e-ISBN: 978-93-6252-164-4
IIP Series

PRIVACY PRESERVATION TECHNIQUES AND VULNERABLE ATTACKS IN DATA MINING

Abstract

Numerous industries, including marketing, sports, data and network security, signal processing, and medical, can profit from data mining approaches. However, "privacy" has grown to be a significant issue, particularly in applications such entail the collecting and distributing individual data, despite the fact that data mining techniques are utilized in security issues like identifying intrusions.Two sections of this chapter look at the previously suggested PPDM approaches. Methods for gathering, cleaning, integrating, choosing, and transforming the input information that will be used for data mining are covered in the first section, while techniques for processed data are covered in the second. This chapter concludes with assaults against data mining applications' privacy.These factors make the issue of privacy protection in the environment of data mining different from the problem with conventional data privacy protection since data mining can be both an ally and an adversary. This chapter covers methods recommended for processed information (the result of the data mining techniques) as well as methods suggested for the initial information that would be subjected to data mining. Highlights assaults on data mining applications' privacy as well. A review of future safeguarding privacy data mining applications for both individuals and organizations closes the chapter.

Keywords: PPDM techniques, association rules, and privacy attacks, Anonymity, vulnerability.

Authors

Shailesh Kumar Vyas
Department of Computer Science & Engineering
G H Raisoni University
Saikheda, India.
shailesh.pk.29@gmail.com

Swapnili Karmore
Department of Computer Science & Engineering
G H Raisoni University
Saikheda, India
swapnilikarmore@gmail.com

I. INTRODUCTION

Particularly in light of the 2019 pandemic, rapid and extensive data sharing is made possible by social media, which sadly works against privacy in a society where commerce and education are conducted electronically over the internet. Interest in privacy has also grown as a result of the data mining techniques' quick and broad adoption in industries like marketing, sports, medical, and signal processing. Here, defining the parameters of the term "privacy" and offering a precise definition are crucial.People use the phrase "keep data concerning myself from being made available to others" to define privacy. People, however, are not alarmed by this circumstance and do not believe that their privacy has been breached when it comes to the use of their private information in a research project that is thought to be well-intentioned [1]. The difficulties of stopping abuse once the material is made public is overlooked in this situation.

The majority of privacy-preserving data mining techniques apply privacy preservation through various data transformations. To maintain anonymity, these techniques typically reduce the level of detail in the description.They can, for example, extrapolate data from individual users to user groups. Data is lost as a result of this granularity decrease, and the usefulness of the statistical mining findings is likely also lost. This is how privacy and data loss are traded off.

Information pertaining to a recognized or recognizable person is known as personal data. The elements of this concept are that the information relates to an individual and that the individual may be identified. The concept of "personal data" is associated with the "ego" and encompasses a wide range of information, including names, preferences, feelings, and thoughts.A person who can be directly or indirectly identified, especially by using a credential or a combination of features unique with their biological, mental, cultural, or social identity, is said to be identifiable. Because of this, losing control over one's data also means losing one's independence, autonomy, and privacy—in other words, the thing that makes one uniquely myself. Removing an individual's capacity to be identified is the primary method of ensuring that this data be used without endangering their privacy.

Data is turned into a commodity and has economic worth through data analysis techniques like data mining. Unquestionably, the digital world raises the possibility of losing autonomy and infringing upon one's right to informational privacy—aside from the ethical arguments surrounding this. The primary conundrum here is that the idea of being an individual is in opposition to the freedom in information flow, interest relationships, and benefits supplied through the information source that technology offers [2].

Governments also create legal restrictions that protect personal data by defining its intended use (historical, statistical, commercial, or scientific), how it should be gathered, and how long it should be kept. The US HIPAA regulations, for instance, are designed to safeguard personally identifiable health information. These data comprise demographic information gathered from an individual and are a subset of health data [3].The European Parliament & Council permit the processing of private information in the EC95/46 [4] regulation if: (i) the information subject has expressly granted consent; or (ii) the use is necessary to achieve a result that the individual has requested. Concerning corporate privacy, this is also applicable. Corporate privacy problems are accompanied with privacy concerns.

However, there aren't many differences between privacy issues pertaining to individuals and corporations. Information about an organization being made public may constitute a privacy violation. In this instance, generalizing knowledge regarding a particular group of data includes considering both points of view.

It is important to remember that, in addition to the exposure information provided by subjects, the organization's secrets held by data providers should also be considered. For instance, taking into account that in an academic study, data mining investigations were conducted on student data from multiple universities. Even though the techniques employed safeguard students' privacy, some information that is university-specific and they choose to keep private might be disclosed.

Data mining solutions that consider data privacy are being proposed in the literature. Information about collecting as a whole can still be published using a mechanism that guarantees no personal data is revealed. Data mining is frequently used to obtain this kind of corporate information, but since certain outcomes can be identified, many methods of data concealing has been explained in this chapter.

Privacy and security have always been top priorities. It sought to forecast the future by utilizing historical data. Let's say we purchase anything, in which case they focus our personal information and create predictions based on previous transactions. The ongoing advancement of data mining technology poses significant risks to the confidentiality and security of data, which must be preserved. The actual risk is that it will be impossible to prevent misuse if information is made available to unapproved parties. As a result, we require a system that can safeguard information as well as its resources in terms of integrity and authenticity.

Association rules search for correlations among the variables in a data set. Companies can often comprehend how many items relate to one another thanks to the rules, which can appear frequently across a data set. Gaining additional insight into consumers' purchasing and consumption patterns is one of the more popular applications employing association rules in the field of data mining. For instance, it's possible that female consumers in the 30–40 age range are more inclined to buy shampoo and conditioner together. With that knowledge, a shop could suggest conditioner to consumers who are considering shampoo, or the other way around.

There are further situations where using association norms could be advantageous besides shopping. For example, when diagnosing patients, physicians can apply association rules. A healthcare professional can document a patient's diagnosis after comparing the patient's present symptoms to those of others who have had comparable conditions. Over time, the practice will become increasingly accurate as an increasing number of cases are diagnosed and the algorithm may adjust itself to better represent such relationships. It should be noted that in this instance, information about patients would have to remain anonymous due to stringent healthcare privacy laws. Association rules are also frequently used by recommendation engines. Users who have comparable viewing habits and their own watch histories are used by platforms as Netflix and Hulu to suggest episodes and films to other users.

Before engaging in data mining, a person or business must obtain pertinent data. Sometimes people give out their info freely. For instance, when users register for social media accounts, they divulge a great deal of information. They frequently add their location, date of birth, college, and names of their kids to their online profiles. In other cases, obtaining pertinent information requires further investigation on the part of the person wishing to undertake data mining. They could go through open records or purchase a data set through a business that monitors customer behavior.

II. INFORMATION CONFIDENTIALITY

An important need of data privacy in this procedure is the purpose for which the data is going to be examined the people who receive it's going to be shared, the location of its transfer, and the data subject's ability to govern it transparently and controllably. However, there is no precise definition of privacy; instead, it might be defined in a way that is application-specific.Data controllers are presumed to be trustworthy and to be bound by law when it comes to taking privacy safeguards to prevent data breaches. They are also required to retain and utilize the data obtained through digital apps appropriately, sharing anonymized versions of the data when needed. Gathered information is characterized as following [7];

1. **Identifiers (ID):** This category includes details like a person's complete name & their federal identification number that can be used to directly and uniquely identify them.

2. Identifiers that, when paired with outside information, provide an indirect means of identifying a person are known as quasi-identifiers, or QIDs. These characteristics include non-unique information like age, gender, and postal code.

3. **Susceptible Attributes (SA):** It includes information like health and salary that is personal and sensitive to an individual.

4. **Insensitive Characteristics:** It includes generic, safe data that isn't addressed by other characteristics.

III. PRIVACY PARAMETERS

Measuring privacy using one indicator is insufficient since various applications may require distinct definitions, necessitating the evaluation of several factors. Depending on which component of privacy is investigated, the suggested indicators for PPDMs [8, 9] can be examined as data quality metrics and privacy level metrics.To assess the degree of privacy/data quality upon the information being provided (data criterion) and data miningoutputs (result criteria).

1. **Binding Knowledge:** By introducing noise or generalizing the data, it can be converted into limited data that is considered as binding of knowledge.

2. **Important to Know:** This statistic helps prevent privacy issues by preventing pointless data from entering the system. Additionally, it guarantees data access control, including access rationale and access authorization.

3. **Secured from Transparency:** Certain procedures (such verifying the searches) can be performed on the outcomes in order to provide privacy and preserve the private information that may surface as the outcome of data mining. One of the most efficient ways to guarantee privacy is to use the categorization method to stop data exposure that is one of the requirements [11].

4. **Metrics for Measuring the Quantity of Data:** They quantify information loss and benefit. Within this scope, complexity criteria measuring the effectiveness and scalability of various strategies are assessed.

5. **Data Mining's Effects on Society:** Data mining has significantly changed the way we work, shop, and get information. It also saves us time by providing personalized suggestions for products based on our past purchases from sites like Amazon and Flipkart.

Emerging across every industry, including social media, healthcare, finance, and marketing, is data mining. Raising operational efficiency, lowering expenses, increasing patient happiness, and delivering better patient-centered care by using data mining, insurance companies can identify instances of health insurance fraud and misuse that minimizes their losses.

Different types of transactions have been accepted by an outdated payment system based on factors like availability, acceptability, technology, techniques, and usage. It converts real-world financial transactions into digital ones. Thus, data mining monitors fraudulent transactions while concentrating on successful ones.

Additionally, it is a component of web-wide monitoring technology that monitors users' interests across all websites they visit. As a result, every website has information that is recorded and can be utilized to send marketers information about your interests.

Additionally, it is utilized for customer relationship management, which aids in giving each consumer individualized, more customized service. Businesses can target promotions and ads to customers who fit their profile and are more likely to be interested in receiving them than to become irritated with unsolicited emails by looking up their browsing and purchase histories on online stores. This lowers expenses, eliminates time wastage, and boosts productivity at work.

IV. DATA MINING WHILE MAINTAINING PRIVACY

The development of Privacy Protected Data Mining (PPDM) approaches has made it possible to extract information from data sets without disclosing sensitive information or the names of data subjects. Furthermore, several researchers can work together on a dataset using PPDM [11, 12]. Moreover, PPDM is the process of conducting data mining on sets of data that are to be acquired from databases holding private and sensitive information within a multilateral setting without sharing each party's data with third parties [13].

Methods based on statistics and cryptography have been proposed to safeguard privacy in data mining. To preserve privacy, the great majority of these methods use original

data. This is known as the inherent trade-off between the degree of privacy and the quality of the data.

In order to provide a certain degree of privacy while doing efficient data mining, PPDM techniques are being researched. For these techniques, numerous taxonomies have been suggested. The literature categorizes data according to the stages of the data life cycle (data collection, publication, distribution, as well as outcome of data mining) [10] and according to the technique employed (anonymization, perturbation, randomization, condensation, and cryptography) [14].

This study looks into PPDM approaches using a basic taxonomy as techniques for handling input and transformed data and the result information that is going to be used for data mining.

V. TECHNIQUES USED WITH THE PROVIDED DATA

The techniques for gathering, sanitizing, integrating, choosing, and transforming the input information that will be used for data mining are included in this section.It is advised for the initial values not be saved and be utilized solely in the conversion process to avoid confidentiality leakage. For instance, data gathered from sensors which are increasingly frequently employed in conjunction with the internet of things can be modified at the point of collection, generating random values from the received data, and changing the unprocessed data.

VI. DATA PERTURBATION

By considerably altering the data while maintaining its statistical integrity, it is possible to create data that is robust to privacy attacks [15, 16]. In data perturbation, it is common practice to randomize the original data [17, 18,19]. The Micro aggregation method is an additional strategy [20].

By adding noise signals that have a predefined statistical distribution to the data, data providers send their data to the data receiver after first randomizing it. The data receiver then uses distribution reconstruction techniques to determine the distribution after receiving this random input.

It can be computed separately for every data during the data collecting phase, and the statistical features of the data are maintained when the initial distribution is recreated. For instance, if A is the initial distribution of data and B is a widely available noise distribution that is independent of A, the outcome of randomizing P using Q is R (R = P + Q). Then, A might be rebuilt using the formula "P= R− Q." However, if Q has a high variance and R's sample size is insufficient, this reconstruction procedure might not work.Methods that apply the EM [22] or Bayes [21] formulas can be applied as a remedy. The randomization method conceals outliers by requiring a large amount of noise, even if it restricts the use of data to the distribution of R. Because, with this method, values in denser sections of the data are less susceptible to attacks than outliers. This lessens the value for the data for mining, but if you want to prevent information loss, it might be required to add excessive noise to every record in the data [7].

The micro aggregation approach divides the data set into a predetermined number of subsets after all of the records are first placed in a meaningful order. The value of that characteristic within the subset will be substituted with the average value, which is calculated by averaging the values of each subset of the given attribute. As a result, the attribute's average value throughout the whole data set won't change.

In utility-based data models, data perturbation procedures are typically not favored since they negatively affect data utility and are not robust to attacks.

VII. DATA SUPPRESSION APPROACH

The goal of the data suppression approach is to stop sensitive data from being revealed by substituting unique values for some of the original values. It can sometimes include erasing the record as a whole or specific cell values [24]. This allows private information to be modified, rounded, combined, or generalized before being made available for use in applications that use data mining [25].

These techniques may lead to altered general statistics and decreased data quality in big data, which could render the data useless [26]. The purposeful distortion of facts for the purpose of concealment is another issue. With the stated values, data providers can draw erroneous false inferences that accomplish a goal [27].

Conversely, if full knowledge of sensitive values is needed for data mining, suppression shouldn't be employed. Restricting the identification link of a record could be a better option if the record contains sensitive information.

VIII. DATA SWAPPING TECHNIQUE

By changing values between various records, a method seeks to prevent the leakage of private information. When multiple data providers are involved, data swapping refers to the process by which each supplier jumbles data by trading it with other providers. The method's benefit is that the data has no effect on the sub-order sums, enabling precise and thorough collective computations.

It is advised to utilize this technique only in secure contexts because it might readily disclose private data within the system as a result of data exchanges. Private definitions can be upheld when combined with other techniques like k-anonymity.

IX. COLLABORATIVE ANONYMIZATION

Table 1 compares different collaboration anonymity techniques.

Table 1: Collaborative Anonymization Techniques

Approach	Focused on	Susceptible	Robust against
k-anonymity	Disclosure of sensitive data	Attack on homogeneity	only using record linkage

l-diversity	Similarities in meaning between sensitive data	Attack of skewness	The linking of records and attributes
t-closeness	Measures of distance	Attack with attribute linkage	Attack probabilistic and attribute linking

1. ***K-anonymity:*** Certain attributes can be swapped out for more generic numbers (data exchanging), certain data points may be eliminated, while data that is descriptive can be suppressed (removed from the representation) in order to lessen the degree of detail in the data. Nevertheless, whereas k-Anonymity shields against identity revelation assaults, it is not impervious to attribute disclosure attacks.

2. ***l*-diversity:** In 2007, Ashwin Machanavijjhala introduced the l-diversity technique as a solution to the homogeneity attack flaws in the k-anonymity paradigm [34].The objective of this approach is to guarantee that every QID group contains a minimum of l well-represented sensitive values, hence preventing the indirect disclosure of sensitive information.L-diversity does not address the issue of diverse values belonging to the same category of important characteristics.Stated differently, it is not impervious to attacks that arise from similarities in semantics between values.

X. MINING ASSOCIATION RULES ACCORDING TO PERTURBATION

Rules develop in the data collection and are evaluated using statistical significance, support, and confidence as metrics. Every association rules are higher than or equal to the support and confidence that users have established; however, certain rules are sensitive from the user's perspective, while others are not. The association rules concealing method is to

To purify the original data set, apply the next procedure.All sensitive criteria are restricted to appearing on the original data mining; once the data set has been cleaned, they cannot appear concurrently with (or higher than) the confidence and support. That non-sensitive rules that can be found in the initial data set can be found with the same level of confidence and support on the cleaned data set.

Sensitive rules that were not discoverable in the initial set of data were not discoverable with the same level of confidence or support in the purification data set. Large item sets can be hidden using association rules mining, but the best filtering is NP hard [7]. Citation[8] suggested a significant project to clear sensitive rules and purify them. In order to minimize the assurance of the vulnerable rules by bringing them under a user-specified threshold or to prevent the vulnerable rules from being produced altogether, the frequently used item sets from which they originate were hidden in this work. Three methods for concealing delicate regulations were developed as a result of these two techniques.

XI. BLOCK-BASED MINING ASSOCIATION RULES

The data block is another perturbation for the association rules of the data modification approach [6]. The blocking method replaces a data item's property value with a question mark. In medicine, it is common practice to use unknown values in place of genuine values rather than false values. A blocking technique for association rules mining was

presented in Reference [7]. It appropriately modifies the definition of the minimum support and substitutes it with an appropriate supporting gap and minimal confidence, which is then replaced with a confidence interval. As far as support for sensitive rules is under the middle of the support interval and confidence in sensitive rules is under the middle of the confidence interval, we consider that privacy has not been infringed.

Whether a question mark needs to be linked to a 1-value or a 0-value; otherwise, the question mark's original value will be revealed. A thorough explanation of the blocking method's efficacy—a technique for reconstructing the text based on disruption rules—can be found in reference [8].

Mining Classification Rules Using Block a new framework integrating parsimonious downgrading and classification rule analysis is provided in Reference [12]. The data administrator's objective in this framework is to block values for the class label. This will prevent the information receiver from developing insightful models for the downgraded data.

The process of removing information from a data collection in order to downgrade information is known as parsimonious downgrading, and it has a formal structure. A cost measure is used to potentially degraded information in parsimonious downgrading to ensure it is not sent too low. The primary objective is to be the goal of this effort is to determine if the additional secrecy justifies the functionality loss that results from not downgrading the data.

XII. TECHNIQUES USED IN PROCESSED DATA

Without granting public access to the underlying data collection, information might be revealed via the results of data mining algorithms. Research on the outcomes provides access to sensitive data. Data mining results must therefore safeguard privacy.

1. Query Audit with Controlled Inferences

This approach is investigated as query auditing and query inference control. The query's output or input data are regulated within the query inference control. The queries that are run on the data mining outputs are audited in query auditing. The audited query request is rejected if it makes private information available. It actively contributes to maintaining privacy even while it restricts data mining. You can perform query auditing offline or online. It is determined whether the results infringe privacy because the inquiries along with outcomes have been identified in offline control.

2. Differential Privacy

The comprehensive approaches of k-anonymity, l-diversity, and t-closeness aim to safeguard the privacy of all data. In certain situations, record-level data privacy protection is required. Dwork has therefore suggested a differential privacy technique to safeguard the confidentiality of database query outcomes [37]. This model focuses on possible attacks that could happen in the interim between submitting a database query and getting a response. The presence of one record between databases won't be revealed if it isn't possible to determine which database supplied the response to the same query done in several databases.

3. Association Rule Mining

Certain rules may specifically reveal specific details during data mining. Certain connections may involve rules that are superfluous or reveal confidential information. By concealing all sensitive rules, the Association rule concealment technique which was first put forth by Atallah [38] aims to safeguard privacy.

4. Mining Data Association Rules with Horizontal Partitioning

The transactions in a database with horizontal partitioning are split among n locations. The total of all the local support counts for an item set is its total support count. If X's global support count exceeds s% of all items, then that item set is considered globally supported. The size of the transaction database overall.

5. Mining of Vertically Divided Data Association Rules

Different properties for each item at different sites are part of a vertically partitioned data collection. Finding an item set's support count allows for the mining of confidential association rules using vertically partitioned data. If such an item set's support count can be safely Therefore, we can determine whether the item set is frequent by seeing if the support is higher than the criterion. The crucial step to computing the vector dot product is for each party participating in the calculation to determine the total number of the item set support formed of a vector within the sub-item set. As a result, supports can be computed in a secure manner if the result of dots can be processed securely.

6. Restored Technology

Many of the recently proposed privacy-preserving data mining techniques employ data reconstruction or perturbation at the data convergence layer. Reference [10] investigated how to build a decision tree predictor with training data consisting of each record's perturbation value. Since the initial principles of each because records are not reliable, the author believes that an accurate approximation of the original distribution is possible. Bayesian approach is considered to rebuild the original distribution.

Reference [11] uses the EM algorithm in the scattered data to enhance the Bayesian reconstruction process. In other words, the author not only illustrated how the EM algorithm determines the maximum estimate that is pretty consistent with the initial information on the disruption distribution, but also demonstrated that when substantial amounts of data can be acquired, the robust original distribution can be estimated by the EM technique. Reference [10] also demonstrates that the estimation of privacy will decline if the data miner knew the background through the reconstruction distribution.

7. Protection of Anonymous Privacy

Anonymous publication chose to disclose the raw data. Sensitive data does not publish or distribute sensitive data with less precision in order to safeguard privacy. The current study concentrated on the technical aspects of data anonymity, specifically, Create trade-offs between the dangers of privacy exposure and data value, which

Release of sensitive information and data selectively, taking care to guarantee that the danger of sensitive information and privacy disclosure is kept within a reasonable bound. Data anonymity is concerned with two things: first, improving anonymization techniques is one of the guiding principles. This way, data published in accordance with the concept can better safeguard privacy while also being extremely useful. Creating more effective anonymity algorithms for a particular anonymous principle is the second option.The focus of research shifts towards how to accomplish the practical application of anonymous data as a result of the depth of anonymity research.

The k-anonymity principle, which was developed by Samarati and Sweeney, states that no record in the published table may be distinguished from any other k-1 record [12]. When k records are indistinguishable, we refer to them as an analogous class. In terms of insensitive properties, there is no distinction to be made here. Generally speaking, higher kValues improve privacy protection to a greater extent, while information loss rises. One of k-anonymity's shortcomings is that it does not impose any constraints on sensitive material. An attacker may identify private information or sensitive data by using background knowledge attacks and protocol against attacks [13], which could result in privacy breaches.

Based on this, (π, k)-anonymity [14] provide an upgrade that guarantees both the satisfaction of k-anonymity publishing and the fact that the percentage of π is not exceeded in any record pertaining to any attribute value in any disclosed equivalency class.

Generally speaking, generalization approaches are used in data publication methods like k-anonymity, l-diversity, t-closeness [15] as well as anonymous release, which significantly reduces accuracy and data utility. In the context of gathering information, if all sensitive data in data set D, which was made public by data owners, has disclosure risks, are smaller than the cutoff , [0,1], the data set's disclosure risk is referred to as . The privacy risk of released data sets is guaranteed to be less than 1/l by means of static data release variety [13] and less than 1/m by use of dynamic data publication principles m-invariance [16].

8. Vulnerable Attacks

This section provides a summary of the typical attack types that prompted the creation of the aforementioned techniques and resulted in privacy violations [6].

9. Attacks using Sementic Similarities

Attacks that take advantage of the sensitive attribute values' perceived resemblance within anonymous groupings. In this instance, privacy protection requires more than just the sensitive characteristic's values to differ from one another [40]. By determining how similar sensitive attributes are inside an identical anonymous group and by offering ways to include comparable values for sensitive attributes across groups, this attack can be stopped.

10. Attacks using Prior Knowledge

Background knowledge is insensitive knowledge that can be found using social engineering techniques or data released by various organizations, social networks, and the media. Attackers' background information leads to privacy breaches and attacks.The use of

data binding techniques to link background information with other records results in a breach of the privacy of the data subject [41].

To calculate the anatomized tables, the authors also suggested an anatomizing algorithm. First, the sensitive property is used by the algorithm to hash the records into buckets; that is, records that have the same sensitive information are placed in the same bucket. The algorithm then chooses a single entry from each of the! Buckets to create a group by repeatedly obtaining the Buckets with the most records at the time. Next, each record that remains is allocated to an already-existing group.

11. Attacks on Homogeneity

The privacy of information owners may be violated when a majority of the sensitive characteristics in the categories that comprise the anonymous databases are comparable. Avoiding homogeneity attacks requires either reproducing varied records by diluting homogeneous characteristics using the record duplication strategy, or prohibiting identical sensitive attributes inside the groups of records within the anonymous database from belonging to the same group [34].

12. Attacks with Skewness

Skewness attacks on privacy can be successful if the normal distribution of private attribute values in shared or published anonymous data sets is known. When these values become overly prevalent, anonymous records become vulnerable and the normal distribution of critical traits becomes distorted [35].

13. Matching Attack Without Sorting

This attack is caused over time by publicly declaring generalized data that has already been released. Because of this, it is best to use tables that have already been published and to avoid sharing any new records that can result in data disclosure [44].The tuples that appear in the published table in that sequence are the basis for this attack. Although the sequence of arrays cannot be presumed because we have preserved the application of a relational framework, this is frequently problematic in real-world applications. Of course, it can be fixed by sorting the solution's tuples at random. If not, private information may be revealed through the publication of a related table.

14. Attack with Complementary Release

It is more typical for the attributes that make up the quasi-identifier to be a subset of the released attributes. Therefore, when a k-minimal solution is made public, it ought to be seen as incorporating additional outside data. Consequently, unless the subsequent releases are also generalizations, every of the published attributes must be regarded as a quasi-identifier to prevent linking in subsequent releases of generalizations of the same privately owned information.

XIII. DISCUSSION

Data flow has accelerated as a result of the Covid-19 epidemic, which made digitization necessary worldwide. It is now much more crucial to get the required data, do accurate analysis, and produce trustworthy information. In practically every industry, this circumstance has led to the employment of data mining techniques to boost output and deliver superior goods and services. When using data mining techniques, it is clear that irreparable harm to people, institutions, and organizations will result from privacy not being taken into account throughout the data life cycle.

To optimize the utilization of data mining techniques and expand its reach, it is imperative that "privacy" be clearly defined, measurement criteria be established, and the outcomes assessed using these metrics prior to implementing PPDM techniques. This is why the concept of privacy was the main topic of this study. There is no accepted definition for the broad phrase "privacy." Since there is no accepted definition of privacy, quantifying privacy is very difficult. This chapter discusses a few measuring measures; however, metrics are typically defined by application.

Protecting one's privacy must be done both on an individual and an organizational level. Individual privacy protection is contingent upon the ways in which a person is shaped by their culture, religion, and social mores. Because of this, the idea of personalized privacy has been put forth, giving people some degree of control over their information. Personalized privacy implementation has proven to be challenging, nevertheless, as individuals tend to believe that giving up their privacy for what they believe to be well-meaning applications won't do any harm. Thus, new approaches to the contradiction among privacy and utility are needed in the environment of personalized privacy.

Organizational policy makers ought to promote privacy-enhancing technical designs and models for safe data collection, analysis, and sharing in order to successfully preserve data privacy at the organizational level [7]. Organizations should examine privacy-related laws, rules, and core values. Organizations must evaluate their security and privacy policies with input from the data owners. Owners of data should be involved in all aspects of the process, including what information is gathered, how it is processed, and why it is used.

Techniques for maintaining secrecy in data analytics remain in their infancy from a technical standpoint. An interdisciplinary study on PPDM is necessary, even though research in fields like cryptography, database administration, and data mining is still being done by many scientific communities. For instance, a legal viewpoint should be taken into consideration while addressing the challenges this process presents. Thus, academic academics and industry practitioners can jointly establish an improved roadmap towards future privacy-preserving data mining architecture.

The homogeneity attack & prior knowledge attack, as well as the temporal, complementary release, and unsorted matching attacks, are presented in this section and their potential uses for breaching a k-anonymous dataset are demonstrated.Thus, a new definition of diversity emerges here. Even in situations when the data provider is unaware of the type of knowledge the adversary possesses, variety preserves privacy. The fundamental tenet of l-

diversity is that each group must have a strong representation of the results of the sensitive characteristics.

Despite taking adequate precautions to determine the QI, attackers can nevertheless compromise the k-anonymity. Unsorted matching assaults, complimentary release attacks, and temporal attacks are examples of common attacks. Thankfully, there are best practices that can stop these attacks. However, the two main attacks—the homogeneity and background attacks—disclose private information about the individuals. Because k-anonymity can produce groups that leak information, it is not impervious to attacks based on prior knowledge.Thus, it may be concluded that background knowledge assaults are not prevented by k-anonymity. Considering the homogeneity and prior knowledge attacks, we have shown that sensitive information can be revealed in a k-anonymous table. Given the likelihood of the two of these attacks in the real world, a more thorough description of privacy that considers prior knowledge and diversity is required.

The drawback of the k-anonymity approach is that sensitive information can be derived for the modified data whenever there is a homogeneity of those values within a group. In order to achieve anonymization, the l-diversity framework was created to manage this vulnerability by requiring intragroup variation of sensitive parameters. The goal is to make it difficult enough for adversaries to precisely identify individual records using combinations of data properties.

XIV. CONCLUSION

In order to better serve their clients and people, businesses and even governments gather data via a variety of online channels (such as social media, e-health services e-commerce, entertainment, e-government, etc.). Sensitive information may be gathered, stored, examined, and most likely anonymized before being shared with others. Explaining a privacy authorization and the rationale for data access is required during research wherein data can be utilized at any point in the life cycle, for any cause. Techniques such as Privacy Preserving Data Mining (PPDM) are being developed to enable data extraction without revealing sensitive information.

No single PPDM technique is ideal for every phase of the data chain. The appropriate PPDM technique varies based on the needs of the application, including the required level of privacy, the volume and size of data, the amount of acceptable information loss, the complexity of transactions, etc. because the laws, presumptions, and requirements pertaining to privacy vary throughout application locations.

The techniques of k-anonymity and l-diversity these two techniques both modify individual data to prevent precise identification. The degree of data visualization is sufficiently lowered in the k-anonymity approach such that a particular piece of data maps onto at least k more records within the data set. It requires strategies like suppression and generalization.

Regarding privacy, security, as well as social effects, data mining offers a number of benefits and drawbacks. Before putting data mining techniques into practice, it's crucial to thoroughly assess the risks and effects, even though they can increase efficiency and yield

insightful information. Organizations must take action to reduce risks, safeguard people's security and privacy, and make sure data mining is done with an ethical and accountable way.

REFERENCES

[1] Clifton C, Kantarcioglu M, Vaidya J, Defining privacy for data mining. In National science foundation workshop on next generation data mining. 2002; Vol. 1, No. 26, p. 1
[2] İzgi M. C, The concept of privacy in the context of personal health data. TürkiyeBiyoetikDergisi, 2014. (S 1), 1
[3] Centers for Disease Control and Prevention. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. MMWR: Morbidity and mortality weekly report, 200352(Suppl 1), 1-17
[4] Data P, Directive 95/46/EC of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L, 1995; 281(23/11), 0031-0050
[5] Belsey A, Chadwick, R. Ethical issues in journalism and the media. Routledge. (Eds.) 2002
[6] Vural Y, VeriMahremiyeti: Saldırılar, KorunmaVeYeni Bir ÇözümÖnerisi. UluslararasıBilgiGüvenliğiMühendisliğiDergisi, 4(2), 21-34
[7] Pramanik M. I, Lau R. Y, Hossain M. S, Rahoman M. M, Debnath S. K, Rashed, M. G., Uddin M. Z.,. Privacy preserving big data analytics: A critical analysis of state-of-the-art. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021; 11(1), e1387
[8] Bertino E, Lin D, Jiang W, A survey of quantification of privacy preserving data mining algorithms, in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 183-205
[9] Dua S, Du X, Data Mining and Machine Learning in Cybersecurity. Boca Raton, FL, USA: CRC Press, 2011
[10] Mendes R, Vilela J. P, Privacy-preserving data mining: methods, metrics, and applications. IEEE Access, 2017; 5, 10562-10582
[11] Vaidya J, Clifton C, Privacy-preserving data mining: Why, how, and when. IEEE Security & Privacy, 2004; 2(6), 19-27
[12] Nayak G, Devi S, A survey on privacy preserving data mining: approaches and techniques. International Journal of Engineering Science and Technology, 2011; 3(3), 2127-2133
[13] Lindell Y, Pinkas B, Privacy Preserving Data Mining, In: Proceedings of the 20th Annual International Cryptology Conference, 2000; California, USA, 36- 53
[14] Rathod S, Patel D, Survey on Privacy Preserving Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT) 2020; Vol. 9 Issue 06
[15] Hong T. P, Yang K. T, Lin C. W, Wang S. L, Evolutionary privacy-preserving data mining. In: Proceedings of the World Automation Congress 2010; (pp. 1-7). IEEE
[16] Qi X, Zong M, An overview of privacy preserving data mining. Procedia Environmental Sciences, 2011; 12, 1341-1347
[17] Muralidhar K, Sarathy R, A theoretical basis for perturbation methods. Statistics and Computing, 2003; 13(4), 329-335
[18] Evfimievski A, Randomization in privacy preserving data mining. ACM Sigkdd Explorations Newsletter, 2002; 4(2), 43-48
[19] Kargupta H, Datta S, Wang Q, Sivakumar K, On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the Third IEEE international conference on data mining 2003; (pp. 99-106). IEEE
[20] Domingo-Ferrer J, Torra V, Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 2005; 11(2), 195-212
[21] .Agrawal R, Srikant R, Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data 2000; (pp. 439-450)
[22] Agrawal D, Aggarwal C. C, On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems 2001; pp. 247-255
[23] Niranjan A, Nitish A, Security in Data Mining-A Comprehensive Survey. Global Journal of Computer Science and Technology 2017
[24] Oliveira S, Zaiane O, Data perturbation by rotation for privacy-preserving clustering, Technical Report 2004

[25] Verykios V. S, Bertino E, Fovino I. N, Provenza L. P, Saygin Y, Theodoridis Y, State-of-the-art in privacy preserving data min ing. ACM SIGMOD Record, 2004; 33, 50-57
[26] Aggarwal C. C, On randomization, public information and the curse of dimensionality. In: Proceedings of the IEEE 23rd International Conference on Data Engineering; Istanbul, Turkey, 2007, pp. 136-145
[27] Zhu D, Li X. B, Wu S, Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. Decision Support Systems, 2009; 48, 133-140
[28] Yang Y, Zheng X, Guo W, Liu X, Chang V, Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. Information Sciences, 2019; 479, 567-592
[29] Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. IEEE Network, 28, 46-50
[30] Yao A. C, How to generate and exchange secrets. In: Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, 1986; 162-167. IEEE
[31] Goldreich O, Micali S, Wigderson A, How to play any mental game - a completeness theorem for protocols with honest majority. In: Proceedings of the 19th ACM Symposium on the Theory of Computing, 1987; 218-229
[32] Sweeney L, k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002; 10(05), 557-570
[33] Samarati P, Sweeney L, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI International, Technical Report, 1998; SRI-CSL-98-04
[34] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M, ℓ-Diversity: Privacy beyond k-anonymity, In: Proceedings of the The 22nd International Conference on Data Engineering, 2006; Atlanta, USA
[35] Li N, Li T, Venkatasubramanian S, t-Closeness: Privacy beyond k-anonymity and ℓ-diversity, In: Proceedings of the International Conference on Data Engineering (ICDE), İstanbul, Turkey, 2007; 106-115
[36] Rubner Y, Tomasi C, Guibas L. J, The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000; 40(2), 99-121
[37] Dwork C, Differential privacy: A survey of results. In: Proceedings of the International conference on theory and applications of models of computation Springer, Berlin, Heidelberg. 2008; (pp. 1-19)
[38] Atallah M, Bertino E, Elmagarmid A, Ibrahim M, Verykios V, Disclosure limitation of sensitive rules. In Knowledge and Data Engineering Exchange Workshop (KDEX'99), 1999; 25-32
[39] Evfimievski A, Srikant R, Agrawal R, Gehrke J, Privacy preserving mining of association rules. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002; 217-228
[40] Wang H, Han J, Wang J, Wang L, (l, e)- Diversity - A Privacy Preserving Model to Resist Semantic Similarity Attack, Journal of Computers, 2014; 59-64
[41] Chen B. C, LeFevre K, Ramakrishnan R, Privacy skyline: Privacy with multidimensional adversarial knowledge. University of Wisconsin-Madison Department of Computer Sciences. 2007
[42] Kifer D, Attacks on privacy and deFinetti's theorem", In: Proceedings of the ACM SIGMOD International Conference on Management of data, Rhode Island, ABD, 2009; 127-138, 2009
[43] Wong R. C. W, Fu A. W. C, Wang K, Pei J, Minimality attack in privacy preserving data publishing. In: Proceedings of the 33rd international conference on Very large data bases 2007; (pp. 543-554)
[44] Sanjita B. R, Nipunika A, Desai R, Privacy Preserving In Data Mining, Journal of Emerging Technologies and Innovative Research 2019; vol6 Issue 5.

