an oblivious transfer from the cryptography technique. The oblivious transfer protocol involves two parties, the sender and the receiver, where the receiver learns (and nothing else), and the sender learns nothing (Lindell & Pinkas, 2002). Besides this method, k-anonymity is explored to provide without compromising the integrity of the information released by using different techniques (Samarati, 2001). In addition, Sweeney (2002) proposed a k-

TABLE 2 Techniques of PPDM

Technique name	Definition of technique	Studies that analyzed the technique
Data anonymization	Data anonymization (re-identification) is a technique that is composed of various methods that are used to efficiently prevent the disclosure of personal data.	Revathi (2017); Herman and Custers (2004); Rajesh et al. (2016); Brickell and Shmatikov (2008); Silva, Basso, and Moraes (2017)
Data perturbation	The data perturbation technique uses a mechanism that distorts data to protect the privacy of personal data.	Revathi (2017); D. J. Patel and Swati (2015); N. Patel and Patel (2015); Liu, Giannella, and Kargupta (2008); N. et al., (2016)
Cryptography	Cryptography is a process encrypting data that can be decrypted only by an intended receiver.	Manikandan, Porkodi, Mohammed, and Sivaram (2018); Xu (2011); Revathi (2017); Schlitter (2008); Clifton, Kantarcioglu, Vaidya, Lin, and Zhu (2002); Rajesh et al. (2016); Herman and Custers (2004)
Fuzzy logic and algorithms	Fuzzy logic is a generalization of the classical mathematical logic.	Chris and Marks (2008); Manikandan et al. (2018); Gupta and Joshi (2009); Kumar, Varma, and Sureka (2011); Rajesh et al. (2016); Zadeh (2015)
Neural networks	Neural networks are representing the learning process that is used for discovering the unknown data patterns.	Kaushal and Shukla (2014); Schlitter (2008); Rajesh et al. (2016)

anonymity protection model that explored related attacks and provided ways in which these attacks can be thwarted.

The final cluster derived trough the research is a *blue cluster* entitled "Key strategies, steps and terms for more efficient PPDM" and deals with the definition of key terms related to the research area as well as identification of strategies. First, four different types of users were identified that are involved in data mining applications: data provider, data collector, data miner, and decision-maker (Xu, Jiang, Wang, Yuan, & Ren, 2014). Second, strategies or steps are identified in order to create the concept of data mining, bearing in mind privacy preserving (Liew et al., 1985; Verykios, Elmagarmid, et al., 2004). Detailed information is available in Appendix A.

4.3 | Identification of methods and techniques

PPDM uses various techniques and methods. Throughout the process of the systematic literature review, five dominant PPDM techniques were defined. Those five techniques, along with their definitions and the studies that were used to define the techniques, are represented in Table 2: Techniques of PPDM.

Throughout the process of the systematic literature review, seven PPDM methods that were mentioned the highest number of times in the literature were defined. Seven privacy preserving methods, along with their definitions, techniques on which they are based on and the studies that were used to define the methods are represented in Table 3: Methods of PPDM.

In addition to the 12 described and defined PPDM methods and techniques, there are also three additional methods or techniques that were recognized throughout the systematic literature review and that are possible to categorize and included as relevant methods. Three of them were identified in the field of privacy preserving and were added in the classification, and those methods/techniques are (Table 4):

1. Data swapping (method based on anonymization technique);
2. Semi-Honest Parties (method based on cryptography technique);
3. Secure Cloud Computing based PPDM technique.

Figure 6 was created based on the PPDM Classification Hierarchy from the study Privacy Preserving Data Mining: Techniques, Classification, and Implications—A Survey by Shah and Gulati (Shah & Gulati, 2016). Figure 6: Classification of PPDM techniques and methods is modified and adapted by adding few more defined methods that were

TABLE 3 Methods of PPDM

Method name	Name of the technique on which the method was based	Definition of the method	Studies that analyzed the method
Probabilistic neural network PNN	Neural networks	PNN represents the special version of the neural networks that are using in classification and for data pattern recognition.	Rajesh et al. (2016)
Apriori algorithm	Fuzzy logic and algorithms	A priori algorithm is one of the standard algorithms that are used in the process of data mining. It is used for the process of searching through the data sets.	Manikandan et al. (2018); Shmueli, Bruce, Yahav, Patel, and Ke (2018); Xu (2011); Kavitha and Elango (2017); Rajesh et al. (2016)
K-anonymity	Data anonymization	The most accurate definition of the k-anonymity was given by the Latanya Sweeney and Pierangela Samarati in their paper generalizing data to provide anonymity when disclosing information where they wrote:" A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 {\displaystyle k-1} individuals whose information also appear in the release."	Kawano, Honda, Kasugai, and Notsu (2013); Rajesh et al. (2016); Hussien, Hamza, and Hefny (2013); Qi and Zong (2012); Bradić-Martinović and Zdravković (2013); Samarati and Sweeney (1984)
SMC (secure multiparty computations)	Cryptography	SMC ensures the computation of the function while at the same time preserving the privacy of the input and output of the function.	Schlitter (2008); Rajesh et al. (2016); Shah and Gulati (2016); Rajesh et al. (2016); XU (2011)
Oblivious transfer	Cryptography	The oblivious transfer is actually a protocol that is used in cryptography, and its purpose is to send one part of the bigger message that is being transported, but in a way that is unclear or oblivious which part of the message is being transported at the moment.	Aggarwal and Yu (); Clifton et al. (2002); Shah and Gulati (2016); Mendes and Vilela (2017); Sharma and Ojha (2010); Poovammal and Ponnavaikko (2010)
Noise addition	Data perturbation	Noise addition is a perturbation-based method in which the noise is added in the data set to distort the data.	Revathi (2017); Kargupta, Datta, Wang, and Sivakumar (2005)
Randomization	Data perturbation	The randomization method is a method that uses a randomization algorithm to recover the properties of the data, whereas the single entities are distorted.	Kargupta et al. (2005); Aggarwal and Yu (); Rajesh et al. (2016); Garg and Chhinkaniwala (2011); N. Patel and Patel (2015); Yadav and Yadav (2016)

identified in the process of the literature review in this study. The methods that were added in the classification were not initially included in the classification described by Shah and Gulati (2016).

The classification of the PPDM methods and techniques presented in Figure 6 represents the proposed new classification with the inclusion of two methods and one technique that were not included in the previous classification proposed by Shah and Gulati (2016). These are methods and techniques that, to the best of our knowledge and currently available scientific material, could be classified. Certainly, given the rapid development of technology, there are many methods and techniques that are just being tested to serve the need for PPDM and they relate to the concepts of Big data, Internet of Things, or Cloud. Certainly, updated and more detailed classification will be needed in the future.

TABLE 4 Newly identified PPDM methods

Method name	Name of the technique on which the method was based	Definition of the method	Studies that analyzed the method
Data swapping	Anonymization	The data swapping method exchanges the values of the attributes between different records in the database. The goal of this method is to exchange enough amounts of attribute values to secure the individual records (it is impossible to find any sensitive information among the individual records).	Patel and Swati (2015); Mendes and Vilela (2017); Gupta and Joshi (2009); Herman and Custers (2004)
Semi-honest parties	Cryptography	Both parties in the semi-honest model have to follow the exact predefined protocol, and the protocol dictates that there is no interchangeable knowledge about input and output.	Lindell and Pinkas (2000); Rajesh et al., 2016
/	Secure cloud computing-based PPDM	Since the cloud computing platforms are frequently used because of their storage capability, privacy issues and data protection are the first two terms that have to be discussed to prevent information leakage or unwanted harmful privacy issues. All of the methods and techniques can be used to support privacy and security issues that the user can encounter using one of the cloud computing platforms.	Kumaraswamy and Venugopal (2017)

4.4 | Quinquennial trend analysis

In addition to identifying the methods and techniques, a brief analysis was also made through a 5-year timeframe for the papers published in this field for the period of 2000–2020 (Figure 7). At the outset, where a small number of papers were recorded, the authors relied on the following methods and techniques: data perturbation, decision tree, and cryptography with the most common application in business applications, databases, and online statistical databases (Agrawal & Srikant, 2000; Fukasawa, Wang, Takata, & Miyazaki, 2004; Kantarcioglu & Clifton, 2004).

With the development of technology and greater use of the Internet, in the next quinquennial, a much broader use in e-commerce, telecommunications, internet elections, law enforcement, or direct marketing is visible. In addition to the methods/techniques already mentioned, data anonymization, and neural networks (Duan & Canny, 2009; Zhan & Matwin, 2007) are introduced.

The next two periods we analyzed make up more than 70% of the scientific material and are full of new methods and techniques, and combinations thereof, to ensure data privacy in data mining. So, in 2011–2015, we still have the most common methods and techniques recorded with the addition of fuzzy logic and algorithms, but now their application has shifted to medical and financial data and user data generated on the web (Sah & Gunasekaran, 2014; Wu & Xiao, 2013). Furthermore, what has characterized the most recent period is the use of techniques and methods in emerging technological concepts and circumstances such as Big data, cloud technologies, twitter data, and the Internet of Things (IoT). Certainly, scientific research in the field of medicine and finance is still present where the most sensitive data lies and need to be protected (Domadiya, Kumar, & Rao, 2019; Johnsana, Rajesh, Sangeetha, & Kishore Verma, 2016).

All this leads us to the conclusion that there is no optimal method/technique for protecting the privacy of data using data mining, and that their application changes with the advent and change of trends in technology. Certainly, much will be written on this topic in the coming period, given the comprehensive measures being taken to protect privacy around the world.

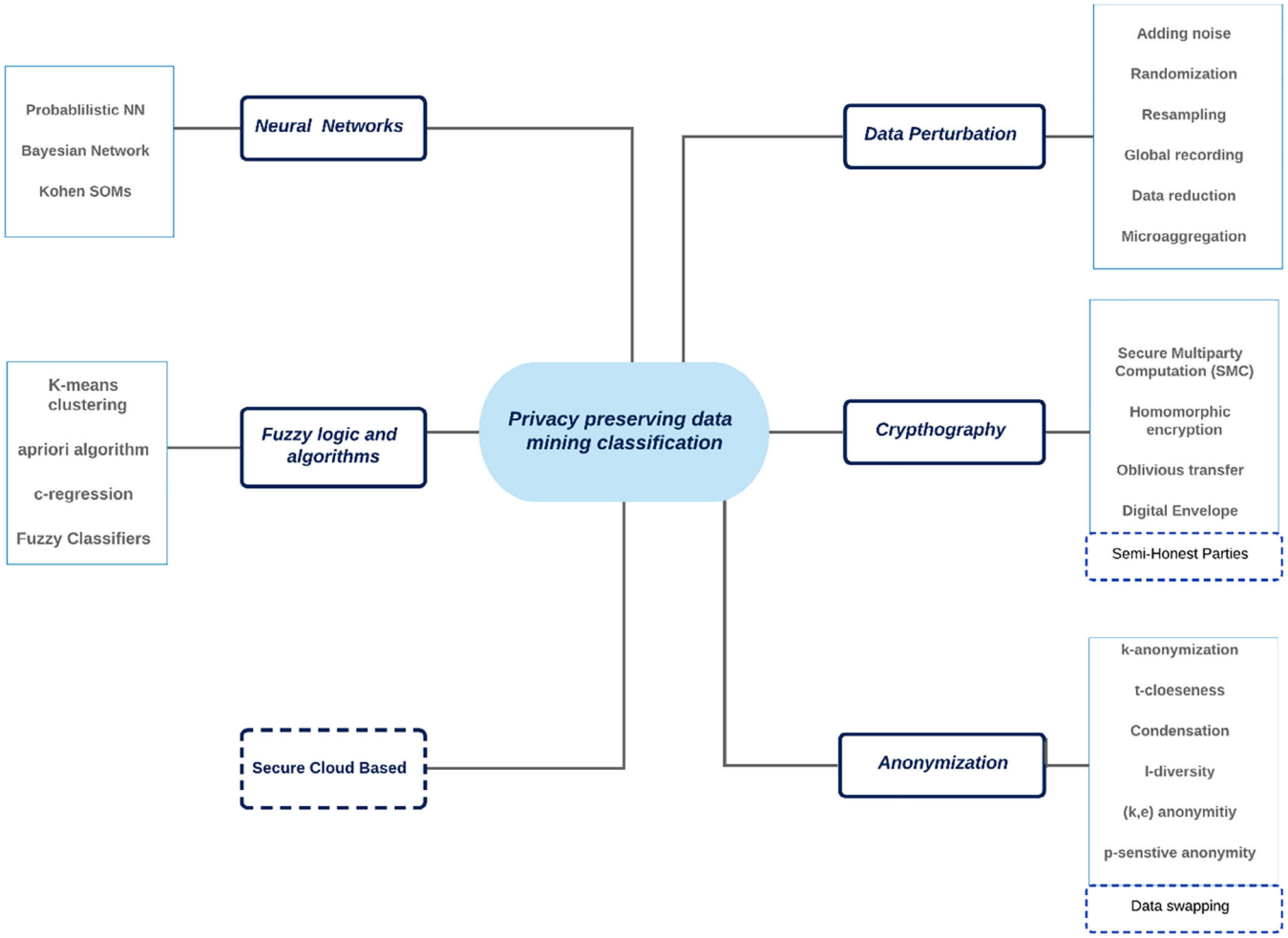

FIGURE 6 Classification of the PPDM techniques and methods

4.5 | Future research

One of the main goals of the systematic literature review process was to discover new suggestions and directions that future researches should follow. To be able to predict the future direction of development of the PPDM field, it is of the utter importance to explore the past researches, the previous solutions that were done in the previous sections of this article. In the following section, the focus is on the identification of the new directions in which the field of the PPDM should be developed. Detail analysis is available in Appendix B.

The studies that were in the focus of the analysis were the ones that were published in the last 3 years (2017–2019). There were in total of 28 studies published in the proposed period. Further, in the analysis, 21 studies were included, and seven studies were excluded because they did not offer any future research directions or suggestions.

Recommendations for future research offered in these studies were singled out, creating a text dataset, and then a detailed thematic analysis of the text was conducted. The results of the thematic analysis indicate the three most common directions that should be pursued by the future researches:

- Development of new and more efficient PPDM techniques, methods, approaches, and solutions.
- Identification and optimization of the most optimal PPDM techniques.
- Quantify the PPDM methods and techniques.

From the total of 21 studies, the first research direction development of new and more efficient PPDM techniques, methods, approaches, and solutions was supported by 11 studies, the second proposed research direction identification

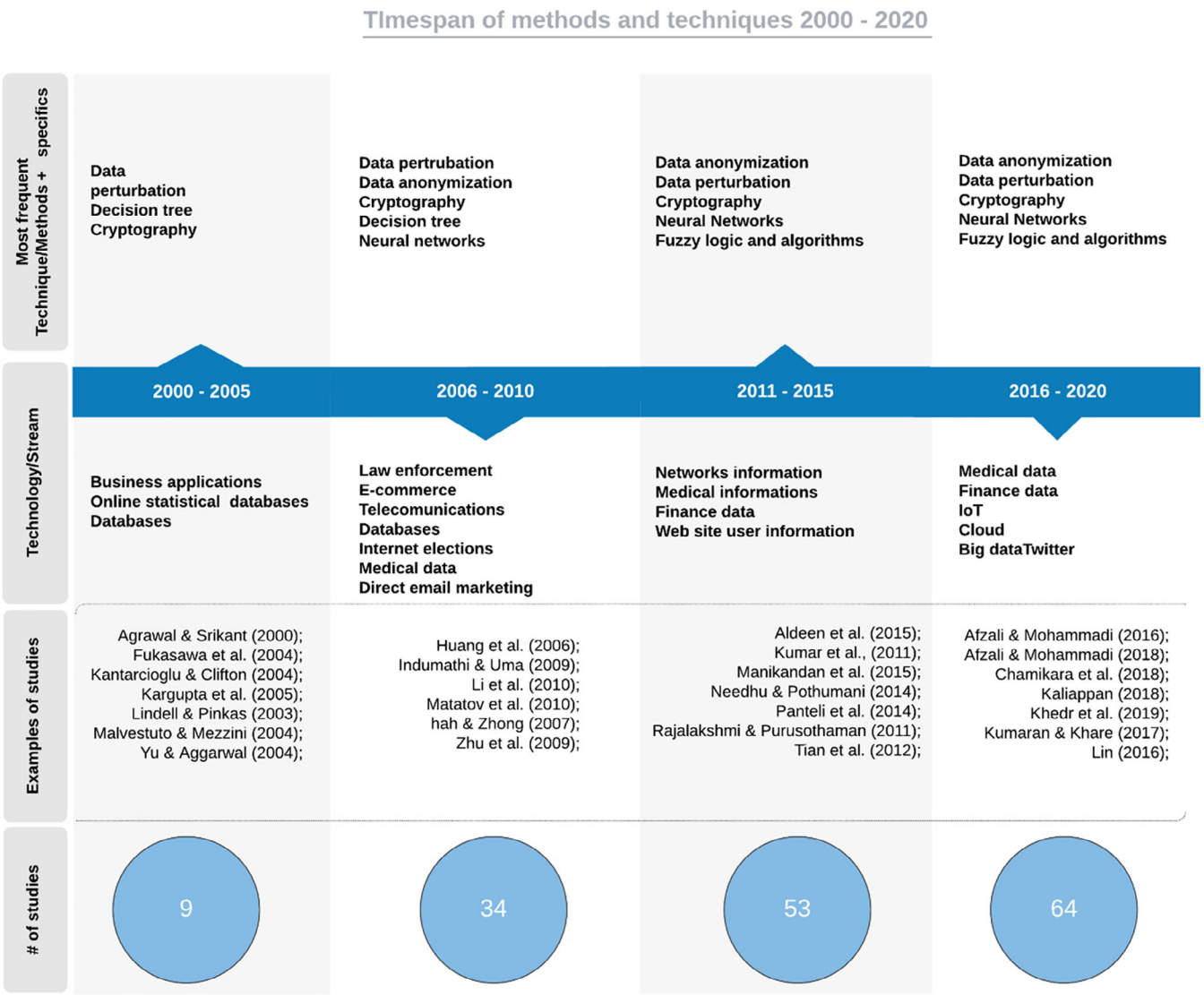

FIGURE 7 Timespan of techniques and methods of PPDM

of the most optimal PPDM techniques was supported by seven studies and the third research direction considering the quantification of the PPDM methods and techniques was supported by three studies.

When it comes to the development of new and more efficient PPDM techniques in the future researches, Agrawal and Srikant (2000) state that the main focus should be on the development new techniques that will incorporate all the privacy concerns, whereas both Kantarcioglu and Clifton (2004) and Verykios, Elmagarmid, et al. (2004) both suggest main focus to be on the newly programmed secure algorithms for data mining classification and clustering. According to Wang, Zhu, Chen, and Chang (2018), the best way to improve the field of PPDM is to focus on the investigation of the other techniques that could possibly have a better level of privacy protection.

On the other hand, the second proposed direction claims that the identification and optimization of the most optimal PPDM techniques would be necessary in future researches. According to Li and Xue (2018), the different privacy protecting algorithms have to be analyzed to conclude which one is the most optimal one. Samarati (2001) states that it is necessary to focus on the investigation of the algorithms characterized as efficient to be able to optimize those algorithms in which optimization is necessary. Afzali and Mohammadi (2018), Mendes and Vilela (2017), and Ilavarasi and Sathiyabhama (2017) state that it Is of utter importance to optimize the methods, techniques, and algorithms that were presented in their work.

According to Liu et al. (2006), quantifying the PPDM methods and techniques as a possible future research direction is absolutely necessary since there is still no efficient quantification of the privacy and data utility. Xu et al. (2014) and

Kalyani, Rao, and Janakiramaiah (2018) suggest that the focus of future research should be on the new techniques and on the quantification of the privacy and utility at the same time.

After analyzing the proposals for future research, we believe that additional research direction should concern the application of PPDM in the field of information systems and knowledge management in the industrial engineering and manufacturing industry. To be more specific, the goal is to apply the PPDM through information systems to simplify the production process, product design, and at the same time, factory automation. The reason why the manufacturing industry is using information technology in the first place is to enhance their competitiveness in the global market and to make supply chain management more efficient (Harding et al., 2006). One of the goals is to automate all production processes and business functions. The global market and to make supply chain management more efficient. One of the goals is to automate all production processes and business functions (Harding et al., 2006).

The role of data mining in the engineering and industry is extremely important, because it improves the quality of the process (Köksal, g., Batmaz, I. and Testik, M. C, 2011). Unfortunately, the data mining problem in this field is not much different than the data mining problem in every other field—privacy invasion. This is where the PPDM should be introduced to make sure the data privacy is preserved, and at the same time, the quality of the product is still being maintained (Wang, Tong, Eynard, Roucoules and Matta, 2007). The three main functionalities that were recognized by the author on which the PPDM would focus on future work and try to make them more efficient are: manufacturing system modeling, quality control, and process planning and scheduling (Saad, 2018). The point of the PPDM techniques and methods, classification was to try to systematize all the techniques, methods, and their characteristics to able to determine which technique/method is the most suitable technique/method to be used in different cases, whether it is manufacturing systems modeling, quality control or process planning and scheduling, which again depends on the data type that are used in the mentioned modules (Harding et al., 2006). The reason why PPDM should be used in these cases rather than classic data mining is that in most cases, companies and industrial facilities want to secure and preserve their data (Testik, 2011). PPDM should be able to quickly analyze large databases where all the data used by manufacturing system modeling, quality control, and process planning and scheduling are stored (Ur-Rahman and Harding, 2012). The analysis would not harm the data, would not decrease the data utility, and the PPDM technique or method or even a combination of the different techniques and methods will be determined based on the data types of the data being used in the modules of the manufacturing system modeling, quality control and process planning and scheduling (Wang, Roucoules and Matta, 2007). The PPDM classification system in this article will help to correctly determine which PPDM technique or method is the most suitable to be used.

5 | CONCLUSION

This study investigates and uncovers the development of the PPDM research field from 2000 to 2020 through a systematic literature review and bibliometric analysis, offering a special contribution by dividing the period examined into four quinquennials, and presenting an extensive analysis of research focus for each period. According to 104 articles identified in WoS and additional 75 papers identified through cross-referencing, we designated publishing trends through the years, important publications, most influential authors, and analyzed the research field network identifying three dominant research topics. Also, we identified PPDM methods and techniques and offered an updated classification. We also presented the recommendations of recent studies for future research and predicted potential trends and hot topics in the future.

Science and scientific progress can be monitored by statistical analysis of publication outputs. In addition, the statistical analysis offers various tools for mapping scientific research and locating the most prominent actors on stage (Bellis, 2016). Guided by current research aimed at mapping a particular scientific field, as well as presenting scientific trends, this article offers a comprehensive analysis of the scientific field of PPDM in the last two decades and contributes to the existing knowledge in the following ways. First, using the systematic literature review technique, relevant papers in the scientific field were identified. Second, using bibliometric analysis, the scientific field was mapped, and the most prominent studies and authors were identified, and the co-citation analysis was presented. In addition, the co-citation analysis of identified studies yielded three clusters, which also represent the three thematic areas of the field. Third, content analysis of primary studies identified through a systematic literature review provides an overview of the most relevant PPDM techniques and methods. A framework that represents the most comprehensive classification to date has also been identified and supplemented by newer techniques and methods. In this way, this article is a synthesis of the scientific field with special reference to publication trends, as well as the PPDM techniques and methods used. Finally, the paper, with a thorough content analysis of recent studies, also provides guidance for future research in the field.

This study answers two research questions. Various techniques were employed to answer the first research question related to the trends and research agenda for the future (systematic literature review, bibliometric analysis, qualitative content analysis). This article identified 104 available papers in the Web of Science citation database from 2000 to 2020 (75 papers are added through cross-referencing identification). Our study revealed that the study of author Rakesh Agrawal is the most cited author in the field of data privacy and data mining and coined the term "privacy preserving data mining" (Agrawal & Srikant, 2000). Immediately after Agrawal's most cited work, the other most cited authors are Lindell and Pinkas (2000) and Kantarcioglu and Clifton (2004). Using VosViewer software tool, we were able to identify three key clusters analyzed in detail and derived one of the main contributions of this article. The first cluster identified in the research "The most frequently used techniques in PPDM" deals with cryptography as the most frequently used technique in PPDM. The second cluster, "Different methods for PPDM " identified three methods: randomization, oblivious transfer, and k-anonymity for achieving PPDM. Finally, the third cluster "Key strategies, steps, and terms for more efficient PPDM" deals with the definition of key terms related to the research area as well as identification of strategies. Besides current trends, this article offered future direction of this field trough three main topics: (a) Development of new and more efficient PPDM techniques, methods, approaches, and solutions, (b) Identification of optimal DM techniques, and (c) quantify the PPDM methods and techniques.

The systematic literature review and qualitative content analysis enabled us to answer the second research question to identify the most important and most frequently used techniques and methods in PPDM. Five main PPDM techniques were defined: data anonymization, data perturbation, cryptography, fuzzy logic and algorithms, and neural networks. Besides techniques, seven PPDM methods were identified: Probabilistic Neural Network, Apriori algorithm, K-anonymity, SMC, Oblivious transfer, Noise addition, and Randomization. The additional contribution of this article lies in the discovery and upgrade of Shah and Gulati (2016) classification of methods and techniques by adding data swapping (Gupta & Joshi, 2009; Herman & Custers, 2004; Mendes & Vilela, 2017; Patel & Swati, 2015), semi-honest parties (Lindell & Pinkas, 2000; Rajesh et al., 2016), and also Secure Cloud Computing based PPDM (Kumaraswamy & Venugopal, 2017). The extracted data in the process of data extraction prove that there is no generic solution when it comes to privacy preserving and security in the data mining process. There is no uniform method, technique, nor a solution that would cover all the security issues and privacy preserving issues in the data mining process. The articles and studies included in the data extraction process suggest that there could be specific solutions for specific privacy or security issues, but there is no uniform solution.

Through the use of several literature reviews and bibliometric techniques, and a detailed analysis of the content of identified papers by the authors, this study is useful for presenting a comprehensive framework for PPDM research. It should be noted that, although this study focused on articles identified from WoS, we tried to avoid bias for high-quality publications by adding articles that were identified through the cross-referencing technique, so that the results represented the research field as realistically as possible. Although the bibliometric analysis performed using specialized software is completely objective, it can be pointed out that the interpretation of the results is somewhat subjective, which is a potential limitation of the study. However, through multiple discussions by three authors in this article, we have tried to overcome the subjective interpretation of the results. Given the dynamics of the development of data mining techniques, future research should continuously update the classification of PPDM methods and techniques.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

AUTHOR CONTRIBUTIONS

Inda Kreso: Conceptualization; formal analysis; methodology; data curation; resources; software; writing-original draft. **Amra Kapo:** Conceptualization; data curation; formal analysis; methodology; project administration; visualization; writing-original draft. **Lejla Turulja:** Conceptualization; formal analysis; methodology; resources; software; writing-original draft.

ORCID

Inda Kreso https://orcid.org/0000-0002-5556-4669
Amra Kapo https://orcid.org/0000-0001-5066-7696
Lejla Turulja https://orcid.org/0000-0003-1493-8318

FURTHER READING

Adam, N.R., & Wortmann, J.C. (1989). Security-control methods for statistical databases : a comparative study. *ACM Computing Surveys*, *21* (4), 515–556.

Afzali, G. A., & Mohammadi, S. (2016). Privacy preserving big data mining: Association rule hiding. *Journal of Information Systems and Telecommunication*, *4*(2), 70–77. https://doi.org/10.7508/jist.2016.02.001

Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *Springerplus*, *4*(1), 1–36. https://doi.org/10.1186/s40064-015-1481-x

Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios, (1999). Proceedings 1999 Workshop on Knowledge and Data Engineering Exchange. *No.PR00453*,

Chandrakanth, P., & Anbarasi, M. S. (2017). A comprehensive survey of privacy preserving data mining techniques. *Journal of Advanced Research in Dynamical and Control Systems*, *9*(Special Issue 12), 2239–2253.

Custers, B, Calders, T, Schermer, B, & Zarsky, T (2013). Data Dilemmas in the Information Society Introduction and Overview. *Discrimination and Privacy in the Information Society*, Data Mining and Profiling in Large Databases, Springer.

Harding J. A., Shahbaz M., Srinivas, Kusiak A. (2006). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969–976. http://dx.doi.org/10.1115/1.2194554.

Huang, Y., Lu, Z., Hu, H., & Li, R. (2006). Privacy preserving distributed data mining association rules of frequent itemsets. *Jisuanji Gongcheng/Computer Engineering*, *32*(13), 12–14.

Indumathi, J., & Uma, G. V. (2009). A novel framework for optimised privacy preserving data mining using the innovative desultory technique. *International Journal of Computer Applications in Technology*, *35*(2–4), 194–203. https://doi.org/10.1504/IJCAT.2009.026596

Köksal Gülser, Batmaz İnci, Testik Murat Caner (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, *38*(10), 13448–13467. http://dx.doi.org/10.1016/j.eswa.2011.04.063.

Kaliappan, S. (2018). A hybrid clustering approach and random rotation perturbation (RRP) for privacy preserving data mining. *International Journal of Intelligent Engineering and Systems*, *11*(6), 167–176. https://doi.org/10.22266/IJIES2018.1231.17

Kao, Y.-H., Lee, W.-B., Hsu, T.-Y., Lin, C.-Y., Tsai, H.-F., & Chen, T.-S. (2015). Data perturbation method based on contrast mapping for reversible privacy-preserving data mining. *Journal of Medical and Biological Engineering*, *35*(6), 789–794. https://doi.org/10.1007/s40846-015-0088-6

Keqin, Wang, Shurong, Tong, Benoit, Eynard, & Roucoules, Lionel (2007). *Fuzzy Systems and Knowledge Discovery*, Volume: *4*,

Khedr, A. M., Osamy, W., Salim, A., & Salem, A.-A. (2019). Privacy preserving data mining approach for IoT based WSN in smart city. *International Journal of Advanced Computer Science and Applications*, *10*(8), 555–563.

Kumaran, U., & Khare, N. (2017). Feature selection for privacy preserving in data mining with linear regression using genetic algorithm. *Journal of Advanced Research in Dynamical and Control Systems*, *9*(Special Issue 2), 1059–1067.

Kumaraswamy S, S H Manjula, Venugopal K R (2017). Secure Cloud based Privacy Preserving DataMinning Platform. *Indonesian Journal of Electrical Engineering and Computer Science*, *7*(3), 830. http://dx.doi.org/10.11591/ijeecs.v7.i3.pp830-838.

Li, G., Wang, Y.-D., & Su, X.-H. (2010). Privacy preserving data mining on decision tree. *Tien Tzu Hsueh Pao/Acta Electronica Sinica*, *38*(1), 204–212.

Lin, C.-Y. (2016). A reversible data transform algorithm using integer transform for privacy-preserving data mining. *Journal of Systems and Software*, *117*, 104–112. https://doi.org/10.1016/j.jss.2016.02.005

Lindell, Y., & Pinkas, B. (2003). Privacy preserving data mining. *Journal of Cryptology*, *15*(3), 177–206. https://doi.org/10.1007/s00145-001-0019-2

Malvestuto, F. M., & Mezzini, M. (2004). Privacy preserving and data mining in an on-line statistical database of additive type. In J. Domingo-Ferrer & V. Torra (Eds.), *Privacy in Statistical Databases. PSD 2004. Lecture Notes in Computer Science* (Vol. 3050, pp. 353–365). Berlin: Springer.

Manikandan, G., Sairam, N., Sathya Priya, M., & Madhuri, S. R. (2015). A general critical review on privacy preserving data mining techniques. *Global Journal of Pure and Applied Mathematics*, *11*(4), 1899–1906.

Matatov, N., Rokach, L., & Maimon, O. (2010). Privacy-preserving data mining: A feature set partitioning approach. *Information Sciences*, *180*(14), 2696–2720. https://doi.org/10.1016/j.ins.2010.03.011

Nathiya, S., Kuyin C, Sundari j.D. (2016). Providing Multi Security In Privacy Preserving Data Mining. *International Journal Of Engineering And Computer Science*, http://dx.doi.org/10.18535/ijecs/v4i12.50.

Needhu, C., & Pothumani, S. (2014). An emblematic study of privacy preserving data mining using cryptographic techniques. *International Journal of Applied Engineering Research*, *9*(22), 5833–5840.

Panteli, A., Maragoudakis, M., & Gritzalis, S. (2014). Privacy preserving data mining using radial basis functions on horizontally partitioned databases in the malicious model. *International Journal on Artificial Intelligence Tools*, *23*(5), 1450007. https://doi.org/10.1142/S0218213014500079

Rajalakshmi, M., & Purusothaman, T. (2011). Privacy preserving distributed data mining using randomized site selection. *European Journal of Scientific Research*, *64*(4), 610–624.

Rajalakshmi, V., & Anandha Mala, G. S. (2014). Anonymization by data relocation using sub-clustering for privacy preserving data mining. *Indian Journal of Science and Technology*, *7*(7), 975–980.

Saad Hamza (2018). The Application of Data Mining in the Production Processes. *Industrial Engineering*, *2*(1), 26. http://dx.doi.org/10.11648/j.ie.20180201.14.

Shah, D., & Zhong, S. (2007). Two methods for privacy preserving data mining with malicious participants. *Information Sciences*, *177*(23), 5468–5483. https://doi.org/10.1016/j.ins.2007.07.013

Sun Zhaohao, Strang Kenneth David, Pambel Francisca (2020). Privacy and security in the big data paradigm. *Journal of Computer Information Systems*, *60*(2), 146–155. http://dx.doi.org/10.1080/08874417.2017.1418631.

Tian, H., Zhang, W., Xu, S., & Sharkey, P. (2012). A knowledge model sharing based approach to privacy-preserving data mining. *Transactions on Data Privacy*, *5*(2), 433–467.

Ur-Rahman N., Harding J.A. (2012). Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*, *39*(5), 4729–4739. http://dx.doi.org/10.1016/j.eswa.2011.09.124.

Yadav, Shivali, & Yadav, K. P. (2015). Case Study on Data Mining Security Issues and Remedies in Privacy Preservation. *International Journal of Information Technology and Management*, *IX*(XIV),

Yu, P. S., & Aggarwal, C. (2004). A Condensation Approach to Privacy Preserving Data Mining Conference Paper in Lecture Notes in Computer Science, March 2004. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-540-24741-8

Zhu, D., Li, X.-B., & Wu, S. (2009). Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. *Decision Support Systems*, *48*(1), 133–140. https://doi.org/10.1016/j.dss.2009.07.003

REFERENCES

Afzali, G. A., & Mohammadi, S. (2018). Privacy preserving big data mining: Association rule hiding using fuzzy logic approach. *IET Information Security*, *12*(1), 15–24. https://doi.org/10.1049/iet-ifs.2015.0545

Agrawal, D., & Aggarwal, C. C. (2001). *On the design and quantification of privacy preserving data mining algorithms*. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 247–255. https://doi.org/10.1145/375551.375602

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. *ACM SIGMOD Record*, *29*(2), 439–450.

Bellis, N. D. (2016). *Leadership Lessons for Health Care Providers*. Amsterdam: Elsevier. https://doi.org/10.1016/c2014-0-00967-7

Bourvil, & Levi. (2017). Multi-level trust privacy preserving data mining to enhance data security and prevent leakage of the sensitive data. *Bonfring International Journal of Industrial Engineering and Management Science*, *7*(2), 21–25. https://doi.org/10.9756/bijiems.8327

Bradić-Martinović, A., & Zdravković, A. (2013). Zaštita privatnosti—Anonimizacija podataka. *IASSIST Quarterly*, *2013*, 206–213.

Brickell, J., & Shmatikov, V. (2008). *The cost of privacy: Destruction of data-mining utility in anonymized data publishing*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; pp. 70–78. https://doi.org/10.1145/1401890.1401904

Chamikara, M. A. P., Bertok, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing*, *48*(May), 1–19. https://doi.org/10.1016/j.pmcj.2018.05.003

Chris, C., & Marks, D. (2008). Security and privacy implications of data mining. *Knowledge and Information Systems*, *14*(2), 161–178. https://doi.org/10.1007/s10115-007-0073-7

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. *ACM SIGKDD Explorations Newsletter*, *4*(2), 28–34. https://doi.org/10.1145/772862.772867

Diffie, W., & Hellman, M. E. (1976). New directions in croptography. *IEEE Transactions on Information Theory*, *22*(6), 159. https://doi.org/10.1007/3-540-44709-1_14

Domadiya, N. H., Kumar, A., & Rao, U. P. (2019). Improving healthcare using privacy preserving association rule mining in distributed healthcare data. *International Journal of Engineering and Advanced Technology*, *8*(4), 592–596.

Duan, Y., & Canny, J. (2009). How to deal with malicious users in privacy-preserving distributed data mining. *Statistical Analysis and Data Mining*, *2*(1), 18–33. https://doi.org/10.1002/sam.10029

Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2004). Privacy preserving mining of association rules. *Information Systems*, *29*(4), 343–364. https://doi.org/10.1016/j.is.2003.09.001

Fetscherin, M., Voss, H., & Gugler, P. (2010). 30 years of foreign direct investment to China: An interdisciplinary literature review. *International Business Review*, *19*(3), 235–246. https://doi.org/10.1016/j.ibusrev.2009.12.002

Fukasawa, T., Wang, J., Takata, T., & Miyazaki, M. (2004). An effective distributed privacy-preserving data mining algorithm. In Z. R. Yang, H. Yin, & R. M. Everson (Eds.), *Intelligent Data Engineering and Automated Learning – IDEAL 2004. IDEAL 2004. Lecture Notes in Computer Science* (Vol. 3177, pp. 320–325). Berlin, Heidelberg: Springer.

Garg, S., & Chhinkaniwala, H. (2011). *Privacy Preserving Data Mining Techniques: Challenges & Issues*. International Conference on Computer Science & Information Technology. p. 2014. https://doi.org/10.13140/RG.2.1.1811.3526

Goldwasser, S. (1997). *New directions in cryptography: Twenty some years later (or cryptography and complexity theory: A match made in heaven)*. Annual Symposium on Foundations of Computer Science - Proceedings, 314–324. https://doi.org/10.1109/sfcs.1997.646120

Gupta, M., & Joshi, R. C. (2009). Privacy preserving fuzzy association rules hiding in quantitative data. *International Journal of Computer Theory and Engineering*, *1*(4), 382–388. https://doi.org/10.7763/ijcte.2009.v1.60

Herman, B., & Custers, M. (2004). *The power of knowledge: ethical, legal, and technological aspects of data mining and group profiling in epidemiology*. Oisterwijk, Netherlands: Wolf Legal Publishers.

Huang, Z., Du, W., & Chen, B. (2005). *Deriving private information from randomized data*. Proceedings of the ACM SIGMOD International Conference on Management of Data, 37–48. https://doi.org/10.1145/1066157.1066163

Hussien, A. A., Hamza, N., & Hefny, H. A. (2013). Attacks on Anonymization-based privacy-preserving: A survey for data mining and data publishing. *Journal of Information Security*, *04*(02), 101–112. https://doi.org/10.4236/jis.2013.42012

Ilavarasi, A. K., & Sathiyabhama, B. (2017). An evolutionary feature set decomposition based anonymization for classification workloads: Privacy preserving data mining. *Cluster Computing*, *20*(4), 3515–3525. https://doi.org/10.1007/s10586-017-1108-9

Johnsana, J. S. A., Rajesh, A., Sangeetha, S., & Kishore Verma, S. (2016). Value and pattern anonymization of time series data for privacy preserving data mining. *Journal of Chemical and Pharmaceutical Sciences*, *9*(4), 2221–2228.

Kalyani, G., Rao, M. V. P. C. S., & Janakiramaiah, B. (2018). Privacy-preserving classification rule Mining for Balancing Data Utility and Knowledge Privacy Using Adapted Binary Firefly Algorithm. *Arabian Journal for Science and Engineering*, *43*(8), 3903–3925. https://doi.org/10.1007/s13369-017-2693-x

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering*, *16*(9), 1026–1037. https://doi.org/10.1109/TKDE.2004.45

Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2005). Random-data perturbation techniques and privacy-preserving data mining. *Knowledge and Information Systems*, *7*(4), 387–414. https://doi.org/10.1007/s10115-004-0173-6

Kaushal, A., & Shukla, M. (2014). Comparative analysis to highlight pros and cons of data mining techniques-clustering. *Neural Network and Decision Tree.*, *5*(1), 651–656.

Kavitha, G., & Elango, N. M. (2017). An overview of data mining techniques and its applications. *International Journal of Civil Engineering and Technology*, *8*(12), 1013–1020.

Kawano, A., Honda, K., Kasugai, H., & Notsu, A. (2013). A greedy algorithm for k-member co-clustering and its applicability to collaborative filtering. *Procedia Computer Science*, *22*, 477–484. https://doi.org/10.1016/j.procs.2013.09.126

Kim, J., & McMillan, S. J. (2008). Evaluation of internet advertising research: A bibliometric analysis of citations from key sources. *Journal of Advertising*, *37*(1), 99–112. https://doi.org/10.2753/JOA0091-3367370108

Krauskopf, E. (2018). A bibiliometric analysis of the journal of infection and public health: 2008–2016. *Journal of Infection and Public Health*, *11*(2), 224–229. https://doi.org/10.1016/j.jiph.2017.12.011

Kumar, P., Varma, K. I., & Sureka, A. (2011). Fuzzy based clustering algorithm for privacy preserving data mining. *International Journal of Business Information Systems*, *7*(1), 27–40. https://doi.org/10.1504/IJBIS.2011.037295

Li, G., & Xue, R. (2018). A new privacy-preserving data mining method using non-negative matrix factorization and singular value. *Wireless Personal Communications*, *102*, 1799–1808. https://doi.org/10.1007/s11277-017-5237-5

Liew, C. K., Choi, U. J., & Liew, C. J. (1985). A data distortion by probability distribution. *ACM Transactions on Database Systems (TODS)*, *10*(3), 395–411. https://doi.org/10.1145/3979.4017

Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. In M. Bellare (Ed.), *Advances in Cryptology—CRYPTO 2000. Lecture notes in computer science* (Vol. 249, pp. 517–524). Berlin: Springer. https://doi.org/10.1007/978-3-319-03095-1_55

Lindell, Y., & Pinkas, B. (2002). Privacy preserving data mining. *Research Journal of Applied Sciences, Engineering and Technology*, *9*(8), 616–621. https://doi.org/10.19026/rjaset.9.1445

Liu, K., Giannella, C., & Kargupta, H. (2008). A survey of attack techniques on privacy-preserving data perturbation methods. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-Preserving Data Mining. Advances in Database Systems* (Vol. 34, pp. 359–381). Boston, MA: Springer. https://doi.org/10.1007/978-0-387-70992-5_15

Liu, K., Kargupta, H., & Ryan, J. (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on Knowledge and Data Engineering*, *18*(1), 92–106. https://doi.org/10.1109/TKDE.2006.14

Manikandan, V., Porkodi, V., Mohammed, A. S., & Sivaram, M. (2018). Privacy preserving data mining using threshold based fuzzy Cmeans clustering. *ICTACT Journal on Soft Computing*, *6956*(October), 1813–1816. https://doi.org/10.21917/ijsc.2018.0252

Mendes, R., & Vilela, J. P. (2017). Privacy-preserving data mining: Methods, metrics. *And Applications. IEEE Access*, *5*(March), 10562–10582. https://doi.org/10.1109/ACCESS.2017.2706947

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Grp, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement (reprinted from annals of internal medicine). *Physical Therapy*, *89*(9), 873–880. https://doi.org/10.1371/journal.pmed.1000097

Paillier, P. (1999). *Public-Key Cryptosystems Based on Composite Degree Residuosity Classes*. Proceedings of the BT—Advances in Cryptology—EUROCRYPT '99, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, May 2–6, 1999. pp. 223–238. https://doi.org/10.1007/3-540-48910-X_16

Patel, D. J., & Swati, P. (2015). A survey on data perturbation techniques for privacy preserving in data mining. *International Journal for Scientific Research & Development*, *3*(01), 52–54.

Patel, N., & Patel, S. (2015). A study on data perturbation techniques in privacy preserving data mining. *International Research Journal of Engineering and Technology*, *2*, 2120–2124.

Persson, O., Danell, R., & Schneider, J. W. (2009). *How to use Bibexcel for various types of bibliometric analysis*. Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday, pp. 9–24. Retrieved from http://lup.lub.lu.se/record/1458990/file/1458992.pdf#page=11.

Poovammal, E., & Ponnavaikko, M. (2010). Utility independent privacy preserving data mining -horizontally partitioned data. *Data Science Journal*, *9*(July), 62–72. https://doi.org/10.2481/dsj.008-040

Qi, X., & Zong, M. (2012). An overview of privacy preserving data mining. *Procedia Environmental Sciences*, *12*(Icese 2011), 1341–1347. https://doi.org/10.1016/j.proenv.2012.01.432

Rajesh, Sujatha, & Arul. (2016). Survey on privacy preserving data mining techniques using recent algorithms. *International Journal of Computer Applications*, *133*(7), 30–33. https://doi.org/10.5120/ijca2016907917

Revathi, T. (2017). Data privacy preservation using data perturbation. *International Journal of Soft Computing and Artificial Intelligence*, *2*, 10–12.

Sah, H. R., & Gunasekaran, G. (2014). Privacy preserving collaborative data mining using steganography and encryption. *Journal of Theoretical and Applied Information Technology*, *68*(3), 411–415.

Samarati, P. (2001). Protecting respondents' identities in micro-data release. *IEEE Transactions on Knowledge and Data Engineering*, *13*(6), 1010–1027.

Samarati, P., & Sweeney, L. (1984). Generalizing data to provide anonymity when disclosing information. *Urology*, *23*(3), 27–28. https://doi.org/10.1016/S0090-4295(84)80111-9

Schlitter, N. (2008). A Protocol for Privacy Preserving Neural Network Learning on Horizontally Partitioned Data. Proc. Privacy Statistics in Databases (PSD '08). Retrieved from: http://www.nicoschlitter.de/downloads/Schlitter_PSD2008.pdf

Shah, A., & Gulati, R. (2016). Privacy preserving data mining: Techniques, classification and implications—A survey. *International Journal of Computer Applications*, *137*(12), 40–46. https://doi.org/10.5120/ijca2016909006

Shah, S. H. H., Lei, S., Ali, M., Doronin, D., & Hussain, S. T. (2019). Prosumption: Bibliometric analysis using HistCite and VOSviewer. *Kybernetes*, *49*(3), 1020–1045.

Sharma, A., & Ojha, V. (2010). Implementation of cryptography for privacy preserving data mining. *International Journal of Database Management Systems*, *2*(3), 57–65. https://doi.org/10.5121/ijdms.2010.2306

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Jr., Ke C. L. (2018). Data Mining for Business Analytics.

Silva, H. D. O., Basso, T., & Moraes, R. L. D. O. (*2017*). Privacy and Data Mining: Evaluating the Impact of Data Anonymization on Classification Algorithms. Proceedings - 2017 13th European Dependable Computing Conference, EDCC 2017, 111–116. https://doi.org/10.1109/EDCC.2017.17

Sweeney, L. (2002). A model for protecting privacy. *International Journal of Uncertainty, Puzziness and Knowledge-Based Systems*, *10*(5), 557–570.

Vaghashia, H., & Ganatra, A. (2015). A survey: Privacy preservation techniques in data mining. *International Journal of Computer Applications*, *119*(4), 20–26. https://doi.org/10.5120/21056-3704

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *SIGMOD Record*, *33*(1), 50–57. https://doi.org/10.1145/974121.974131

Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., & Dasseni, E. (2004). Association rule hiding. *IEEE Transactions on Knowledge and Data Engineering*, *16*(4), 434–447. https://doi.org/10.1109/TKDE.2004.1269668

Wahono, R. S. (2015). A systematic literature review of software defect prediction: Research trends, datasets, methods and frameworks. *Journal of Software Engineering*, *1*(1), 1–16.

Wang, N., Liang, H., Jia, Y., Ge, S., Xue, Y., & Wang, Z. (2016). Cloud computing research in the IS discipline: A citation/co-citation analysis. *Decision Support Systems*, *86*, 35–47. https://doi.org/10.1016/j.dss.2016.03.006

Wang, R., Zhu, Y., Chen, T. S., & Chang, C. C. (2018). An authentication method based on the turtle shell algorithm for privacy-preserving data mining. *Computer Journal*, *61*(8), 1123–1132. https://doi.org/10.1093/comjnl/bxy024

Wu, L.-F., & Xiao, J. (2013). A privacy preserving data mining scheme based on network user's behavior. *Journal of Theoretical and Applied Information Technology*, *47*(2), 671–677.

Xu, L., Jiang, C., Wang, J., Yuan, J., & Ren, Y. (2014). Information security in big data: Privacy and data mining. *IEEE Access*, *2*, 1151–1178. https://doi.org/10.1109/ACCESS.2014.2362522

Xu, Z. (2011). Analysis of privacy preserving distributed data mining protocols. Victoria, 1–85. Retrieved from http://vuir.vu.edu.au/16047/1/Zhuojia_Xu_Masters.pdf.

Yao, A. C. C. (1986). How to generate and exchange secrets. *Annual Symposium on Foundations of Computer Science (Proceedings)*, *1*, 162–167. https://doi.org/10.1109/sfcs.1986.25

Zadeh, L. A. (2015). Fuzzy logic - a personal perspective. *Fuzzy Sets and Systems*, *281*, 4–20. https://doi.org/10.1016/j.fss.2015.05.009

Zhan, J., & Matwin, S. (2007). Privacy-preserving data mining in electronic surveys. *International Journal of Network Security*, *4*(3), 318–327.

How to cite this article: Kreso I, Kapo A, Turulja L. Data mining privacy preserving: Research agenda. *WIREs Data Mining Knowl Discov*. 2020;e1392. https://doi.org/10.1002/widm.1392

APPENDIX A

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
The most frequently used techniques in PPDM	Agrawal and Srikant (2000)	Privacy-preserving data mining	Technical feasibility of realizing privacy-preserving data mining The basic premise was that the sensitive values in a user's record will be perturbed using a randomizing function so that they cannot be estimated with sufficient precision.	• Focus on the development of techniques that would incorporate privacy concerns. • Focus on developing the accurate aggregated data models that would avoid accessing the sensitive data from the individual data records. • Focus on the effectiveness of PPDM method randomization with reconstruction for categorical attributes.	Privacy-preserving data mining
	Agrawal and Aggarwal (2001)	On the design and quantification of privacy preserving data mining algorithms	This article discusses an expectation maximization (EM) algorithm for distribution reconstruction which is more elective than the currently available method in terms of the level of information loss. Specifically, we prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. We show that when a large amount of data is available, the EM algorithm provides robust estimates of the original distribution. We propose metrics for quantification and measurement of privacy-preserving data mining algorithms. Thus, this article provides the foundations for measurement of the effectiveness	• Testing the effectiveness of the PPDM algorithms using the proposed privacy metrics that was based on mutual information between the original and perturbed records.	Algorithms, experimentation, theory

(Continues)

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
			of privacy-preserving data mining algorithms.		
	Evfimievski et al. (2004)	Privacy preserving Mining of Association Rules	Present a framework for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. While	• What are the theoretical limits on discoverability for a given level of privacy (and vice versa)? • Extend the proposed approach to encompass not only restricted classes, but also kinds of breaches and other assumptions on the dataset. • Examine the possibility of combining randomization and cryptographic protocols to get better results.	Privacy-preserving data mining, association rule
	Kantarcioglu and Clifton (2004)	Privacy-preserving distributed Mining of Association Rules on horizontally partitioned data	This article addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task.	• Focus researches on the new secure algorithms for classifications, clustering, etc. • Examine secure approximate data mining algorithms. • Predicting the value of information for a particular organization, allowing tradeoff between disclosure cost, computation cost, and benefit from the result.	Data mining, security, privacy
	Lindell and Pinkas (2000)	Privacy-preserving data mining	Paper addresses the issue of privacy preserving data mining. Specifically, paper considers a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information.	• Optimize honest parties in terms of finding the most efficient solution to the malicious party.	Class attribute cryptographic protocol oblivious transfer small circuit evaluation

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
	Paillier (1999)	Public-key cryptosystems based on composite degree Residuosity classes	This article investigates a novel computational problem, namely the composite Residuosity class problem, and its applications to public-key cryptography. We propose a new trapdoor mechanism and derive from this technique three encryption schemes: a trapdoor permutation and two homomorphic probabilistic encryption schemes computationally comparable to RSA	• Possible modifications when it comes to proposed schemes based on cryptography. • Exploring the homomorphic properties of the proposed systems to design distributed cryptographic protocols (multi-signature, secret sharing, threshold cryptography, and more).	Cryptography
	Verykios, Elmagarmid, et al. (2004)	Association rule hiding	Paper presents three strategies and five algorithms for hiding a group of association rules, which is characterized as sensitive. One rule is characterized as sensitive if its disclosure risk is above a certain privacy threshold. Sometimes, sensitive rules should not be disclosed to the public since, among other things, they may be used for inferring sensitive data, or they may provide business competitors with an advantage	• Expanding the proposed algorithms in the study on the broader data mining context.	Privacy-preserving data mining, association rule mining, sensitive rule hiding
	Yao (1986)	How to generate and exchange secrets	In this article, we introduce a new tool for controlling the knowledge transfer process in cryptographic protocol design	• Focus on tradeoffs among complexity measures.	Cryptographic

(Continues)

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
Different methods for PPDM	Agrawal and Srikant (2000)	Privacy-preserving data mining	Technical feasibility of realizing privacy-preserving data mining The basic premise was that the sensitive values in a user's record will be perturbed using a randomizing function so that they cannot be estimated with sufficient precision.	• Focus on the development of techniques that would incorporate privacy concerns. • Focus on developing the accurate aggregated data models that would avoid accessing the sensitive data from the individual data records. • Focus on the effectiveness of PPDM method randomization with reconstruction for categorical attributes.	Privacy-preserving data mining
	Huang et al. (2005)	Information from randomized data	This study proposes a modified randomization scheme, in which we let the correlation of random noises "similar" to the original data. Our results have shown that the reconstruction accuracy of both PCA-based and BE-based schemes becomes worse as the similarity increases.	• Focus on examining Bayes estimate using numerical methods (gradient descent methods)	Privacy-preserving data mining, randomization, PCA, Bayes estimate.
	Lindell and Pinkas (2002)	Privacy-preserving data mining	We focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient than generic solutions and demands both very few rounds of communication and reasonable bandwidth.	• Optimize semi-hones parties in terms of finding the most efficient solution to the malicious party.	Secure two-party computation, oblivious transfer, oblivious polynomial evaluation, data mining, decision trees.
	Liu et al. (2006)	Random projection-based multiplicative data perturbation for privacy-preserving distributed data mining	This article explores the possibility of using multiplicative random projection matrices for privacy-preserving distributed data mining. Experiments demonstrate that the proposed technique is effective and can be successfully used for different types of privacy-	• Since there is still no efficient way to quantify the privacy and data utility, so the study suggests focusing on the computing the maximum mutual information between the original signals and the received	Random projection, multiplicative data perturbation, privacy-preserving data mining.

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
			preserving data mining applications.	signals in a multiplicative model that is actually work related to the model proposed in the paper.	
	Samarati (2001)	Protecting Respondents' identities in microdata release	Paper illustrates how k-anonymity can be provided without compromising the integrity (or truthfulness) of the information released by using generalization and suppression techniques. We introduce the concept of minimal generalization that captures the property of the release process not to distort the data more than needed to achieve k-anonymity, and present an algorithm for the computation of such a generalization.	• Investigating the efficient algorithms to enforce the proposed techniques. • Focus on updating and modifying the stored data. • Focus on multiple releases and consequent possibilities of collusions by multiple recipients or by the same recipients through multiple queries • Explore the application of the techniques at the finer granularity level of cell • Investigate the new techniques for k-anonymity. • Focus on developing a model for the specification of protection requirements of the data with respect to different possible classes of data recipients.	Privacy, data anonymity, disclosure control, microdata release, inference, record linkage, security, information protection.
	Sweeney (2002)	k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY	This article includes a formal protection model named fc-anonymity and a set of accompanying policies for deployment. A release provides fc-anonymity protection if the information for each person contained in the release cannot be distinguished from at least k-\individuals whose information also appears in the release.	No suggestions on future research.	Data anonymity; data privacy; re-identification; data fusion; privacy

(Continues)

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
Key strategies, steps, and terms for more efficient PPDM	Adam & Worthmann, (1989)	Security-control methods for statistical databases: A comparative study	Paper considers the problem of providing security to statistical databases against disclosure of confidential information.	• Focus on developing new methods that prevent exact disclosure and provide statistical-disclosure control and are not affected by the bias.	Security, integrity, protection
	Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios (1999)	Disclosure limitation of sensitive rules	Deals with the problem of limiting disclosure of sensitive rules. In particular, it is attempted to selectively hide some frequent itemsets from large databases with as little as possible impact on other, non-sensitive frequent itemsets	• Evaluate different selection criteria and impact on the set of rules that are marked as non-sensitive. • Examine optimization possibilities to both time and space complexity of the proposed. • Investigate application of the proposed solutions on classification rule mining, correlation rule mining, etc.	Association rules
	Liew et al. (1985)	A data distortion by probability distribution	This article presents a data distortion by probability distribution that requires three steps: (1) identification of the underlying density function, (2) generation of a distorted series from the density function, and (3) mapping of the distorted series onto the original series.	No suggestions on the future research.	Management, security compromisability, point distortion, probability distortion, individual privacy, statistical database, microdata tile
	Verykios, Elmagarmid, et al. (2004)	Association rule hiding	Paper presents three strategies and five algorithms for hiding a group of association rules, which is characterized as sensitive. One rule is characterized as sensitive if its disclosure risk is above a certain privacy threshold. Sometimes, sensitive rules should not be disclosed to the public since, among other things, they may be used for inferring sensitive data, or they may provide business competitors with an advantage	• Expanding the proposed algorithms in the study on the broader data mining context.	Privacy preserving data mining, association rule mining, sensitive rule hiding

Cluster	Author	Publication title	Main findings	Future research recommendation	Keywords
	Xu et al. (2014)	Information security in big data: Privacy and data mining	Paper explored the privacy issues related to data mining from a wider perspective and investigate various approaches that can help to protect sensitive information. In particular, we identify four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. F	• Focus on personalized privacy preserving. • Focus on data customization. • Explore provenance for data mining. • How the data provider can discover the unwanted disclosure of his sensitive information as early as possible? • How to formulate personalized privacy preference in a more flexible way? • How to obtain such preference with less effort paid by data providers? • Focus on new types of data mining applications emerge, finding appropriate ways to quantify privacy and utility.	Data mining, sensitive information, privacy-preserving data mining, anonymization, provenance, game theory, privacy auction, anti-tracking.

APPENDIX B

Studies and authors	Research recommendation from the studies
Agrawal and Srikant (2000)	**Focus on the development of techniques that would incorporate privacy concerns**
Verykios, Elmagarmid, et al. (2004)	Expanding the proposed algorithms in the study on the broader data mining context.
Paillier (1999); Samarati (2001)	Possible modifications when it comes to propose schemes based on cryptography.
Kantarcioglu and Clifton (2004)	Focus researches on the new secure algorithms for classifications, clustering, etc.
Adam and Worthmann (1989)	Investigate the new techniques for k-anonymity. Focus on developing a model for the specification of protection requirements of the data with respect to different possible classes of data recipients.
Verykios, Elmagarmid, et al. (2004)	Focus on developing new methods that prevent exact disclosure and provide statistical-disclosure control and are not affected by the bias.
Xu et al. (2014)	Expanding the proposed algorithms in the study on the broader data mining context
	How to formulate personalized privacy preference in a more flexible way? Focuses on new types of data mining applications emerge, finding appropriate ways to quantify privacy and utility.
Chamikara, Bertok, Liu, Camtepe, and Khalil (2018)	Further studies on increasing the efficiency of P2RoCAl using sampling techniques and parallel implementations can be investigated. This would allow P 2RoCAl to work with high-speed data streams
R. Wang et al. (2018)	Future work can investigate other data hiding techniques to achieve a higher level of privacy preservation and speed the execution time of perturbation. In addition, techniques for preserving the information of categorical attributes are under development
Kalyani et al. (2018)	A new algorithm is projected for defending the sensitive classification rules from disclosure.
Lindell and Pinkas (2000)	Optimize semi-hones parties in terms of finding the most efficient solution to the malicious party.
Samarati (2001)	Investigating the efficient algorithms to enforce the proposed techniques.
Atallah, Bertino, Elmagarmid, Ibrahim, & Verykios (1999)	Examine optimization possibilities to both time and space complexity of the proposed
Afzali and Mohammadi (2018)	As future work, we will try to decrease undesired side effects of the proposed model to gain less information loss.
Mendes and Vilela (2017)	Fully homomorphic encryption is suggested has a future solution to this limitation, however, the computational cost is currently prohibitive.
Ilavarasi and Sathiyabhama (2017)	Examining the algorithm with other data mining tasks such as clustering and association rules. Extend SADM for sequential partitioning structures. Revising the algorithm to function as an ensemble classifier. Extend SADM to machine learning with big data. When a partition does not comply with k-anonymity, the algorithm excludes it from producing offspring. This over-anonymity issue has to be addressed.
Li and Xue (2018)	In the future, we want to study the action mechanisms of different privacy protection algorithm to find an optimized combination to maximize privacy protection.

Studies and authors	Research recommendation from the studies
Agrawal and Srikant (2000)	**Focus on the development of techniques that would incorporate privacy concerns**
Liu et al. (2006)	Since there is still no efficient way to quantify the privacy and data utility, so the study suggests focusing on the computing the maximum mutual information between the original signals and the received signals in a multiplicative model that is actually work related to the model proposed in the paper.
Xu et al., 2014	Focus on new types of data mining applications emerges, finding appropriate ways to quantify privacy and utility.
Kalyani et al. (2018)	The future work on this can be extended by considering the data set with more number of values and in numerical format