International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 06 | June - 2021 ISSN: 2582-3930

A Review: Anonymization Techniques for Preserving Privacy in Data Mining

[1]Prof. Rajyalakshmi Jaiswal Computer Engineering L.D. College of Engineering
[2]Namrata Govind Ambekar Computer Engineering L.D. College of Engineering

---

Abstract - Now a day in the era of internet, there is widespread expansion of information, knowledge and individual sensitive data. This has led to concern that the individual data may be violates and get misused for different purposes. In that concern, Number of researchers has been working in this area of privacy preservation data mining and they have proposed many techniques in order to preserve privacy to perform data mining task. Privacy preservation data mining (PPDM) is the process of masking and Erase the information available in data after released data securely. Several techniques of PPDM have been studied extensively which helps to preserve sensitivity of individual data to achieve privacy. Anonymization is one of the technique helps to preserve the privacy in data publishing. Various anonymization methods, models, algorithms, frameworks have been developed for privacy preserving in data publishing to protect the users privacy. This paper surveys different anonymization techniques and our aim is to give deeper insight to PPDM problem.

Key Words: Anonymization, Data Mining, Generalization, Information Loss, Privacy Preserving, Suppression, Permutation.

1 Introduction-

In recent days data mining can be seen as threat to the privacy because data mining is the process of collecting valuable and relevant information from large volume of available database. When internet become the part of life with social media platform, blogs, forum, net banking , ecommerce etc. helps to give data mining sector massive attention to the available large amount of data collected easily and saved through the internet. The concern is personal privacy violation is growing due to proliferation of internet users. People at threat that there personal data may gained and misused by unauthorized person. Various data owners such as banks, hospitals, insurance companies, credit card companies, educational institutes they itself anonymized their data values before releasing data to required clients or organization for analysis and other data mining activities. Data owner want a way to transfer data consisting sensitive information in secured way to protect personal sensitive data, therefore different techniques of privacy preservation have been developed such as Anonymization, Perturbation, Randomization, Distributed privacy preservation method, Cryptography methods etc. In this survey paper, concern is to provide enhanced privacy with the help of different anonymization techniques. Data anonymization is a technology that converts clear text into a non-human readable form. Data anonymization technique for privacy-preserving data publishing has received a lot of attention in recent years. Through this comprehensive overview, we hope to give a solid foundation for future studies in PPDM area.

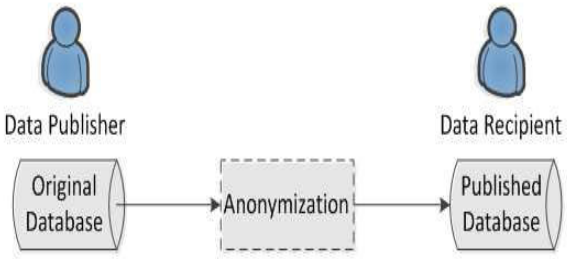

Figure: Anonymization of data

2 Backgrounds

Privacy is a prime factor to be considered when data is published on web and other data repositories. Recent day's big data is mined in large scale to find out relevant and beneficial information and concern is about the privacy of individual sensitive data. Privacy is all about to keeping personal information away from publically access. Privacy is needed for personal respect and individualism. This survey focuses on information privacy of published data. We present the various anonymization methods that were proposed to anonymized users' data to protect user's privacy.

2.1 Anonymization Technique

Anonymization is a data processing technique that removes or modifies personally identifiable information; it results in anonymized data that cannot be associated with any one individual due to this identification of distinct individual is not possible. Anonymization aims at making individual records be indistinguishable among the group of records and allow to sharing such data without compromising privacy of individuals. In particular records of database consist of different attributes such as explicit identifiers, quasi identifiers, sensitive attributes, non-sensitive attributes etc. Explicit identifiers are also known as unique identifiers or direct identifiers which can directly identify particular record owner in database. For example name, SSNo. Quasi identifiers are those attributes which could potentially identify record owner. If they are linked with some external attributes of database that can easily re-identify individual record. For example age, sex, zip code etc. Sensitive attributes are consist of persons specific information which is private and do not want to share it publically. For example attributes like salary, disability status, Disease etc.

2.2 Actors involved in Anonymization Process

Actors involves in typical privacy preserving data publishing (PPDP) or anonymization process are as follows

Record Owner: An entity that hold one or more record in a data. For example users, customers, patients.

Data Holder: A person or organization(s) that holds user records. For example banks, hospitals.

Data Publisher: A person or organization who publish the collected data. For example government non- government agencies.

Data Recipients: An entity who has access to anonymized datasets. For example researchers, data miners, analyst, Competitors.

Adversary: A person or an entity whose aim is to access sensitive information and misuse it [9]

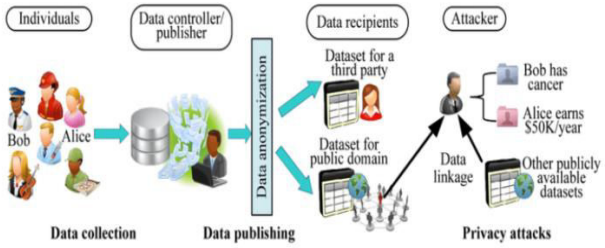

Figure: Privacy preserving data publishing overview

2.3 PPDP differs from PPDM in several ways

Privacy preserving data publishing (PPDP) and Privacy preserving data mining (PPDM) are two techniques which can be used to preserve privacy.

1) Main difference between PPDM and PPDP is that PPDM focuses on extracting relevant and useful information from large amount of data. While maintaining privacy whereas PPDP focuses on publishing information collected from data providers instead of data mining outcomes.
2) PPDM focuses on data without sensitive information whereas PPDP focuses on anonymized data by hiding identity of individual without hiding sensitive information.
3) PPDM allows data mining techniques such as association rule mining, classification and clustering whereas PPDP used hiding techniques like K-anonymity, L-diversity, T-closeness etc.

2.4 Attacks on anonymization based techniques

Attacks are happen if adversaries have some additional information about the anonymization mechanism involved in database and propel an attack based on that knowledge of the database.

Homogeneity attack

When non-sensitive attribute is known by adversary then it is possible to identify sensitive information of record owner. This is known as positive disclosure.

Background knowledge attack

In this background knowledge attack, the adversary can use combination between one or more quasi-identifier attributes with the sensitive attribute in order to find out actual individual sensitive record value. It is called negative disclosure.

3 Literature survey

In this section, surveyed on different anonymization techniques proposed by various research

3.1 Anonymization techniques

1. K-anonymity

K- anonymity is one of the most popular models. This model helps to prevent combining attacks by generalization and suppression to the required attributes and after that published the Microdata so that no individual can be uniquely identified from a group of size k.(JIS attacks on anonymization) The K-anonymity model was developed because of the possibility of indirect identification of records from public databases. The concept of K-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published in 1998 as an attempt to solve the problem: "Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful. A release of data is said to have the K-anonymity property if the information for each person contained in the release cannot be distinguished from at least K-1 individuals whose information also appear in the release. A table is said to be k-anonymous if it is indistinguishable from minimum (k-1) records with every quasi identifiers that it is selected [7]. To reduce the granularity of data representation we can use generalization and suppression on quasi identifier attributes. K-anonymity model suffers record linkage attack. Authors show that k-anonymity does not provide sufficient protection against attribute linkage

Table 1: Patient Microdata

Name	Age	Gender	Disease
Allis	65	Patient	Cancer
Alex	50	Patient	Flu
Bobby	70	Patient	Gastritis
Leena	35	Patient	Pneumonia
Linda	20	Patient	Flu

Table 2: Anonymous Patient Microdata data of Table 1

For example, as shown in above Table 1 we have patient Microdata in this attributes age, gender are quasi identifiers. If we combining the values any of age, gender values, we can re-identified the actual record holder. In Table 2 we make the data anonymous by generalization so the attribute value less likely to be re-identified.

2. Generalization

Generalization is the technique which is used in k-anonymity model to anonymized data values. In this individual values of attributes are replaced with broader categories. This operation transforms the original QI's values into less-specific but semantically consistent
Values during anonymization process [6]. We generalize the attributes age, and gender. In generalization there are two categories like categorical data and numerical data. According to this age is numerical data value represents age in number form. And categorical data includes alphabets like in gender attributes values consisting male or female categories. For example we shown in able Table 1, attribute age values can be transforms into broad category. Allis age 65 is generalized into range values 50-90 as shown in Table 2.

3. Suppression

Group ID	Name	Age	Gender	Disease
1	*	50-90	Patient	Cancer
1	*	50-90	Patient	Flu
1	*	50-90	Patient	Gastritis
1	*	50-90	Patient	Pneumonia
1	*	50-90	Patient	Flu
2	*	10-50	Patient	Corona
2	*	10-50	Patient	TB
2	*	10-50	Patient	Aids
2	*	10-50	Patient	Cancer

Suppression technique is also used in k-anonymity model to remove the values of direct identifier attributes so that they are not released. In this suppression techniques explicit identifier attributes are replaced with some characters like '?', '*'. For example Table 1, Name is explicit identifier or direct identifier so it have to remove from the table so in Table 2 Name attribute is replaced by '*'.

Table 3: Anonymized Data

4. L- Diversity

K- Anonymity model is effective in preventing identification of record but it may not always be effective in preventing inference of sensitive values of the attributes of records. Therefore, to overcome shortcomings of k- anonymity model L-diversity model came into existence which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes [8].

5. T- Closeness

A k-anonymous and l-diverse dataset is said to have t-closeness if the distance
Between the sensitive attribute in the equivalent class is maximum or less than some threshold value t. While implementing k-anonymity and l-diversity model data publisher must effectively
choose the value of k and l. In t-closeness, the value "t" is just an abstract distance between two distributions, which could have different meanings in different contexts [5].

6. Permutation

A Permutation is defined as an act of arranging members of a set into in sequence or order [4]. In this operation, the records are partitioned into several groups, and values of the SA are shuffled within each group. Hence, the SA and QIs relationships are de-associated within each group. This operation may yield inaccurate analysis in terms of anonymous data utility, but user's privacy is significantly preserved. For example Table 3, In attribute Group Id we groups the attributes into 2 equivalence class so that we can apply permutation [3]

4 ARX Tool

ARX is comprehensive open source software for anonymization of sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes [2]. Many biomedical scientists are using this simulation tool for data de-identification model for implementation of different privacy models.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface.

5 Conclusion

In this review paper, we have addressed the problem the release of sensitive information and present different anonymization technique so that private information of particular entities will not

revel to public. We have illustrated the K-anonymity technique by using generalization and suppression technique. One more notable technique of anonymization is permutation. By this permutation we can shuffle the records such that adversaries will not able to identify the record owner. We have explained one of most famous open source ARX tool of anonymization which is also known as De-identification tool. Many biomedical scientists are using this tool for anonymize datasets.

5 References

[1] Ajit Singh, "Data Publishing and privacy ptreserving," International Journal for Information Security Research (IJISR), Volume 9, Issue 3, September 2019, pp 881-890

[2] https://arx.deidentifier.org/

[3] ABDUL MAJEED, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," 10.1109/ACCESS.2020.3045700, IEEE Access, pp 1-35

[4] Shastri, M.D., Pandit, A.A. Remodeling: improved privacy preserving data mining (PPDM). Int. j. inf. tecnol. (2020). https://doi.org/10.1007/s41870-020-00531-8.

[5] Ajit Singh, "Data Publishing and privacy ptreserving," International Journal for Information Security Research (IJISR), Volume 9, Issue 3, September 2019, pp
88-890

[6] ABDUL MAJEED, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," 10.1109/ACCESS.2020.3045700, IEEE Access, pp 1-35

[7] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int'l Journal on Uncertainty, Fuzziness, and Knowledge-based Systems,*2002.

[8] Charu C. Aggarwal, *"A General Survey of Privacy-Preserving Data Mining Models and Algorithms",pp 12-52*

[9] ABDUL MAJEED, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey," 10.1109/ACCESS.2020.3045700, IEEE Access, pp 1-35