Data Science 0 (0) 1
IOS Press

A Systematic Review on Privacy-Preserving Distributed Data Mining

Chang Sun [a,*], Lianne Ippel [a], Andre Dekker [b], Michel Dumontier [a], Johan van Soest [b]
[a] *Institute of Data Science, Maastricht University, Maastricht, The Netherlands*
[b] *Department of Radiation Oncology (MAASTRO), GROW School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands*

Abstract. Combining and analysing sensitive data from multiple sources offers considerable potential for knowledge discovery. However, there are a number of issues that pose problems for such analyses, including technical barriers, privacy restrictions, security concerns, and trust issues. Privacy-preserving distributed data mining techniques (PPDDM) aim to overcome these challenges by extracting knowledge from partitioned data while minimizing the release of sensitive information. This paper reports the results and findings of a systematic review of PPDDM techniques from 231 scientific articles published in the past 20 years. We summarize the state of the art, compare the problems they address, and identify the outstanding challenges in the field. This review identifies the consequence of the lack of standard criteria to evaluate new PPDDM methods and proposes comprehensive evaluation criteria with 10 key factors. We discuss the ambiguous definitions of privacy and confusion between privacy and security in the field, and provide suggestions of how to make a clear and applicable privacy description for new PPDDM techniques. The findings from our review enhance the understanding of the challenges of applying theoretical PPDDM methods to real-life use cases, and the importance of involving legal-ethical and social experts in implementing PPDDM methods. This comprehensive review will serve as a helpful guide to past research and future opportunities in the area of PPDDM.

Keywords: Survey, Data mining, Privacy preserving, Distributed learning

1. Introduction

Mining distributed, sensitive data offers tantalising potential for new insights and a wide variety of applications, but is generally fraught with concerns of model accuracy and data privacy. Consider the case of analyzing patient data in the healthcare domain: hospitals have used patient data to improve diagnostic accuracy and efficiency [1, 2] and to fuel the transition to preventive [3] and precision medicine [4–6]. However, learning patient data from a single hospital might cause limited model performance and incomplete knowledge discovery [7]. Patients' health are not only affected by genetic and biological factors, but also by individual behaviour and social circumstances [8]. Combining various patient data from multiple sources offers one pathway to obtain more accurate and reliable analytical models for health outcomes [9, 10]. However, combining distributed sensitive data faces a number of challenges including: data protection compliance to one or more legal jurisdictions, privacy concerns, security, and trust issues. Beyond the healthcare domain, this also applies to applications in many other fields, such as finance and law [11, 12]. Conventional centralised data mining techniques are challenged in this environment and require viable alternatives.

*Corresponding author. E-mail: chang.sun@maastrichtuniversity.nl; ORCID: https://orcid.org/0000-0001-8325-8848.

Privacy-preserving distributed data mining (PPDDM), which focuses on the analysis of decentralised data without leaking sensitive information from any party to the other parties, offers one way forward for multiple data parties to overcome the challenges posed by centralising the data for analysis [13]. PPDDM techniques, whether data mining or machine learning, aim to make it technically or mathematically infeasible to deduce the original data from a communication message, and certainly from the final analysis result. To make use of PPDDM in practical applications, we should consider the data problems (e.g., classification, regression), the adversarial concerns the involving data parties have (e.g., malicious, honest), and the balance between data privacy and model performance. PPDDM is sometimes referred to privacy-preserving federated learning after Google first proposed the concept in 2016 [14–16]. However, privacy-preserving federated learning can be regarded as a specific category of PPDDM, in which there is a federation of autonomous organisations that express an interest to contribute to a joint analysis [17].

A number of PPDDM methods have been reported in the last 20 years. The existing survey papers have compared the theoretical backgrounds, strengths, and limitations. However, the analysis of distributed data has been poorly addressed as only one special case of privacy-preserving data mining [18–21]. The distributed data problem has been addressed to a limited extent in the survey of Hina Vaghashia [22] and Suchitra Shelke [23]. Vassilios S. et al [21] presented five dimensions of state-of-the-art privacy-preserving data mining algorithms where the problem of analysing distributed data was merely considered to be addressed by cryptography-based techniques and only the association rule mining problem and decision tree induction were presented in this survey. Several surveys summarized the evaluation parameters to assess privacy-preserving techniques including privacy level, hiding failure, data quality, complexity, efficiency, and resistance of different data mining algorithms [19, 21, 24, 25]. Others have a major focus on the definition and construction of Secure Multiparty Computation (SMC) and how SMC can be combined with data mining algorithms [13, 26, 27]. In a recent survey [28], privacy-preserving approaches were summarized for data collection, data publishing, data mining output, and distributed learning. The majority of the published surveys have typically treated PPDDM as a specialised subtopic of either distributed data mining or privacy-preserving data mining. As an emerging field, PPDDM is under-reported in the existing surveys and now requires a more comprehensive and complete analysis.

Accordingly, the main aim of this systematic review is to provide an overview of existing approaches and identify outstanding challenges in the field of PPDDM. This paper reports the results and findings of a comprehensive review of PPDDM techniques from 231 scientific articles published in the past 20 years. We present the characteristics of the 18 most cited studies and analyze their influence on other studies in the field. The results show a wide range of privacy-preserving methods and data mining algorithms have been well-studied. We highlight the findings showing a lack of standard evaluation criteria in the field, the ambiguous definition of privacy, and insufficient experimental information in some studies. These findings enhance the understanding of the challenges of applying the theoretical PPDDM methods to real-life use cases, and the importance of involving legal-ethical and social experts in implementing PPDDM methods.

The main contributions of this work to the literature in the PPDDM field are:

(1) to propose comprehensive criteria with 10 key factors to evaluate the new PPDDM techniques. The evaluation criteria include adversarial behaviour of data parties, data partitioning, experiment datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability.

(2) to present different definitions of privacy, distinguish information privacy from information security in the PPDDM field, and provide suggestions of how to make clear and applicable privacy descriptions to propose new PPDDM techniques.

(3) to identify the most cited PPDDM articles, analyze their characteristics and how these articles influence other studies in the field, and
(4) to provide a guideline based on the proposed evaluation criteria for researchers to conduct future research and publications in the PPDDM field.

This systematic review offers new insights into the important factors that should be considered to propose and evaluate new PPDDM techniques and how to bridge the gap between theoretical methods and practical applications in the field. We present this review paper as a helpful guide to past research and future opportunities in the area of PPDDM.

The outline of this paper is as follows. In the next section, we present existing privacy-preserving methods and define terms related to PPDDM. In Section 3, we describe the approach in conducting this systematic review. In Section 4, we provide the results of our review, including evaluation criteria. In Section 5, we compare the key influential papers. In the last section, we summarize our main findings, present a list of recommendations, and discuss future directions.

2. Privacy-Preserving Methods

Privacy-preserving methods, as the major component of PPDDM techniques, are used to minimize the release of information during data mining model training and communication among multiple parties. Various privacy-preserving methods have been proposed from different communities such as statistics, cryptography, data mining, and secure data transfer. In this section, we summarize the most commonly-used privacy-preserving methods in PPDDM.

2.1. Secure Multiparty Computation (SMC)

Secure multiparty computation protocols are designed for multiple parties to jointly compute some function over their own data without revealing the original data to any other parties [13]. The foundation for SMC started from cryptography. In addition to protect the participants from being attacked by external parties (who are outside of the system or protocol), SMC also protects the participants from each other. For example, some SMC protocols are implemented to prevent participants from learning private information from other parties or deliberately sending incorrect computation results to other parties. The following sub-sections describe some well-known protocols in SMC.

2.1.1. Building Blocks (primitives) SMC of Protocols.

Secure protocols that are deployed as building blocks of secure computation are used to prevent data being revealed or deduced from the communication and/or computation between data parties [13]. Commonly used encryption protocols include oblivious transfer and homomorphic encryption. Oblivious transfer, first developed by Even et al. [29], considers two data parties, a requester and a sender, where the requester obtains exactly one instance without the sender knowing which element was queried, and without the requester knowing about the other instances that were not retrieved. Oblivious transfer protocols iteratively pass over the data many times during training, and as a result are computationally expensive. Another technique, homomorphic encryption, was introduced by Rivest [30]. This technique supports certain algebraic operations such as additions and multiplications on encrypted text (i.e., ciphertext). The decrypted result from the operations on ciphertext matches the result of the operations performed on the plain text. Homomorphic encryption systems are grouped into fully homomorphic encryption (FHE) or partial homomorphic encryption (PHE) [31]. As the initial scheme of a homomorphic cryptosystem, PHE can only perform a specific algebra operation such as addition or multiplication in each iteration. This limits the usability for data mining algorithms, as the algorithms consist of several complex operations. On the contrary, FHE supports any desirable operation and functionality that can run on the ciphertext. Since the ciphertext is never decrypted, the input from each data party is not revealed. The first generation of FHE system was proposed by Gentry in 2009 [32]. However, FHE systems are not sufficiently efficient due to the high computational cost of performing iterative operations over encrypted data during the training epochs.

2.1.2. Generic SMC Protocols.

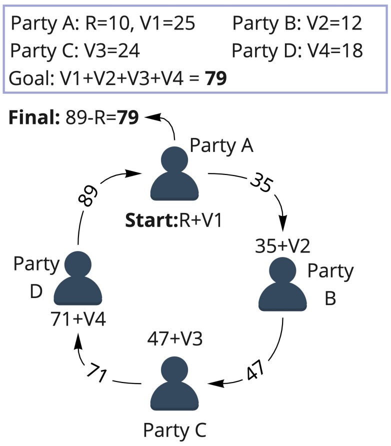
Generic SMC protocols were implemented for any probabilistic polynomial-time function [13]. Unlike homomorphic encryption systems, these generic protocols are sensitive to the number of data parties. The commonly-used protocol of secure two-party computation is Yao's garbled circuit protocol [33]. The protocol is based on evaluating the function that needs to be computed by two data parties as a combinatorial circuit with a collection of gates (e.g., AND, XOR gate). These gates connect with circuit-input wires, circuit-output wires and intermediate wires. Each gate has two input wires and one single output wire. The required communication of the protocol depends on the size of the circuit, while the computation cost depends on the number of input wires. Extensions to more than two data parties, i.e. the cases of multiparty computation, have been developed by Micali et al. [34], Beaver et al. [35], and Ben-Or et al. [36]. Following Yao's theory, these protocols are based on designing the function as a circuit and applying a secure computation protocol to the circuit [13]. Beside computational complexity, communication cost is a considerable factor in these protocols. All protocols need a one-to-one communication channel between every pair of parties. Some require a broadcast channel for all parties.

2.1.3. Specialized SMC Protocols.

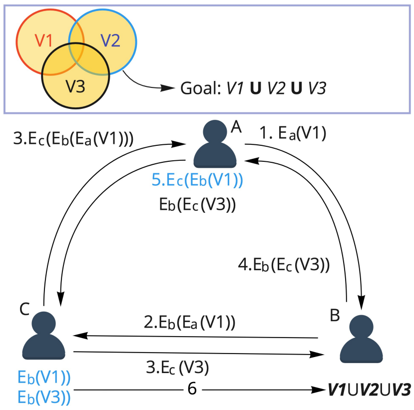
Specialized SMC protocols are commonly used as primitives to the data mining algorithms including secure sum, secure set union, secure size of intersection, and secure scalar product protocols. These protocols allow certain operations without revealing any inputs from any of the participating data parties.

Secure sum as a basic and simple example of secure multiparty computation was introduced by Clifton et al. to obtain the sum of the inputs [26]. The protocol is as follows: data party A has $V1$ local value. Party A generates a random number R and calculates $(R + V1)$ and sends this result to data party B (PB). Then, Party B adds their local value to the received value and sends it $(R + V1 + V2)$ to the next party. In the end, to obtain the final result, the last sum value will be sent back to party A to subtract R. The protocol ends with sending this final result to all participating parties. An example of securely computing a sum among 4 four parties is shown in Fig.1a.

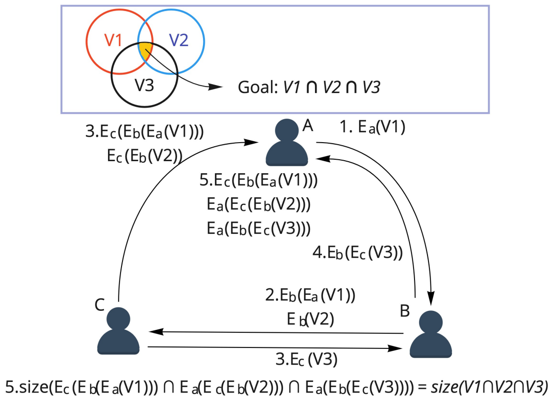
Secure set union has been applied to the case where data parties want to jointly create unions of sets from rules and itemsets shared by multiple parties but not leaking the owner of each set. To guarantee a secure computation, one approach is to apply a commutative encryption system in computing the set union [26, 37]. A commutative encryption system can encrypt original data multiple times using different users' public keys. The final encrypted data can be decrypted without considering the order of the public keys in the encryption process [38]. In the secure set union protocol, one data party encrypts its own itemsets using commutative encryption and transfers them to other parties. The receiver party encrypts both its own sets and the received encrypted sets and passes it to the next party. Once the data is encrypted by all parties, decryption can start at each party in any order. The permutation of the encryption order prevents the participating parties from tracking the ownership of itemsets. However, if one item is present at multiple data parties, then the number of the item will be exposed because of duplication. Fig.1b presents an example of securely computing a set union among three data parties.



(a) An example of secure computation of a sum among four parties. R is a random number generated by Party A. V1, V2, V3, and V4 presents the private data from party A to party D.



(b) An example of secure computation of a set union among three parties. Party A, B, C encrypts their private data using a commutative encryption scheme respectively (E_a, E_b, E_c). Text in blue is decrypted text.



(c) An example of secure computation of a size of set intersection among three parties. Party A, B, C encrypts their private data using a commutative encryption scheme respectively (E_a, E_b, E_c).

Fig. 1. Examples of three secure multiparty computation protocols - Fig.1a. Secure sum protocol, Fig.1b. Secure set union protocol, Fig.1c. Secure size of set intersection protocol.

Secure size of set intersection is solving the problem that multiple data parties want to obtain the size of set intersection of their local datasets without revealing the ownership. Similar to secure set union, each data party encrypts its own item sets by using commutative encryption and sends it to another data party. The receiver encrypts these items, arbitrarily permutes the order, and sends it to the next data party. This process ends when all item sets are encrypted by all data parties. Due to the commutative encryption, if and only if the original inputs are the same, then the final outcomes from two different item sets can be equal. Therefore, the number of values that occur in all encrypted item sets is the size of the set intersection. No input will get exposed since only encryption (no decryption) is required. Fig.1c demonstrates the protocol of securely computing the size of set intersection.

Secure scalar product protocols are essential and powerful. It has been widely applied in many data mining algorithms which can be decomposed to the calculation of scalar products. As a notable example, Vaidya and Clifton extended a secure scalar products protocol to solve association rule mining problems between two parties [39]. The general idea is as follows:

(1) Data party A has $X = \{x_1, ..., x_n\}$, while data party B has $Y = \{y_1, ..., y_n\}$. The goal is to calculate $X * Y = \sum_{i=1}^{n}(x_i * y_i)$ without revealing inputs to the other party. Both parties share a matrix C which is generated by random numbers.
(2) The protocol starts at Party A who generates n random numbers $Ra = \{r_1, ..., r_n\}$. Then, party A calculates $X' = X + C * Ra$ and send to party B.
(3) Party B generates $m(< n)$ random numbers Rb and calculate $Y' = C_1 * Y + Rb_1, \ldots, C_{n/m} * Y + Rb_1, ..., C_{2n/m} * Y + Rb_2, \ldots, C_n * Y + Rb_n$ and $S' = \sum_{i=1}^{n}(x_i' * y_i)$. Y' and S are sent to party A.
(4) Party A calculates $S'' = S' - \sum_{i=1}^{n}(Ra * Y')$ and m sets of sum of Ra which is $Ra' = Ra_1 + Ra_2 + \ldots + Ra_{n/m} + Ra_{n/m+1} + \ldots + Ra_{2n/m}, ..., Ra_{((m-1)n/m)+1} + Ra_{((m-1)n/m)+2} + \ldots + Ra_n$. Party A sends S'' and Ra' for final result calculation.
(5) Party B computes the final scalar product as $S = S'' + Ra' * Rb$.

The security of this secure scalar product protocol is guaranteed by the inability of either side to deduce k equations with more than k unknowns. As with many other existing scalar product protocols [40, 41], it is limited to the collaboration between only two parties because of the lack of efficiency in practice [26].

2.2. Data Perturbation

Data Perturbation preserves data privacy by adding ‘noise" to the individual records but still keeps the key summary information about the data [42]. One major approach of data perturbation is to use statistical techniques to replace the original data with synthetic values which have the same or comparable statistical information (e.g., distributions) as the original values. The synthetic data can be generated by a statistical model which learns from the original data. The other main approach is to distort the values by applying additive noise, multiplicative noise, or other randomization procedures [43]. Data swapping, another method of data perturbation, switches a set of (sensitive) attributes between different data entities to prevent the linkage of records to identities [44, 45]. The major drawback of these methods is the decrease of data quality and accuracy of the learning model. Data perturbation techniques are more commonly used to protect privacy in data publishing problems [28].

2.3. Local Learning and Global Integration

The method that integrates local models to one global model uses the foundation of ensemble learning that trains a set of models in order to enhance the performance of one single model [46, 47]. Each data party can train their own local data miners independently. Then, these local data miners are sequentially or parallelly integrated to compose a center or global data miner which can generate the final results. Consequently, the original data of each party is never transferred to other data parties. A majority of data mining algorithms have been theoretically developed to this approach including Support Vector Machine [48–51], Decision Tree [52–54], Neural Networks [51, 55–57] and so forth. A few of them have been successfully implemented, applied and evaluated in practical use cases such as [7] and [58].

3. Methodology

This paper follows the systematic review procedures described by Kitchenham [59]. In this section, we will detail the workflow. First, we discuss the inclusion and exclusion criteria of study selection, followed by the search strategies, and evaluation criteria for reviewing selected studies.

3.1. Eligibility Criteria

We selected papers that are peer-reviewed publications in English between 2000 and 2020 (August) working on data mining and machine learning techniques that solve problems of classification, regression, clustering, or association rule mining. The eligible papers must take privacy preservation into account when data mining and machine learning models are executed on partitioned data. Partitioned data includes horizontally partitioned/homogeneous data, vertically partitioned/heterogeneous data, and arbitrarily partitioned data (The definitions of different partitioned data are presented in section 3.3). Furthermore, included papers must 1) propose and/or implement a new approach and/or; 2) apply existing approaches to a practical case and/or; 3) improve the performance of existing approaches.

To narrow down the number of publications, we excluded poster and workshop abstracts, survey papers, and articles that only contain discussions on current concerns and future research directions. To set the scope of this survey, the authors screened titles, keywords, and abstracts to exclude the papers that 1) only focus on privacy-preserving data mining/machine learning on centralised data, 2) solve problems of parallel computing, cloud computing, grid computing, edge computing, and fog computing to improve computational performance rather than the complexity of the data analysis problem, 3) solve privacy issues in data collecting, data publishing, data storage, and data querying, and 4) focus on Blockchain, web attacks detection, intrusion detection, data privacy focusing on mobile devices, geographic data privacy, and differential privacy. If the papers could not be identified based on its title, keywords, and abstract, the authors reviewed the full text of the paper.

3.2. Search Strategy

According to the eligibility criteria above, we used the following search engines and digital libraries: IEEE Xplore Digital Library [1], ACM Digital Library [2], Science Direct [3], ISI Web of Science [4], Springer Link [5], PubMed [6]. Based on the inclusion criteria, we formulated the following terms to search in the title, abstract, and keywords of papers. The entire workflow for selecting relevant studies is presented with search results in Figure 2 in Section 4.1.

[1] IEEE Xplore: https://ieeexplore.ieee.org/Xplore/home.jsp/
[2] ACM Digital Library: https://dl.acm.org/
[3] ScienceDirect: https://www.sciencedirect.com/
[4] Web of Science - Clarivate: https://clarivate.com/products/web-of-science/
[5] Springer Link: https://link.springer.com/
[6] PubMed: https://www.ncbi.nlm.nih.gov/pubmed/

(1) *privacy* and (*distributed* or *de-centralized* or *de-centralised* or *partitioned*) and "*machine learning*" (**PPDML**)
(2) *privacy* and (*distributed* or *de-centralized* or *de-centralised* or *partitioned*) and "*data mining*" (**PPDDM**)
(3) *privacy* and (*vertically* or *heterogeneous*) and "*machine learning*" (**PPVML**)
(4) *privacy* and (*vertically* or *heterogeneous*) and "*data mining*" (**PPVDM**)
(5) *privacy* and (*horizontally* or *homogeneous*) and "*machine learning*" (**PPHML**)
(6) *privacy* and (*horizontally* or *homogeneous*) and "*data mining*" (**PPHDM**)

3.3. Evaluation Criteria for Reviewing Papers

To evaluate the paper on PPDDM techniques, conventional data mining evaluation criteria are not adequate [46]. Beside conventional evaluation methods, additional factors such as communication costs, data partitioning, adversary behavior, privacy measures should be considered. To the best of our knowledge, there are no standard criteria for evaluating new PPDDM approaches. Consequently, studies selected a various set of evaluation methods which they think are necessary for their approaches. In this review, we assessed selected papers considering the following 10 factors including adversarial behavior of data party, data partitioning, experimented datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability. The authors initially generated and modified these evaluation criteria by reviewing 10% of the included articles. Then, the evaluation criteria have been discussed by the co-authors in several iterations of reviewing until an agreement has been made on these 10-factor evaluation criteria. Afterwards, all selected papers have been reviewed and assessed again using the criteria.

1) Adversarial behavior of data parties covers the assumed adversarial behavior that involved data parties have. In this review, we consider two types of adversarial behavior of involved parties - semi-honest and malicious. A semi-honest (also called passive, or honest-but-curious) party follows the protocol properly, however is also curious about other parties' data [13]. The semi-honest party will attempt to learn or deduce data from other parties. A malicious (or active) party will arbitrarily deviate from the protocol and will make deliberate attacks to obtain access to data from other parties [60]. For example, possible malicious behavior might be not starting the execution of protocols at all or suspending (or aborting) the execution at any desired point in time. Papers that use ambiguous expressions such as 'untrusted' or 'non-trusting' or 'non-collaborative' are not classified into any category, because they did not clearly indicate the adversarial property of data parties, nor did they provide provide any privacy or security proof of their methods. In addition, we include the situation where a third party was involved. A third party, as another independent entity, can combine data from multiple parties, execute analysis on the joint datasets, or do the final computation based on information from data parties. A third party can be fully-honest, semi-honest, and malicious.

2) Data partitioning Figure 1 shows three scenarios of data partitioning which are considered in this review: 1) Horizontally partitioned data which contains the same attributes from different data instances (see Figure 1a). For example, different hospitals see different patients, though they collect the same patient attributes; 2) Vertically partitioned data which contains the same data instances but with different attributes (see Figure 1b). For example, a hospital has data on the same individuals as the tax office, while the attributes collected differs per data party; 3) Arbitrarily partitioned data, the hybrid situation of horizontally and vertically partitioned data. In this scenario, the data providing institutes hold different attributes for different data instances (see Figure 1c).

3) Dataset information factor indicates whether the study provides adequate information about the applied datasets in their experiments. Basic information of datasets including sources, names, numbers of features and instances, categorical or numeric type (if available) were recorded. Considering the readability, collected information is composed into five categories for this factor:

(1) Datasets that are publicly available (e.g., UCI repository) [61]
(2) Datasets from practical cases such as real patients data from a clinic
(3) Synthetic datasets and datasets which were generated by authors
(4) Experiments are presented in the paper but information about datasets is missing
(5) No experiments are presented in the paper

4) Privacy definition or measurement describes whether the study gave an explicit privacy definition, analyses, or measurements. Due to a lack of a universally accepted standard definition, there are many different definitions of privacy from various aspects such as law and philosophical point of view covering personal information, body, communications, and territory [62, 63]. This review only focuses on information privacy which concerns the control of collection, use, retention, and distribution of personal information. During reviewing, we do not assess if the privacy definitions are correct and the levels of privacy these studies can preserve though whether they gave a sufficient description, measurement, or analysis of privacy.

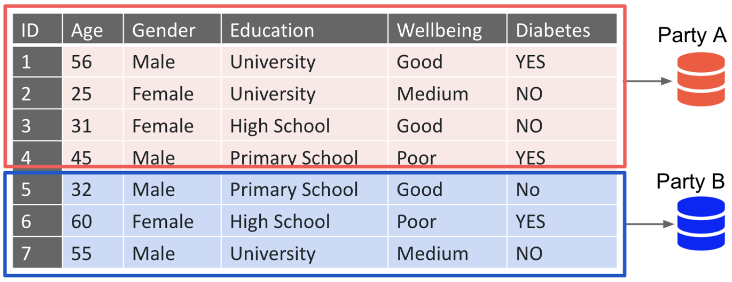
5) Privacy-preserving methods are classified into 5 categories: 1) secure multiparty computation - building blocks, 2) secure multiparty computation - generic and specialized construction protocols, 3) data modification, 4) local learning and global integration, and 5) others. First 4 categories have been explained in detail in the Privacy-Preserving Method Section. The papers which did not use any method from above are categorized to “others”.

6) Types of problems covers four main data mining areas: i.e., classification, regression, clustering, and association rule mining. Classification predicts a class with categorical labels. These categorical labels can be represented by discrete values, where the ordering among values has no meaning. In contrast, regression is to predict continuous-valued function or ordered value. Clustering is to group a set of data objects into multiple groups (clusters) so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Association rule mining is to discover interesting associations and correlations between itemsets in transactional and relational databases [64]. Additionally, we labeled the studies as “general” that solved some mathematical or statistical problems which are applied to classification, regression, and clustering. The studies which worked on outlier detection, record linkage, recommendation system, attribute/dimension reduction, feature selection, and probabilistic graph are categorized into “others”.

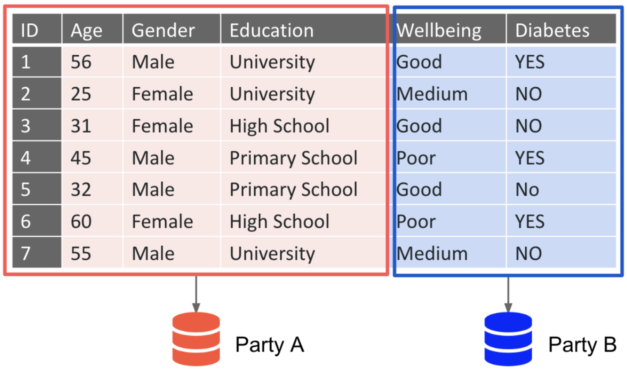
7) Data mining algorithms present the algorithms which have been developed in a privacy-preserving manner and which ones lack attention. There are plenty of algorithms across the data mining and statistics domain [19, 65]. In this review, the top eight algorithms are listed in the result table including decision tree, K-nearest neighbor, bayesian networks, support vector machine, neural networks, K-means, linear/logistic regressions, and A-priori algorithms.

8) Complexity and cost indicates whether the study explicitly measures computational complexity, time cost, and communication cost. The papers which did not present any experiments but only briefly discussed computation, time, and communication costs are counted as ‘No Measurement’.

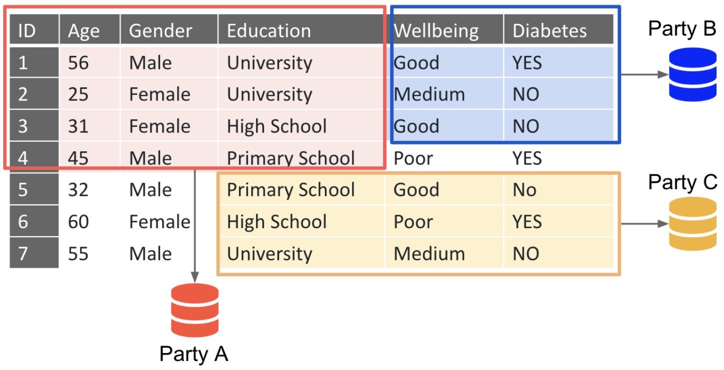
9) Performance measures covers whether the study compared the performance of their approaches with 1) other published PPDDM methods, 2) centralised data mining methods, and 3) distributed without preserving privacy methods. The performance measures include accuracy, precision, recall, F1 score,



(a) An example of horizontally partitioned data.



(b) An example of vertically partitioned data.



(c) An example of arbitrarily partitioned data.

Fig. 2. Examples of three different partitioned data. Fig.2a shows horizontally partitioned data which contains the same attributes/features from different data instances. Fig.2b shows vertically partitioned data which contains the same data instances but with different attributes/features. Fig.2c shows arbitrarily partitioned data which is a hybrid situation of horizontally and vertically partitioned data.

AUC (Area Under the Curve), mean squared error, mean absolute error, and other standard evaluation criteria in the data mining domain [64, 66–69]. Owing to the high degree of heterogeneity in the reporting of performance measures across the reviewed papers, we determine whether any performance measure was applied to evaluate the methods rather than comparing different performance measures. The papers which contained experiments but did not compare their results with other methods are categorized into "No comparison (with experiment)". The studies which did not provide any experiments are classified to

"No experiments".

10) Scalability covers whether the study presented a scalability analysis or the experiments prove the scalability of their approach. The scalability in this review means if the approach can tackle large-size datasets which contain a large number of either features or instances. It is noteworthy that only discussing scalability or mentioning their approaches are scalable were not included.

4. Results

In this section, we first describe the number and distribution of search results retrieved from the six search engines in the last 20 years. Detailed reviews of selected papers based on the evaluation criteria are elaborated in section 4.2. The analysis of the relations among selected papers is described in section 4.3.

4.1. Search Results

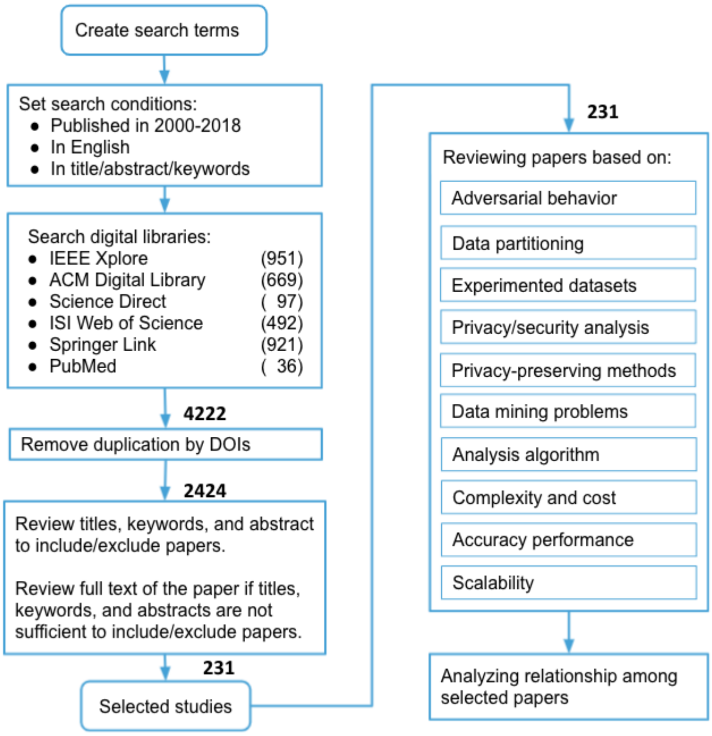


Fig. 3. Workflow of conducting this systematic review

In Figure 3, we present the workflow of this systematic review with the number of papers included in each step. Following the inclusion criteria, 4222 publications including duplicates were retrieved from six search engines. Most papers were from IEEE and Springer Link followed by ACM Digital Library. To remove the duplicates, we used Digital Object Identifiers (DOI) to keep the unique papers. The number of publications was reduced from 4222 to 2424. Furthermore, we filtered out irrelevant papers by screening the titles and abstracts of the retrieved papers. Papers that focused on parallel computing, cloud computing, edge computing, network security, intrusion detection, web attack detection, privacy in mobile data and geographic data, differential privacy, privacy in data collecting, data publishing, data storing, data querying were excluded. In the end, 231 papers were selected to be preliminarily reviewed.

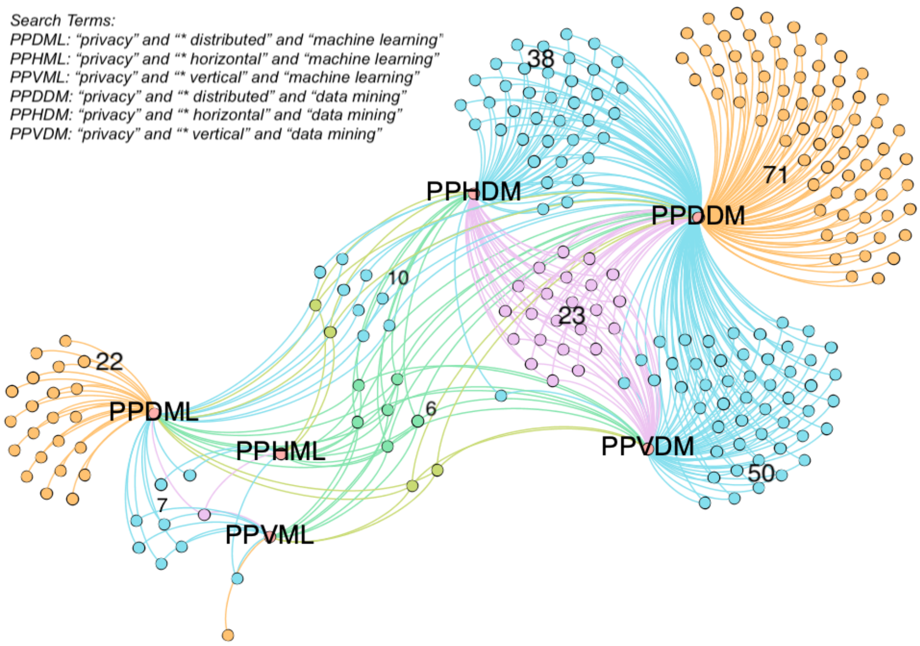
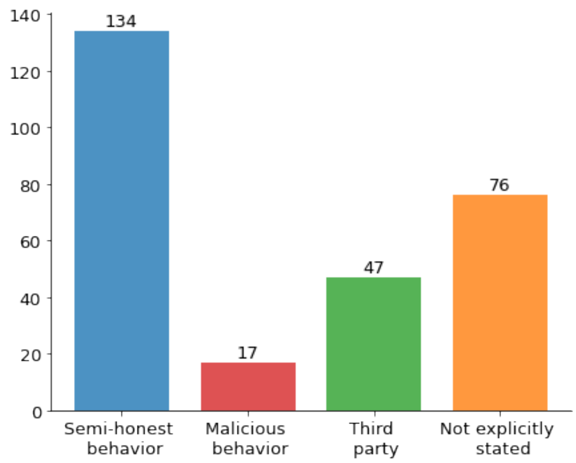


Fig. 4. Numbers and clusters of selected papers from different search terms. Papers are presented as nodes and clustered by the search terms. The number of papers in each cluster is labeled in the figure. The edges show which search terms were used to find the papers. For example, the 23 nodes in the purple cluster were found from using search terms PPDDM, PPHDM, and PPVDM.

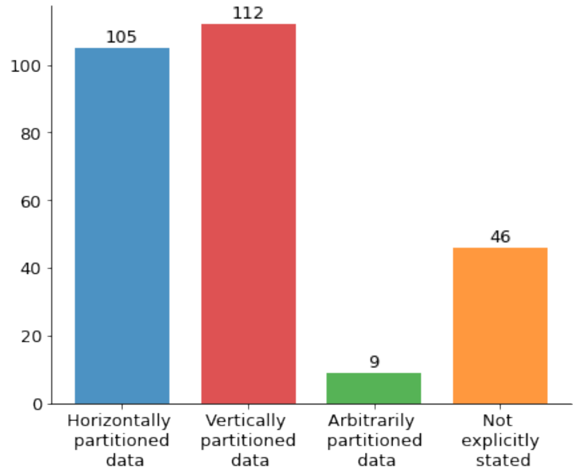
To improve the insight of the search result, we map the selected papers into graphs by using the Gephi visualization tool [70]. In Figure 4, the distribution of 231 selected papers using different search terms is presented. Papers are presented as nodes and clustered by the search terms. For instance, 182 selected papers were found by using the search term - PPDDM, while 38 of them were findable in PPHDM category and 50 of them were findable in PPVDM. It is obvious that data mining papers are the majority of the search outcomes. It is reasonable as data mining covers a larger scope than machine learning. Privacy issues should be considered in the entire data processing procedure instead of only the part of analysis and building machine learning models. Moreover, a large number of papers (71 papers from PPDDM, 22 papers from PPDML) did not indicated what exact data partitioning problems (vertical, horizontal, or arbitrary) their method can solve in their titles, abstracts, and keywords. This increases difficulties for other researchers and practitioners to find the correct papers based on their needs.

4.2. Review Results

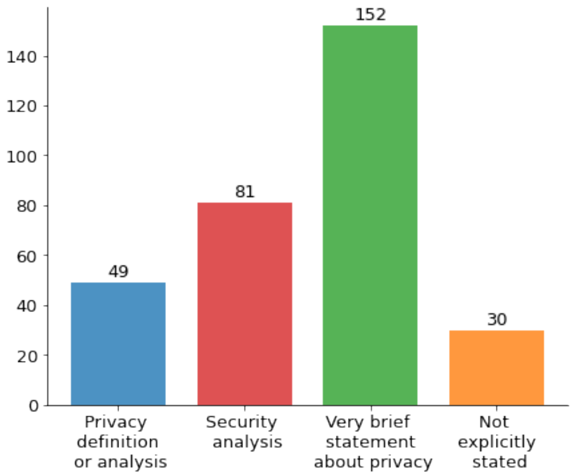
In Fig. 5, we summarize the review results of 231 papers using the 10 evaluation factors we discussed previously. The full review results of 231 papers are publicly available in the data repository: https://figshare.com/s/cbb2317239ecfa48339f. (DOI: 10.6084/m9.figshare.14239937). The following subsection elaborates on the review result of each factor.



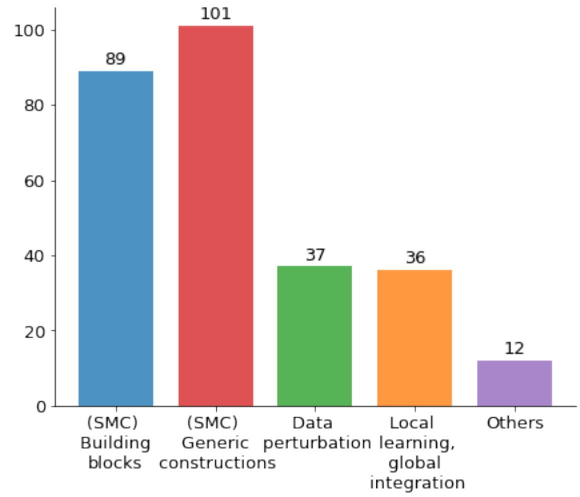
(a) Adversarial behavior of data parties



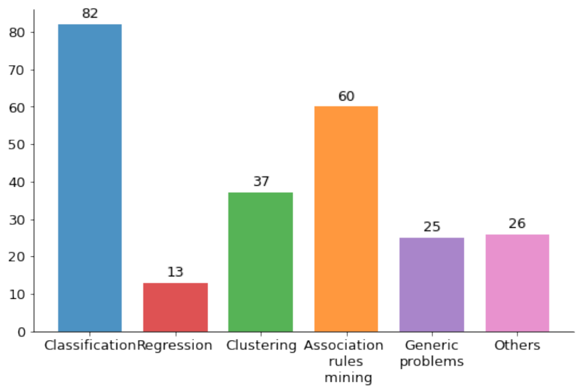
(b) Data partitioning



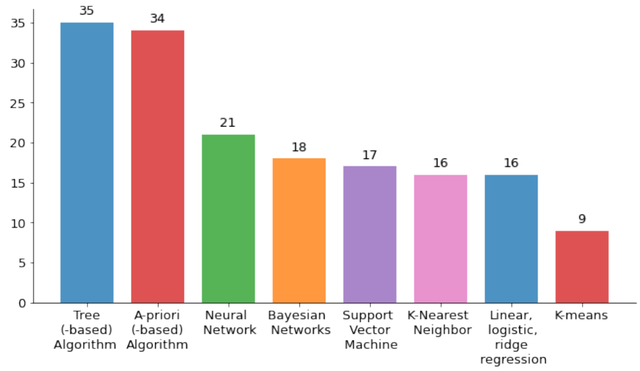
(c) Privacy definition or analysis



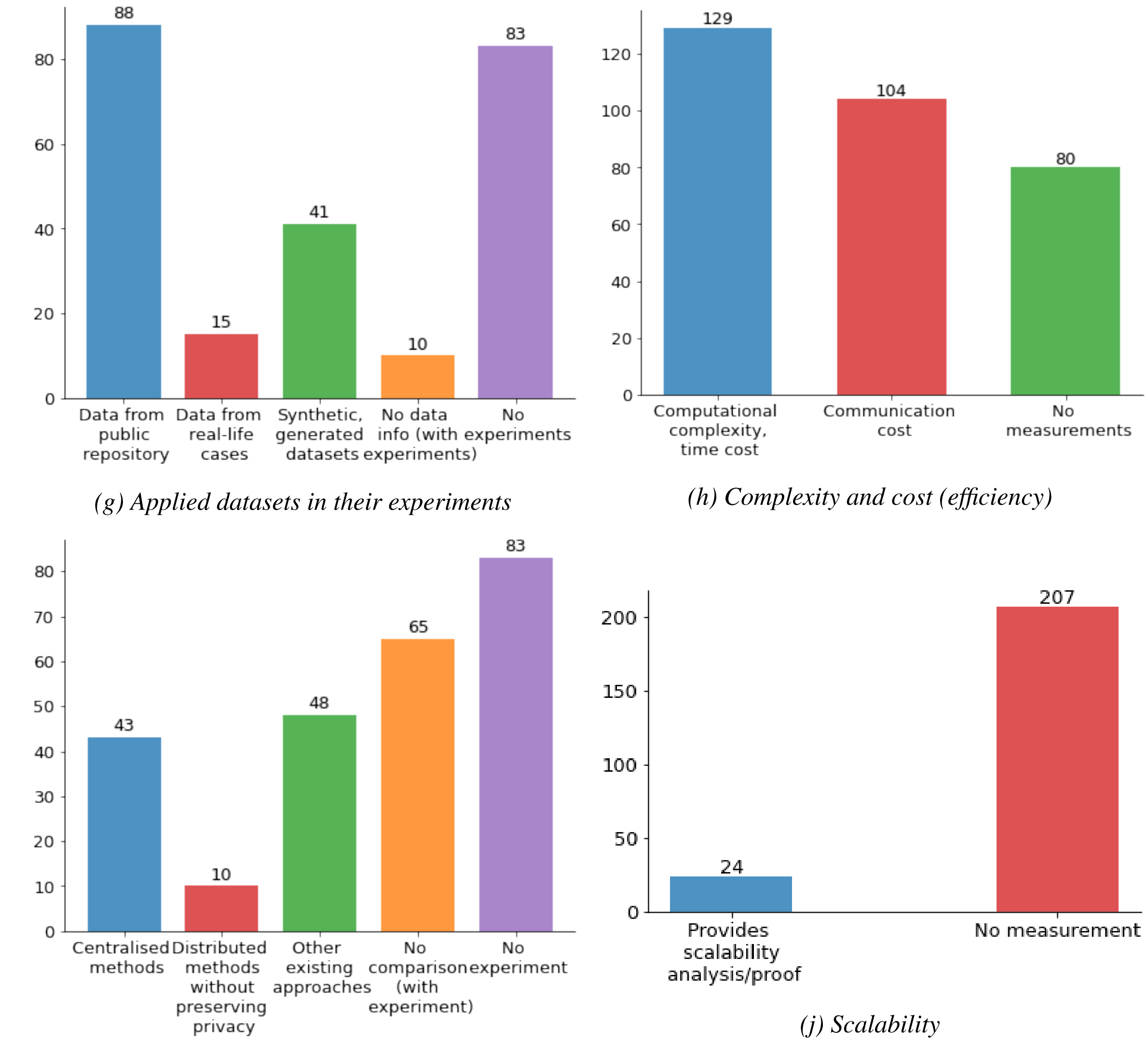
(d) Privacy-preserving methods



(e) Types of data problems



(f) Data mining algorithm



(g) Applied datasets in their experiments

(h) Complexity and cost (efficiency)

(i) Performance measures (compare with)

(j) Scalability

Fig. 5. Bar charts of presenting review results using 10-factor evaluation criteria. Papers can cover one or more items in the factors except Privacy Definition/Analysis and Scalability.

Adversarial behavior of data parties. About half of the reviewed studies assuming their approaches are applicable for the data parties with semi-honest adversary behavior. In contrast, only 17 reviewed studies developed their methods against malicious parties. Third party constructions were applied in the method of 47 studies. More than half of them handled semi-honest behavior data parties together with employing the third party. However, it is worth noting that over 30% of selected papers did not state a clear assumption that which adversarial behavior their approach can deal with.

Data partitioning. Horizontally partitioned data (105 reviewed papers) and vertically partitioned data (112 reviewed papers) seem to be represented equally in the selected literature. There are 35 papers handling both horizontally partitioned data and vertically partitioned data. However, only 9 reviewed studies developed PPDDM methods on arbitrarily partitioned data which can work with semi-honest data parties. Additionally, 20% of selected studies did not indicate in which data partitioning situation their methods can be applied.

Privacy is one of the most important evaluation parameters for PPDDM techniques. However, only one fifth of selected studies describe an explicit definition of privacy and mathematical analysis of how much information is leaked by the proposed method. There are 81 papers proving the security of their approaches rather than a privacy analysis. The difference between security and privacy will be discussed in the next section. The majority of studies describe “privacy preservation” very briefly in their own understanding. These descriptions are heterogeneous: e.g., “not revealing privacy of any database”, “not compromising the privacy of the data owners”, “preserving the confidentiality of datasets”, and “no important information leakage”. The remaining 30 papers proposed new PPDDM methods without indicating any definition or description about privacy.

Privacy-preserving methods. Secure multiparty computation techniques are the most encountered solutions in the PPDDM domain. The generic and specialized protocols were applied in 101 papers, while 89 studies employed homomorphic encryption or oblivious transfer protocols. A minority of reviewed studies used data modification, or methodologies to train local models and combine these local models into a global model. A combination of techniques such as combining data modification and homomorphic encryption protocols has been applied by 41 studies.

Types of data problems and data mining algorithms. Classification problems attracted the most attention from researchers in the PPDDM domain, followed by association rule mining and clustering. By contrast, a minority of studies deal with regression modeling. The most implemented data mining algorithms tackling these data problems are: Tree-based algorithms such as decision tree, random forest (35 papers), A-priori-based algorithms (34 papers), Neural Networks (21), Bayesian Networks (18), Support Vector Machine (17), K-Nearest Neighbor (16), Linear/Logistic/Ridge Regression (16), and K-means (9). There are over 10% of reviewed papers studied on generic algorithms that can be applied to multiple data mining techniques such as gradient descent. About 12% of reviewed papers worked on solving privacy problems in outlier detection, record linkage, recommendation system approaches, attribute/dimension reduction, feature selection, and probabilistic graphs.

Applied datasets in their experiments. From the selected studies, we identified the datasets that were applied in their experiments, measurement of complexity and cost, and performance on accuracy and scalability. We found 88 studies used datasets from public repositories, while 41 studies generated synthetic datasets to conduct their experiments. It is noteworthy that only 15 papers applied real-world datasets in practical use cases. Furthermore, it is remarkable to find that 83 papers proposed new methods by only presenting mathematical theories without any experiments, while 10 papers conducted experiments but did not provide any information about the datasets.

Complexity and cost. To prove the efficiency of proposed methods, 129 papers calculated computational complexity and/or time cost, while 104 papers reported communication cost of their approaches. Among them, 85 papers measured both computational complexity/time cost and communication cost. However, one third of (80) reviewed papers did not have any measurement of computation, running time, or communication cost.

Accuracy performance. We found 83 reviewed papers were lacking in evaluating accuracy performance of their methods because no experiments were conducted in these studies. In the rest of the papers, 43 papers proved their PPDDM methods can achieve comparable accuracy as the centralised data mining methods, while 48 studies proved their methods exceeded other existing PPDDM methods or achieved the same accuracy with higher efficiency. A small proportion of (10) studies proved their privacy-preserving models have comparable performance on learning partitioned data as the non-privacy-preserving models. Lastly, 65 papers conducted experiments but did not compare with any other methods or situations.

Scalability. The last factor - scalability - shows 10% papers proved or analyzed the scalability of their proposed methods. The majority of papers either only provided very brief statements in the discussion and future work section of the paper, or did not consider the scalability challenge.

4.3. Result of Referencing Relationship among Selected Papers

We investigated how selected papers influence each other based on their references and citations. We extracted text from reference sections of all selected studies and recognized titles and authors from the text. As DOIs are not available in the reference section of all papers, only titles and authors were used to recognize different studies. Figure 6 illustrates the citation network, where papers are represented as nodes, and citing relations are represented as edges. The size of nodes are proportional to the number of citations among the 231 papers. Papers [39, 71, 72] are most cited, with 1354, 1320, and 875 citations respectively (until 2021 Feb).

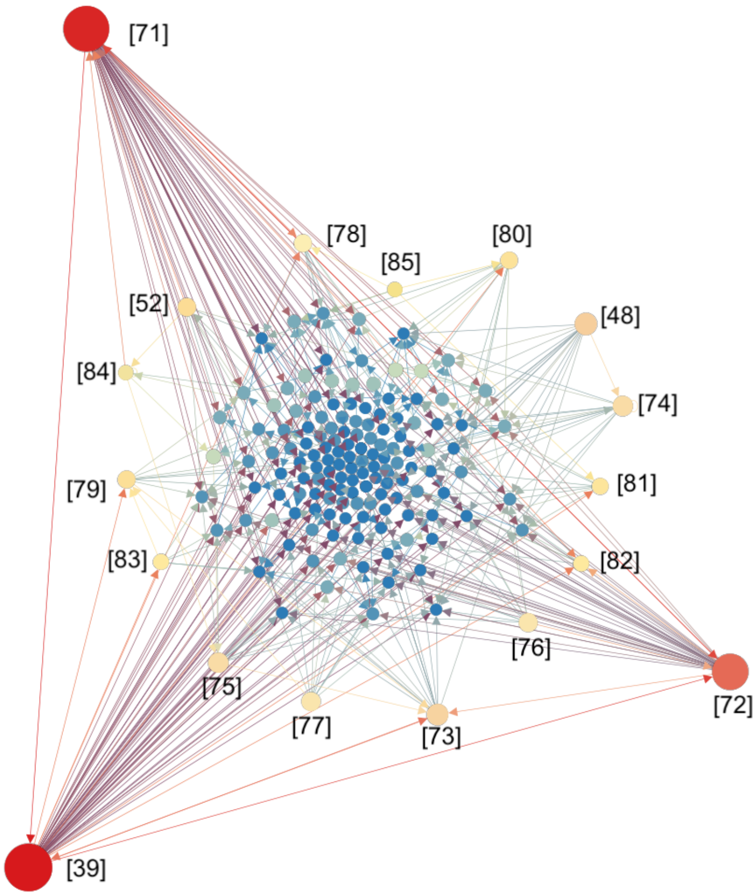


Fig. 6. Citation network among the selected papers. Papers are presented as nodes, while the citing relations are presented as edges. The size of nodes are proportional to the number of citations among the 231 papers.

Ref	User scenario		Data distribution			Privacy/ security analysis	PP method*		Type of problems			Experiment	Cost	
	Semi-honest	Third party	Horizontal	Vertical	Arbitrary		SMC	Local global*	Classification	Clustering	ARM*		computation	communication
[39]	✓			✓		✓	✓				✓			✓
[71]	✓		✓			✓	✓				✓		✓	✓
[72]	✓			✓		✓	✓			✓			✓	✓
[48]	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓
[73]	✓			✓			✓		✓				✓	✓
[74]				✓		✓	✓	✓	✓					
[75]	✓	✓	✓	✓	✓	✓	✓			✓			✓	✓
[76]	✓		✓			✓	✓			✓				
[77]			✓			✓		✓		✓		✓		✓
[52]	✓			✓		✓		✓	✓				✓	✓
[78]				✓		✓	✓				✓		✓	✓
[79]	✓		✓	✓		✓	data perturbation		✓	✓		✓	✓	✓
[80]	✓		✓				✓	✓	✓			✓	✓	✓
[81]	✓		✓			✓	✓			✓				
[82]	✓	✓	✓			✓	✓			✓		✓	✓	✓
[83]				✓		✓	✓		Probabilistic graph			✓		✓
[84]	✓		✓				✓		✓				✓	✓
[85]	✓			✓		✓	✓		✓		✓		✓	✓

Table 1: Review results for the 18 most cited papers in this review.
(PP method: Privacy-preserving methods; Local-global:
Local learning and global integration; ARM: Association Rule Mining.)

Table 1 lists the attributes of the most cited articles. Semi-honest behavior is the most common assumption, while none of these influential papers addressed malicious adversarial behavior. 3 out of 18 studies considered a third party. Two papers [48, 75] took all possible data distribution situations (horizontally, vertically, and arbitrarily partitioned data) into account. Horizontally and vertically partitioned data problems have been covered with a good balance. Although the vertically partitioned data problem is more complicated than the horizontally one [86, 87], our review indicates that they have been developed at the same pace.

A similar balance is apparent in the types of problems as well. Seven papers focused on solving a classification problem by using SVM, decision tree, bayesian networks, while 8 papers looked at clustering problems particularly at K-means, Expectation Maximization algorithms (EM), Local Outlier Factor (LOF) algorithm. Association rule mining problem has fewer influential papers, but the top 2 influential papers [39, 71] both focused on this problem. In contrast to the balance in the types of problems, privacy-preserving solutions from the influential papers are completely dominated by SMC. 16 out of 18 influential papers covered SMC [78, 84] combined SMC with homomorphic encryption, while [48, 74, 80] combined it with structuring local and global data miners. More than half of existing studies in our review applied SMC as the major privacy-preserving method.

It is notable that 12 out of 18 studies did not conduct experiments, but they provided explicit privacy/security analyses and costs measurements instead. These privacy/security analyses have been presented in different ways, but the main objectives were similar. All influential papers described what information their approaches can protect, what information have to be disclosed, and what potential risks, problems or troubles might exist. Moreover, their computational complexity and communication costs of their approaches were clearly presented as one of the evaluation parameters. Hence, the described performance evaluation on privacy and efficiency may be the reasons why these papers are often cited.

5. Discussion

PPDDM has been rapidly developing through active research programs across different scientific communities including data mining and machine learning, mathematics and statistics, cryptography, and data management. The total number of publications in this domain has dramatically increased in the last 20 years. Many of the studies included promising results in the efficiency and accuracy of their models in an experimental environment. These promising experimental results helped move the field forward towards practical applications. In the past five years, use cases have been developed in healthcare [7, 88–90], finance [91], and technology companies [14, 92, 93] to examine different PPDDM methods. Participation of industry partners accelerates the transformation of PPDDM theoretical methods to practical applications. The existing PPDDM methods have been well-developed to solve a wide range of data problems (e.g., classification, clustering, association rule mining) using various data mining algorithms. To achieve the goal of PPDDM methods in practical studies, methods that will preserve privacy require legal, ethical, and social scholars in addition to scientific and technical experts. Successful implementation of PPDDM needs a joint effort from researchers with diverse backgrounds.

5.1. Inadequate definition and measurement of privacy

There are some challenges hindering PPDDM methods to be further developed and widely applied in practice. One of the key issues is the lack of the definition and measurement of (information) privacy. The meaning and operational definition of privacy is commonly ambiguous and subjective in the selected papers. It is not sufficiently expressed by the papers what privacy means to them, and what their proposed approaches can preserve. The three most common definitions of privacy preservation in the selected papers are 1) not revealing sensitive information; 2) not revealing private information; 3) not revealing raw data. However, it is unclear if "sensitive information" or "private information" or "raw data" is equal to personal information privacy. To understand personal information privacy from a legal and ethical perspective, it is the right of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively [94–96]. Similarly, privacy is seen as the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others [97]. In relation to controlling and protecting privacy, two definitions from legal literature state "Privacy, as a whole or in part, represents the control of transactions between person(s) and other(s), the ultimate aim of which is to enhance autonomy and/or to minimize vulnerability" [98] and "Privacy is to protect personal data and information related to a communication entity to be collected from other entities that are not authorized" [99].

According to privacy definitions above, any information about a person can be considered as privacy regardless of its sensitivity, originality, and transformation. It is the data subject that determines what data is private. For instance, a data subject might consider their state of mental health more private than their date of birth. However, existing PPDDM methods have not yet addressed different privacy requirements from each data subject. All data elements have equal treatment for all data subjects. This might cause insufficient privacy preservation for some data elements and data subjects, while over-protection for the others. To personalize the privacy preservation, Xiao and Tao [100] proposed a new generalization framework using personalized anonymity that data subjects can specify the degree of privacy protection for her/his data elements. In the study, Xiao and Tao [100] assume: 1) data subjects can easily set/change their privacy requirements with data parties, 2) data subjects are knowledgeable about the benefits and consequences of setting different degrees of privacy. This method is only applicable when the data is centralized. In the partitioned data scenario, there is no platform yet facilitating data subjects to customize privacy requirements for each data element across multiple parties. Second, privacy requirements can be satisfied when using one single data source. However, analyzing an amount of partitioned data from multiple sources increases risk of privacy violation. As indicated by the 2020 European Commission White Paper on Artificial Intelligence [101], data about persons can be re-identified through the analysis of large amounts of other non-private data.

5.2. Ambiguity between privacy and security

Another ambiguity lies in the difference between (information) privacy and (information) security. Different from privacy, security has an explicit definition and measurement from the cryptography domain, separating the problem into semantic security and technical security [60]. Semantic security is a computational-complexity analogue of Shannon's definition of perfect privacy (which requires that the ciphertext yield no information regarding the plaintext). Technical security is the infeasibility of distinguishing between encryptions of a given pair of messages. Generally speaking, security focuses on maximally protecting information/data from malicious attacks and stealing data. Satisfying security requirements is not always sufficient for addressing privacy issues [102]. However, in the majority of the reviewed papers, the difference between security and privacy is not clearly stated. For example, some studies defined the data privacy but evaluated the methods by conducting security analysis [103–105]. Certain approaches guarantee that the data used for the analyses remain unknown to other parties through secure computation. However, this does not mean that the resulting output from the analyses is equally privacy-preserving [13, 102, 106]. The output can reveal information about the person so that the privacy is still not preserved according to the privacy definition we discussed above. For instance, the outcome of the analysis might portray a harmful profile for individuals sharing certain characteristics. Some essential problems are not taken into consideration, such as how much data or information will be revealed by the output although the output is computed securely [90], whether the models and algorithms are harmless to the data party or individuals, does the purpose of formula or function satisfy the legal and ethical concerns [107, 108]. A typical example is building a decision tree on vertically partitioned data in a privacy-preserving way. The decision tree model can be securely and correctly built up. However, to some extent, the decision tree, as an output, leaks information about the input data [109]. Decision tree algorithm splits nodes based on attributes or features, while the splitting decision is dictated by the data. When the final decision tree is completed, the leaf nodes in the tree might reveal some information about the input data such as class counts. Therefore, releasing the final decision tree to all participating parties could potentially breach privacy.

Providing an applicable privacy description is significant to any PPDDM studies. What data or information should be preserved from mining can be influenced by different legal restrictions, ethical concerns, organizational regulations, personal preference, and application domains. Instead of generalizing the solution of a specific scheme to all situations, it is more reasonable to make a precise statement on the specific scenario to address. Therefore, the authors could provide a clear description to readers about what privacy means to them, and in which situation the proposed approach is privacy preserving by answering the following questions:

(1) *What is the operational definition of privacy-preservation for the work?*
(2) *Which data are deemed sensitive or require protection, and why?*
(3) *What computational operation is intended to preserve privacy, and where does it fail?*
(4) *What is the role or responsibility of each actor (e.g., data collector, data holder, data publisher, data analyst) in the scenario?*

5.3. Inadequate experiments and practical use cases

Half of the reviewed papers did not provide any experiments to evaluate their methods, and as such there were no reports of accuracy, efficiency, and scalability in these papers. This is one of the gaps between the theoretical research and practical use cases in this domain. Solutions based on theory might not solve real world problems. In our review, only a few papers applied real-world use cases to evaluate their methods. It reflects a fact in this domain that many solutions have been proposed by researchers, but only a few of them were implemented in practice. Without experimenting on real data, the proposed approaches might neglect essential problems such as sparse or biased datasets [55, 110], or record linkage problems in vertically partitioned data [111–113]. Future research in PPDDM should consider conducting experiments using real-world datasets and provide adequate information about the experiments. Meanwhile, we observed most real-life use cases to examine existing PPDDM approaches from the healthcare domain [88, 90, 107]. We suggest researchers apply the PPDDM methods to practical cases also in other research domains such as social science and finance. In addition to developing new theories, implementing and improving existing approaches in practice can also make a meaningful contribution to the PPDDM domain.

5.4. Challenge of linking data in vertically partitioned data scenario

The accurate linking of entities across distributed datasets is of crucial importance in vertically partitioned data mining. Data parties must link their data and/or order them in an identical manner prior to data analysis. However, most papers assume this correspondence between data entities (records) exist by default. Matching data entities from multiple datasets can be error-prone particularly where the use of direct identifiers - even encrypted - are prevented by law, as is the case in the use of the national Citizen Service Number (‘Burgerservicenummer’) in the Netherlands [114]. Sharing such identifiers compromises privacy as the sole information that a data subject is known to another data entity might be sensitive. Furthermore, one often assumes that records can be linked by doing exact matching on this unique identifier. However, exact matching can be very difficult due to the unstable and incorrect identifiers. Winkler and Schnell showed that 25% true matches would have been missed by exact matching in a census operation [115, 116]. In another case, two data parties do not share the unique identifiers but have some features in common. As an alternative solution, two parties can match the data entities based on their common features. The matching accuracy will be affected by the correctness, completeness, and updating promptness of these common features from both data parties. In addition, privacy needs to be preserved in the matching procedure. Some efficient and privacy-compliant algorithms for the field of privacy-preserving entity matching have been developed [117–120] in the past 10 years.

5.5. A recommendation list of key parameters for PPDDM studies

It is challenging to compare similar PPDDM methods where there is a lack of key parameters presented. For instance, approaches which are designed for semi-honest parties might not be comparable with the approaches aiming to handle malicious behavior. The privacy-preserving methods for semi-honest parties will fail if involved parties show malicious behavior such as manipulating the input or output or completely aborting the protocol. Thus, the allowed adversarial behavior of participating parties is essential to be explicitly stated in the PPDDM papers. To consider all key parameters in PPDDM techniques, we provide a list of recommendations for the reporting of studies proposing new PPDDM methods or improving existing PPDDM methods as Table 4 shows. The recommendations detail the key parameters that should be described in each section of the paper of PPDDM. The factors in Table 2 refer to the 10 factors in the evaluation criteria which were discussed in the Methodology Section.

Section	Factor	Recommendations
Title and abstract		
Title and keywords	2,7	Identify the study as developing new or improving existing PPDDM algorithms to solve which data problem by using which type of partitioned data in a privacy-preserving manner
Abstract	1,2,4,6,7	Summarize the problems, objectives covering assumed adversarial behavior of data parties, data partitioning, brief description about privacy-preserving method, data mining algorithms, and applied dataset in the experiments.
Introduction		

Table 2 continued from previous page

Section	Factor	Recommendations
Problem statement and background	2,3,5,6	Describe how data partitioned in which domain are considered by this study, what privacy issues are involved in that domain, which data mining algorithm is studied to solve what problems. Additionally, the number of participating parties and if all parties or only some parties have the target class should be also covered by this section.
Objectives and study design	1,3,4,7	Specify the objectives and study design include what level of privacy (or information leakage) is preserved against what adversarial behavior, applied privacy-preserving methods, evaluation criteria (for accuracy, efficiency, and privacy level), applied datasets in the experiments.
Methods		
Method design	4,5,6	Clearly explain which privacy-preserving methods are applied including the specific protocols/structures, proofs of preserving information leakage. Then, describe how certain data mining algorithms are adapted to combine with privacy-preserving methods, what information is communicated among parties, and complexity in different scenarios such as using categorical or numerical data, or involving different numbers of data parties. Lastly, make the code publicly available so that other researchers can reproduce the work.
Data	7	Describe data sources (and where and how other researchers can request the same dataset), the type and size of the datasets, basic description about data, what the target features/attributes are, missing values, and other basic information about the datasets.
Data analysis design	5,6	If real-life datasets are applied in the study, this subsection should describe the pre-processing of features/attributes (such as normalization, re-sampling), data analysis algorithms, parameter setting, and so on with reference to other comparable studies.
Experiment design	7	Describe how the datasets are partitioned (both feature-wise and instances-wise), how data parties communicate/transfer files, what validation is used, and what machine(such as CPU, memory) and software(versions) are used to do the experiments. In addition, experiments should be set up to compare with other existing PPDDM methods, or compare with privacy-preserving centralised data mining methods, or compare with distributed data mining methods without preserving privacy.
Evaluation design	8,9,10	Describe the evaluations of accuracy, efficiency (computational complexity, time cost on computation and communication among parties), privacy/security (such as information disclosure measurement)

Table 2 continued from previous page

Section	Factor	Recommendations
Result		
Discovery from datasets	7	If real-life datasets are applied in the study, this subsection should describe what new knowledge was obtained from their analysis
Model performance	8	Present the performance measures such as accuracy scores of the proposed models in comparison with other existing PPDDM methods, or privacy-preserving centralised data mining methods, or distributed data mining methods without preserving privacy. Performance will be presented based on the evaluation criteria which was described in the methods section.
Privacy and/or security analysis	9	Provide sufficient privacy/security analysis based on the assumed adversarial behavior (semi-honest or malicious). Describe what information is exchanged among parties, what can be learnt from the exchanged information, if the models as a final outcome can cause information leakage, what the potential risks exist during the training process or in the final model.
Scalability analysis	10	Present the computation complexity and time consumption of the methods and describe what the volume (number of instances) and variety (number of features/attributes) of data can be handled by the proposed methods
Discussion		
Limitations	/	Discuss any limitations of proposed methods such as special cases where the methods are not applicable or certain assumptions which are not common in practice.
Interpretation	/	If real-life datasets are applied in the study, this subsection should discuss the findings with reference to any other validation data from other studies. Then, interpret the model performance on accuracy, efficiency, feasibility in practice, strengths and weaknesses with reference to other existing PPDDM methods.
Implementation	/	Discuss what other resources, paperwork, or supports are needed to implement the proposed methods, what potential challenges or risks will appear if apply the methods on real-life data.

Table 2: A list of recommendations for reporting PPDDM studies

5.6. Potential limitations

The findings of this review have to be seen in light of some potential limitations. First, the 231 reviewed studies were searched from only 6 digital bibliographic databases (IEEE Xplore Digital Library, ACM Digital Library, Science Direct, ISI Web of Science, SpringerLink, and PubMed) and must be peer-reviewed publications. Some relevant studies may be missed in this review because they were not findable in these 6 bibliographic databases during searching. Studies that have not been peer-reviewed such as relevant articles published on arXiv.org[7] were excluded. Second, we did not apply an iterative 'snowballing' approach to further identify more relevant studies [121]. 'Snowballing' searching includes 1) reference tracking which identifies relevant studies from the reference lists of the primarily selected papers, 2) citation tracking which identifies relevant articles that cite primarily selected papers. We decided not to apply 'snowballing' approach is because it may introduce a bias in favour of what authors think is relevant to their narrative [122]. The possible effect of not performing the 'snowballing' searching would be some follow-up studies of the reviewed papers were not included in this review if they did not meet our search strategy. Moreover, due to the scope of this review (providing a general overview of existing PPDDM methods and identifying outstanding challenges), more details of some privacy-preserving methods were not extensively discussed. For instance, in the category of 'local learning and global integration', multiple different methods can be applied to integrate the local miner (model) into a global miner (model) such as stacked generalization [123] and meta-learning[124]. In our belief this field warrants a separate in-depth review.

6. Conclusion

Privacy-preserving distributed data mining (PPDDM) techniques consider the issue of executing data mining algorithms on private, sensitive, and/or confidential data from multiple data parties while maintaining privacy. This review presented a comprehensive overview of current PPDDM methods to help researchers better understand the development of this domain and assist practitioners to select the suitable solutions for their practical cases. We discovered there is a lack of standard criteria for evaluating new PPDDM techniques. The previous studies applied a variety of different evaluation methods, which brings challenges to objectively comparing existing PPDDM techniques. Therefore, an comprehensive evaluation criteria was proposed in this review including 10 key factors - adversarial behavior of data parties, data partitioning, experiment datasets, privacy/security analysis, privacy-preserving methods, data mining problems, analysis algorithms, complexity and cost, performance measures, and scalability to assess 231 recent studies published between 2000 to 2020 (August). We highlighted the characteristics of the 18 most cited studies and analyzed their influence on other studies in the field. Furthermore, a variety of definitions of privacy and distinguishment between information privacy and information security in the PPDDM field were discussed in this review, followed by some suggestions of making applicable privacy descriptions for new PPDDM methods. We also provided a list of recommendations for future research such as explicitly describing the privacy aspect under consideration, and evaluating new approaches using real-life data to narrow the gap between theoretical solutions and practical applications.

In PPDDM, there is an important tradeoff between leakage of information and effectiveness or efficiency of learning. Addressing both is crucial in practice. Future research will preferably balance this trade-off depending on their specific use cases and the purpose of the data analysis. For example, people may weigh the different trade-offs when the purpose of data analysis is for commercial use or helping solve an urgent public health challenge. In the outbreak of the COVID-19 pandemic, the processing of sensitive data including personal health data is allowed by the General Data Protection Regulation (GDPR) and other European data protection laws in order to protect against serious cross-border threats to health [125].

[7] arXiv: https://arxiv.org/

Therefore, addressing this trade-off needs collaborations between PPDDM researchers and legal-ethical and social experts to investigate which and how much information revealing is acceptable to achieve the effectiveness and efficiency we require in certain situations.

Acknowledgement

Financial support for this study was provided by a grant from the Dutch National Research Agenda (NWA; project number: 400.17.605). We gratefully acknowledge the time and effort devoted by two reviewers Dr. Abdur Rahim and Dr. Dayana Spagnuelo for their valuable comments. We would like to thank Dr. Leto Peel for his generous feedback and suggestions to help us improve the quality of the manuscript. Special thanks are given to Dr. Amrapali Zaveri for her constructive suggestions on preparing the manuscript.

References

[1] Geoff Dougherty. *Digital image processing for medical applications*. Cambridge University Press, 2009.

[2] Hanaa Elshazly, Ahmed Taher Azar, Abeer El-Korany, and Aboul Ella Hassanien. Hybrid system for lymphatic diseases diagnosis. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 343–347. IEEE, 2013.

[3] EA Clarke. What is preventive medicine? *Canadian Family Physician*, 20(11):65, 1974.

[4] Jacques S Beckmann and Daniel Lew. Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities. *Genome medicine*, 8(1):1–11, 2016.

[5] Shawn Dolley. Big data's role in precision public health. *Frontiers in public health*, 6:68, 2018.

[6] Raphael B Stricker and Lorraine Johnson. Lyme disease: the promise of big data, companion diagnostics and precision medicine. *Infection and drug resistance*, 9:215, 2016.

[7] Arthur Jochems, Timo M Deist, Johan Van Soest, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, Philippe Lambin, and Andre Dekker. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology*, 121(3):459–467, 2016.

[8] Commission on Social Determinants of Health et al. *Closing the gap in a generation: health equity through action on the social determinants of health: final report of the commission on social determinants of health*. World Health Organization, 2008.

[9] Jessica S Ancker, Min-Hyung Kim, Yiye Zhang, Yongkang Zhang, and Jyotishman Pathak. The potential value of social determinants of health in predicting health outcomes. *Journal of the American Medical Informatics Association*, 25(8):1109–1110, 2018.

[10] Suranga N Kasthurirathne, Joshua R Vest, Nir Menachemi, Paul K Halverson, and Shaun J Grannis. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. *Journal of the American Medical Informatics Association*, 25(1):47–53, 2018.

[11] Olga Ohrimenko, Felix Schuster, Cédric Fournet, Aastha Mehta, Sebastian Nowozin, Kapil Vaswani, and Manuel Costa. Oblivious multi-party machine learning on trusted processors. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 619–636. USENIX Association, 2016.

[12] Ruili Wang, Wanting Ji, Mingzhe Liu, Xun Wang, Jian Weng, Song Deng, Suying Gao, and Chang-an Yuan. Review on mining data from multiple data sources. *Pattern Recognition Letters*, 109:120–128, 2018.

[13] Yehida Lindell. Secure multiparty computation for privacy preserving data mining. In *Encyclopedia of Data Warehousing and Mining*, pages 1005–1009. IGI global, 2005.

[14] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[15] Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[17] Sheng Shen, Tianqing Zhu, Di Wu, Wei Wang, and Wanlei Zhou. From distributed machine learning to federated learning: In the view of data privacy and security. *Concurrency and Computation: Practice and Experience*, 2020.

[18] Yousra Abdul Alsahib S Aldeen, Mazleena Salleh, and Mohammad Abdur Razzaque. A comprehensive review on privacy preserving data mining. *SpringerPlus*, 4(1):1–36, 2015.

[19] Elisa Bertino, Dan Lin, and Wei Jiang. A survey of quantification of privacy preserving data mining algorithms. In *Privacy-preserving data mining*, pages 183–205. Springer, 2008.

[20] Alpa Shah and Ravi Gulati. Privacy preserving data mining: techniques, classification and implications-a survey. *Int. J. Comput. Appl*, 137(12):40–46, 2016.

[21] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza, Yucel Saygin, and Yannis Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[22] Hina Vaghashia and Amit Ganatra. A survey: privacy preservation techniques in data mining. *International Journal of Computer Applications*, 119(4), 2015.

[23] Suchitra Shelke and Babita Bhagat. Techniques for privacy preservation in data mining. *International Journal of Engineering Research*, 4(10), 2015.

[24] Elisa Bertino and Igor Nai Fovino. Information driven evaluation of data hiding algorithms. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 418–427. Springer, 2005.

[25] Sam Fletcher and Md Zahidul Islam. Measuring information quality for privacy preserving data mining. *International Journal of Computer Theory and Engineering*, 7(1):21, 2015.

[26] Chris Clifton, Murat Kantarcioglu, Jaideep Vaidya, Xiaodong Lin, and Michael Y Zhu. Tools for privacy preserving distributed data mining. *ACM Sigkdd Explorations Newsletter*, 4(2):28–34, 2002.

[27] Jaideep Vaidya. A survey of privacy-preserving methods across vertically partitioned data. In *Privacy-preserving data mining*, pages 337–358. Springer, 2008.

[28] Ricardo Mendes and João P Vilela. Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access*, 5:10562–10582, 2017.

[29] Shimon Even, Oded Goldreich, and Abraham Lempel. A randomized protocol for signing contracts. *Communications of the ACM*, 28(6):637–647, 1985.

[30] Ronald L Rivest, Len Adleman, Michael L Dertouzos, et al. On data banks and privacy homomorphisms. *Foundations of secure computation*, 4(11):169–180, 1978.

[31] Monique Ogburn, Claude Turner, and Pushkar Dahal. Homomorphic encryption. *Procedia Computer Science*, 20:502–509, 2013.

[32] Craig Gentry et al. *A fully homomorphic encryption scheme*, volume 20. Stanford university Stanford, 2009.

[33] Andrew Chi-Chih Yao. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 162–167. IEEE, 1986.

[34] Silvio Micali, Oded Goldreich, and Avi Wigderson. How to play any mental game. In *Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC*, pages 218–229. Association for Computing Machinery, 1987.

[35] Donald Beaver, Silvio Micali, and Phillip Rogaway. The round complexity of secure protocols. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 503–513. Association for Computing Machinery, 1990.

[36] Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*, pages 351–371. Association for Computing Machinery, 2019.

[37] Stephen Pohlig and Martin Hellman. An improved algorithm for computing logarithms over gf (p) and its cryptographic significance (corresp.). *IEEE Transactions on information Theory*, 24(1):106–110, 1978.

[38] Kaibin Huang and Raylin Tso. A commutative encryption scheme based on elgamal encryption. In *2012 International Conference on Information Security and Intelligent Control*, pages 156–159. IEEE, 2012.

[39] Jaideep Vaidya and Chris Clifton. Privacy preserving association rule mining in vertically partitioned data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 639–644. Association for Computing Machinery, 2002.

[40] Mikhail J Atallah and Wenliang Du. Secure multi-party computational geometry. In *Workshop on Algorithms and Data Structures*, pages 165–179. Springer, 2001.

[41] Ioannis Ioannidis, Ananth Grama, and Mikhail Atallah. A secure protocol for computing dot-products in clustered and distributed environments. In *Proceedings International Conference on Parallel Processing*, pages 379–384. IEEE, 2002.

[42] Rick L Wilson and Peter A Rosen. Protecting data through perturbation techniques: The impact on knowledge discovery in databases. *Journal of Database Management (JDM)*, 14(2):14–26, 2003.

[43] Nabil R Adam and John C Worthmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

[44] Tore Dalenius and Steven P Reiss. Data-swapping: A technique for disclosure control. *Journal of statistical planning and inference*, 6(1):73–85, 1982.

[45] Stephen E Fienberg and Julie McIntyre. Data swapping: Variations on a theme by dalenius and reiss. In *International Workshop on Privacy in Statistical Databases*, pages 14–29. Springer, 2004.

[46] Diego Peteiro-Barral and Bertha Guijarro-Berdiñas. A survey of methods for distributed machine learning. *Progress in Artificial Intelligence*, 2(1):1–11, 2013.

[47] Joost Verbraeken, Matthijs Wolting, Jonathan Katzy, Jeroen Kloppenburg, Tim Verbelen, and Jan S Rellermeyer. A survey on distributed machine learning. *ACM Computing Surveys (CSUR)*, 53(2):1–33, 2020.

[48] Jaideep Vaidya, Hwanjo Yu, and Xiaoqian Jiang. Privacy-preserving svm classification. *Knowledge and Information Systems*, 14(2):161–178, 2008.

[49] Yunmei Lu, Piyaphol Phoungphol, and Yanqing Zhang. Privacy aware non-linear support vector machine for multi-source big data. In *2014 IEEE 13th international conference on trust, security and privacy in computing and communications*, pages 783–789. IEEE, 2014.

[50] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. Privacy-preserving heterogeneous federated transfer learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2552–2559. IEEE, 2019.

[51] Anup Tuladhar, Sascha Gill, Zahinoor Ismail, Nils D Forkert, Alzheimer's Disease Neuroimaging Initiative, et al. Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling. *Journal of biomedical informatics*, 106:103424, 2020.

[52] Jaideep Vaidya, Chris Clifton, Murat Kantarcioglu, and A Scott Patterson. Privacy-preserving decision trees over vertically partitioned data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(3):1–27, 2008.

[53] Eakalak Suthampan and Songrit Maneewongvatana. Privacy preserving decision tree in multi party environment. In *Asia Information Retrieval Symposium*, pages 727–732. Springer, 2005.

[54] Weiwei Fang and Bingru Yang. Privacy preserving decision tree learning over vertically partitioned data. In *2008 International Conference on Computer Science and Software Engineering*, volume 3, pages 1049–1052. IEEE, 2008.

[55] Elena Czeizler, Wolfgang Wiessler, Thorben Koester, Mikko Hakala, Shahab Basiri, Petr Jordan, and Esa Kuusela. Using federated data sources and varian learning portal framework to train a neural network model for automatic organ segmentation. *Physica Medica*, 72:39–45, 2020.

[56] Ye Dong, Xiaojun Chen, Liyan Shen, and Dakui Wang. Eastfly: Efficient and secure ternary federated learning. *Computers & Security*, 94:101824, 2020.

[57] Qi Zhao, Chuan Zhao, Shujie Cui, Shan Jing, and Zhenxiang Chen. Privatedl: Privacy-preserving collaborative deep learning against leakage from gradient sharing. *International Journal of Intelligent Systems*, 35(8):1262–1279, 2020.

[58] Michael Wolfson, Susan E Wallace, Nicholas Masca, Geoff Rowe, Nuala A Sheehan, Vincent Ferretti, Philippe LaFlamme, Martin D Tobin, John Macleod, Julian Little, et al. Datashield: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. *International journal of epidemiology*, 39(5):1372–1382, 2010.

[59] Barbara Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26, 2004.

[60] Oded Goldreich. *Foundations of cryptography: volume 2, basic applications*. Cambridge university press, 2009.

[61] UCI Machine Learning Repository, url = https://archive.ics.uci.edu/ml/index.php.

[62] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press, 1997.

[63] H Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research: an interdisciplinary review. *MIS quarterly*, pages 989–1015, 2011.

[64] Jianwei Han, Micheline Kamber, and Jian Pei. *Data Mining Concepts and Techniques*. Elsevier, third edition edition, 2011.

[65] Alex A Freitas. A survey of evolutionary algorithms for data mining and knowledge discovery. In *Advances in evolutionary computing*, pages 819–845. Springer, 2003.

[66] Johannes Fürnkranz and Peter A Flach. An analysis of rule evaluation metrics. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 202–209, 2003.

[67] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[68] Julio-Omar Palacio-Niño and Fernando Berzal. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*, 2019.

[69] Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018.

[70] Gephi - The Open Graph Viz Platform, url = https://gephi.org/.

[71] Murat Kantarcioglu and Chris Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE transactions on knowledge and data engineering*, 16(9):1026–1037, 2004.

[72] Jaideep Vaidya and Chris Clifton. Privacy-preserving k-means clustering over vertically partitioned data. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. Association for Computing Machinery, 2003.

[73] Rebecca Wright and Zhiqiang Yang. Privacy-preserving bayesian network structure computation on distributed heterogeneous data. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–718. Association for Computing Machinery, 2004.

[74] Hwanjo Yu, Jaideep Vaidya, and Xiaoqian Jiang. Privacy-preserving svm classification on vertically partitioned data. In *Pacific-asia conference on knowledge discovery and data mining*, pages 647–656. Springer, 2006.

[75] Geetha Jagannathan and Rebecca N Wright. Privacy-preserving distributed k-means clustering over arbitrarily partitioned data. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 593–599. Association for Computing Machinery, 2005.

[76] Xiaodong Lin, Chris Clifton, and Michael Zhu. Privacy-preserving clustering with distributed em mixture modeling. *Knowledge and information systems*, 8(1):68–81, 2005.

[77] Srujana Merugu and Joydeep Ghosh. Privacy-preserving distributed clustering using generative models. In *Third IEEE International Conference on Data Mining*, pages 211–218. IEEE, 2003.

[78] Justin Zhan, Stan Matwin, and LiWu Chang. Privacy-preserving collaborative association rule mining. *Journal of Network and Computer Applications*, 30(3):1216–1227, 2007.

[79] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1):92–106, 2005.

[80] Hwanjo Yu, Xiaoqian Jiang, and Jaideep Vaidya. Privacy-preserving svm using nonlinear kernels on horizontally partitioned data. In *Proceedings of the 2006 ACM symposium on Applied computing*, pages 603–610. Association for Computing Machinery, 2006.

[81] Mark Shaneck, Yongdae Kim, and Vipin Kumar. Privacy preserving nearest neighbor search. In *Machine Learning in Cyber Trust*, pages 247–276. Springer, 2009.

[82] Ali Inan, Selim V Kaya, Yücel Saygın, Erkay Savaş, Ayça A Hintoğlu, and Albert Levi. Privacy preserving clustering on horizontally partitioned data. *Data & Knowledge Engineering*, 63(3):646–666, 2007.

[83] Da Meng, Krishnamoorthy Sivakumar, and Hillol Kargupta. Privacy-sensitive bayesian network parameter learning. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 487–490. IEEE, 2004.

[84] Ming-Jun Xiao, Liu-Sheng Huang, Yong-Long Luo, and Hong Shen. Privacy preserving id3 algorithm over horizontally partitioned data. In *Sixth international conference on parallel and distributed computing applications and technologies (PDCAT'05)*, pages 239–243. IEEE, 2005.

[85] Boris Rozenberg and Ehud Gudes. Association rules mining in vertically partitioned databases. *Data & Knowledge Engineering*, 59(2):378–396, 2006.

[86] John Wang. *Encyclopedia of data warehousing and mining*. iGi Global, 2005.

[87] Jaideep Vaidya, Christopher W Clifton, and Yu Michael Zhu. *Privacy preserving data mining*, volume 19. Springer Science & Business Media, 2006.

[88] Timo M Deist, Frank JWM Dankers, Priyanka Ojha, M Scott Marshall, Tomas Janssen, Corinne Faivre-Finn, Carlotta Masciocchi, Vincenzo Valentini, Jiazhou Wang, Jiayan Chen, et al. Distributed learning on 20 000+ lung cancer patients–the personal health train. *Radiotherapy and Oncology*, 144:189–200, 2020.

[89] Hiroaki Kikuchi, Chika Hamanaga, Hideo Yasunaga, Hiroki Matsui, Hideki Hashimoto, and Chun-I Fan. Privacy-preserving multiple linear regression of vertically partitioned real medical datasets. *Journal of Information Processing*, 26:638–647, 2018.

[90] Jin Li, Yu Tian, Yan Zhu, Tianshu Zhou, Jun Li, Kefeng Ding, and Jingsong Li. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. *Artificial intelligence in medicine*, 103:101814, 2020.

[91] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. Federated learning for privacy-preserving ai. *Communications of the ACM*, 63(12):33–36, 2020.

[92] Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

[93] Yaochen Hu, Di Niu, Jianming Yang, and Shengping Zhou. Fdml: A collaborative machine learning framework for distributed features. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2232–2240. Association for Computing Machinery, 2019.

[94] Charles A Shoniregun, Kudakwashe Dube, and Fredrick Mtenzi. *Electronic healthcare information security*, volume 53. Springer Science & Business Media, 2010.

[95] Maria Manuela Cruz-Cunha. *Handbook of research on digital crime, cyberspace security, and information assurance*. IGI Global, 2014.

[96] Fatima-Zahra Benjelloun and Ayoub Ait Lahcen. Big data security: challenges, recommendations and solutions. In *Web Services: Concepts, Methodologies, Tools, and Applications*, pages 25–38. IGI Global, 2019.

[97] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[98] Stephen T Margulis. Conceptions of privacy: Current status and next steps. *Journal of Social Issues*, 33(3):5–21, 1977.

[99] Yacine Djemaiel, Slim Rekhis, and Noureddine Boudriga. Trustworthy networks, authentication, privacy, and security models. In *Handbook of Research on Wireless Security*, pages 189–209. IGI Global, 2008.

[100] Xiaokui Xiao and Yufei Tao. Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 229–240. Association for Computing Machinery, 2006.

[101] European Commission. White paper on artificial intelligence: a european approach to excellence and trust. Technical report, European Commission, 2020.

[102] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. *Journal of Big Data*, 3(1):1–25, 2016.

[103] Wei Jiang and Maurizio Atzori. Secure distributed k-anonymous pattern mining. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 319–329. IEEE, 2006.

[104] Lu Li, Liusheng Huang, Wei Yang, Xiaohui Yao, and An Liu. Privacy-preserving lof outlier detection. *Knowledge and Information Systems*, 42(3):579–597, 2015.

[105] Bin Gu, Zhiyuan Dang, Xiang Li, and Heng Huang. Federated doubly stochastic kernel learning for vertically partitioned data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2483–2493. Association for Computing Machinery, 2020.

[106] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[107] Timo M Deist, Arthur Jochems, Johan van Soest, Georgi Nalbantov, Cary Oberije, Seán Walsh, Michael Eble, Paul Bulens, Philippe Coucke, Wim Dries, et al. Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: eurocat. *Clinical and translational radiation oncology*, 4:24–31, 2017.

[108] Chang Sun, Lianne Ippel, Johan Van Soest, Birgit Wouters, Alexander Malic, Onaopepo Adekunle, Bob van den Berg, Ole Mussmann, Annemarie Koster, Carla van der Kallen, et al. A privacy-preserving infrastructure for analyzing personal health data in a vertically partitioned scenario. In *MEDINFO 2019: Health and Wellbeing E-Networks for All: Proceedings of the 17th World Congress on Medical and Health Informatics*, volume 264, pages 373–377. IOS Press, 2019.

[109] Sam Fletcher and Md Zahidul Islam. Decision tree classification with differential privacy: A survey. *ACM Computing Surveys (CSUR)*, 52(4):1–33, 2019.

[110] Li Huang, Andrew L Shea, Huining Qian, Aditya Masurkar, Hao Deng, and Dianbo Liu. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics*, 99:103291, 2019.

[111] Alexandros Karakasidis and Vassilios S Verykios. A sorted neighborhood approach to multidimensional privacy preserving blocking. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pages 937–944. IEEE, 2012.

[112] Aleksandra B Slavkovic, Yuval Nardi, and Matthew M Tibbits. " secure" logistic regression of horizontally and vertically partitioned distributed databases. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 723–728. IEEE, 2007.

[113] Johan Van Soest, Chang Sun, Ole Mussmann, Marco Puts, Bob van den Berg, Alexander Malic, Claudia van Oppen, David Townend, Andre Dekker, and Michel Dumontier. Using the personal health train for automated and privacy-preserving analytics on vertically partitioned data. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, volume 247, pages 581–585. IOS Press, 2018.

[114] Binnenlandse Zaken en Koninkrijksrelaties. Wet van 21 juli 2007, houdende algemene bepalingen betreffende de toekenning, het beheer en het gebruik van het burgerservicenummer (wet algemene bepalingen burgerservicenummer). 2018-07-28.

[115] R. Schnell. Efficient private record linkage of very large datasets. In *59th World Statistics Congress of the International Statistical Institute*. International Statistical Institute, 2013. Copyright 2013, the authors.

[116] William E Winkler. Record linkage. In *Handbook of statistics*, volume 29, pages 351–380. Elsevier, 2009.

[117] Rob Hall and Stephen E Fienberg. Privacy-preserving record linkage. In *International conference on privacy in statistical databases*, pages 269–283. Springer, 2010.

[118] Peter Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[119] Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38(6):946–969, 2013.

[120] Adrià Gascón, Phillipp Schoppmann, Borja Balle, Mariana Raykova, Jack Doerner, Samee Zahur, and David Evans. Privacy-preserving distributed linear regression on high-dimensional data. *Proceedings on Privacy Enhancing Technologies*, 2017(4):345–364, 2017.

[121] Trisha Greenhalgh and Richard Peacock. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *Bmj*, 331(7524):1064–1065, 2005.

[122] Matthias Egger, George Davey-Smith, and Douglas Altman. *Systematic reviews in health care: meta-analysis in context*. John Wiley & Sons, 2008.

[123] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[124] Phillip K Chan, Salvatore J Stolfo, et al. Toward parallel and distributed learning by meta-learning. In *AAAI workshop in Knowledge Discovery in Databases*, pages 227–240, 1993.
[125] Magdalena Kędzior. The right to data protection and the covid-19 pandemic: the european approach. In *ERA Forum*, pages 1–11. Springer, 2020.