International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST21842

Anonymization Techniques for Privacy Preservation in Social Networks: A Review

Kalpana Chavhan, Dr. Praveen S. Challagidad
Computer Science and Engineering Basaveshwar Engineering College Bagalkot, Karnataka, India

Article Info
Volume 8, Issue 4
Page Number : 14-21

Publication Issue
July-August-2021

Article History
Accepted : 25 June 2021
Published : 02 July 2021

ABSTRACT

Any data that user creates or owns is known as the user's data (For example: Name, USN, Phone number, address, email Id). As the number of users in social networks are increasing day by day the data generated by the user's is also increasing. Network providers will publish the data to others for analysis with hope that mining will provide additional functionality to their users or produce useful results that they can share with others. The analysis of social networks is used in modern sociology, geography, economics and information science as well as in various fields. Publicizing the original data of social networks for analysis raises issues of confidentiality, the adversary can search for documented threats such as identity theft, digital harassment and personalized spam. The published data may contain some sensitive information of individuals which must not be disclosed for this reason social network data must be anonymized before it is published. To do the data in nominate the anonymization technique should be applied, to preserve the privacy of data in the social network in a manner that preserves the privacy of the user whose records are being published while maintaining the published dataset rich enough to allow for the exploration of data. In order to address the issue of privacy protection, we first describe the concept of k-anonymity and illustrate different approaches for its enforcement. We then discuss how the privacy requirements characterized by k-anonymity can be violated in data mining and introduce possible approaches to ensure the satisfaction of k-anonymity in data mining also several attacks on dataset are discussed.

Keywords : Privacy Protection, Data Mining, K-Anonymous Method.

I. INTRODUCTION

Personal data is now entrusted to businesses, which they use to better support consumers and make better decisions, however much of the data's value remains untapped. Many organizations want to share this information while protecting individual privacy because it could help third- party researchers and analysts address questions on everything from community planning to cancer research. However, in order to obtain accurate analytical results, it is also essential to preserve the data's utility.

Data owners want a way to transform a highly sensitive dataset into a low-risk, privacy-preserving data set that they can share with everyone, from researchers to business partners. Organizations, on the other hand, are gradually releasing databases that they assumed were anonymized, only to discover that a large portion of the documents have been re-identified. It's crucial to understand how anonymization strategies operate, when and where they can be used safely, and what advantages and disadvantages they have.

This article discusses k-anonymity, a privacy technique widely used to protect data subjects' privacy in data sharing situations, as well as the benefits of using k-anonymity to anonymized data. Anonymity for data subjects is the end goal of many privacy-preserving systems. When taken at face value, anonymity simply means not being identified, but closer examination reveals that removing names from a dataset is insufficient to achieve anonymity. By comparing anonymized data to another dataset, it is possible to re- identify it. Data containing quasi-identifiers, which are bits of information that aren't unique identifiers but can be used to re-identify people. Data containing quasi-identifiers, which are bits of information that aren't unique identifiers but can be combined with other datasets to recognize individuals.

K-anonymity is a basic principle that has been created to prevent anonymized data from being connected to other datasets and being re-identified. When used correctly and with the required protections in place, such as access control and contractual safeguards, K-anonymization is a powerful tool. Along with other methods like differentially private algorithms, it's an important part of the privacy-enhancing technology arsenal. As big data becomes the standard rather than the exception, we're seeing more data dimensionality and a growing number of publicly accessible datasets that can aid in re-identification.

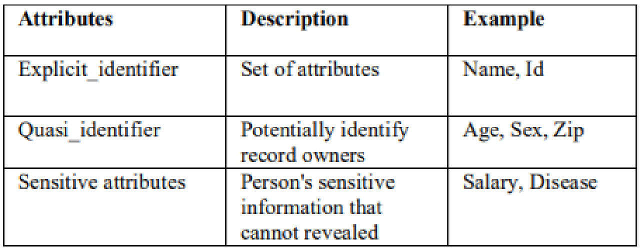
According to study, the majority of conventional k-anonymity approaches rely on generalization and suppression techniques. Since they depend heavily on ordering relations from predefined generalization layers on the attribute domains, they all suffer from substantial information loss. As a result, the anonymization results often result in substantial information loss and, as a result, low availability. Furthermore, current anonymization algorithms are primarily concerned with preserving private information while ignoring the practical usefulness of anonymized data. As a consequence, anonymized data is scarce in real-world situations.

The following are some key k- anonymity concepts.

K-anonymity

To ensure an individual's privacy, k-anonymity notes that there should be at least k tuples with the same quasi- identifier values. Only if each tuple in the table is equal to at least (k-1) other tuples may it achieve k-anonymity.

Table 1. K-anonymity attribute



Attributes	Description	Example
Explicit_identifier	Set of attributes	Name, Id
Quasi_identifier	Potentially identify record owners	Age, Sex, Zip
Sensitive attributes	Person's sensitive information that cannot revealed	Salary, Disease

Generalization and suppression are used to achieve K-anonymity. Table 1 shows the important attributes used in k-anonymity.

What's the Need of K- anonymity?

A health insurance organization from X city had compiled a database of hospital visits by state employees, and had thought that giving it to researchers could encourage innovation and scientific discovery. There were privacy considerations while using that dataset for research purpose. To allow researchers to look at other citizen's health records the organization decided to remove all columns that indicated who a patient was: name, phone number, full address, social security number, etc. Because of this the important information is loss and this didn't end so well.

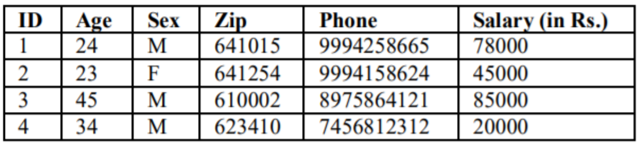
Some demographic information was left in the database, so researchers could still compile useful stats: ZIP code, date of birth, and gender were all part of the data. One of researcher realized that the claims of the X city governor, who insisted that the privacy of state employees was respected were perhaps a little bit over-optimistic. Since the governor himself was a state employee, Researcher decided to do the obvious thing and re-identify which records of the "anonymized" database were the governor's.

Researchers also bought the public voter records from X city, which had both full identifiers (names, addresses) and demographic data (ZIP code and date of birth), and contained the governor's information. With compiling the hospital data and voter's data, the researcher is able to find governors complete data as there are several fields are in common such as governor's gender, ZIP code, and date of birth. Only one, such record found in both dataset. so researcher is able to know which prescriptions and visits in the data were the governor's. Researcher posted all of it to governor's office, showing theatrically that their anonymization process wasn't as solid as it should have been.

K-anonymity Example

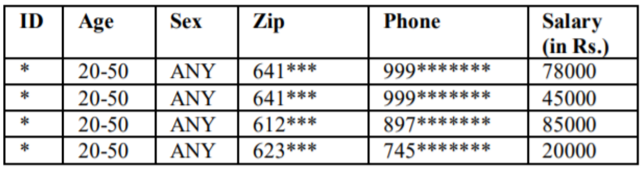
Example: If the above mentioned table is to be anonymized with Anonymization Level (AL) set to 2 and the set of Quasi identifiers as QI = {AGE, SEX, ZIP, PHONE}.Sensitive attribute = {SALARY}. The quasi-identifiers and sensitive attributes are identified by the organization according to their rules and regulations.

Table 2. Dataset Entries



ID	Age	Sex	Zip	Phone	Salary (in Rs.)
1	24	M	641015	9994258665	78000
2	23	F	641254	9994158624	45000
3	45	M	610002	8975864121	85000
4	34	M	623410	7456812312	20000

Table 3. Anonymized Data



ID	Age	Sex	Zip	Phone	Salary (in Rs.)
*	20-50	ANY	641***	999*******	78000
*	20-50	ANY	641***	999*******	45000
*	20-50	ANY	612***	897*******	85000
*	20-50	ANY	623***	745*******	20000

How to prevent attack?

Removing one of the factors should be enough to prevent attacks like these. Which ones can we afford to remove, while making sure that the data can be used for data analysis tasks. The suggestion is the use of idea of k-anonymity.

Generalization

The process of generalization is the transfer of any value to more general form. For example, "Male" and "Female" can be combined to form "Individual." The following stages of generalization can be used: Generalization is done at the column level; a generalization stage generalizes all of the values in the column.

Cell (CG): The generalization is done on individual cells; as a result, a generalized table may have values for a given column at different levels of generalization. The generalizing date, month, and year, for example, form different degrees of generalization in the case of DOB.

Suppression

Suppression is the process of completely removing a value from a data table. The levels of repression are as follows: Suppression is performed at the row stage, and it removes an entire tuple.

Attribute (AS): Suppression is done at the column level; a suppression operation masks all of the values in a column.

Cell (CS): Suppression occurs at the cell level; as a result, a k-anonymized table can only delete the cells of a given tuple/attribute.

II. Literature Review

Swagatika Devi at el in [1] addresses the idea of k-anonymity, from its inception to its implementation by generalization and suppression. They also go through the various ways that generalization and suppressions can be utilized for achieving k-anonymity. By relocating the concept of k-anonymity from data to patterns, they formally define the notion of a challenge to privacy in the context of pattern discovery. In their work, they introduced the l-diversity strategy, which focuses on retaining the diversity of sensitive attributes while maintaining a minimum group size of k. The authors offer a summary of the various strategies and how they are related. Their aim is to offer a new reader an overview of the field from the perspective of the data mining community.

Bettini, Wang, and colleagues et al [2] investigated the K- anonymity property to protect the location's privacy. When location information is published in the sense of location- based services, the author must have a mechanism for determining the privacy of the user's identity. Instead of a database of tuples, k-anonymity is ensured in this case by a group of individuals who can send a message in the same spatiotemporal sense.

To fix the issue of privacy security F. Song, T. Ma et al [3] suggest a modern k-anonymous approach that differs from conventional k-anonymous. Numerical data is made k- anonymous by inserting sounds, while categorical data is made k-anonymous by randomization. The drawback that at least k elements in the k-anonymous data set must have the same quasi identifier has been solved using the above two methods. Since seeking anonymous equivalence is time consuming, the original data set is divided into equivalence groups using a two-step clustering process. The equivalence classes are established in the sub-datasets after the original data set is divided into many different sub-datasets, significantly reducing the computational cost of identifying anonymous equivalence classes. They tested their proposed method on three separate datasets, and the results show that it is more effective, with far less information loss in anonymous datasets.

The de-identification of the 2013-2014 edX data set resulted in significant improvements to the data set, according to Daries et al. [4], including revisions to the profiles and grade distributions of students who completed the course. In the 2015 edX. data set follow-up analysis, Angiuli et al [5] discovered that suppression-heavy anonymization approaches skew column values, whereas generalization- heavy approaches skew column correlations. According to their findings, high-achieving students have rare quasi- identifying characteristics, making high-grade records more likely to be suppressed.

A clustering-based strategy for preventing data loss and thus maintaining high data quality has been proposed by Byun JW, Kamra A, Bertino E, and Li N et al [6]. The main goal is to classify data records that are naturally identical into the same equivalence class. As a result, the authors created the k-member clustering problem, a one-of-a-kind clustering problem. And show that this is an NP-hard problem and suggest a greedy heuristic with an O-complexity (n2). They also created a metric for estimating the loss of knowledge caused by generalizations that works for both quantitative and categorical data as part of our method.

LeFevre et al [7] turn the k-anonymity problem into a partitioning problem. The following are the two moves that make up their strategy. The first step is to partition the d- dimensional space into at least k partitions, where d denotes the number of attributes in the quasi-identifier. The records in and partition are then combined into a single quasi- identifier value. While they have been shown to be effective, these methods have the downside of having a complete order for each attribute domain. In certain cases, involving categorical data with no meaningful order, this makes it impractical.

This paper defines the "Degree Priority" method of visiting Lattice nodes on the generalization tree to boost the K- Anonymity algorithm's performance. A privacy protection analysis of the K-Anonymity algorithm is being worked on by Zhao FeiFei, Dong LiFeng, Wang Kun, and Li Yang et al [8]. In addition, this paper proposed a "Two Times K- anonymity" strategy to reduce information loss in the K-anonymity process. Finally, the authors used experimental findings to demonstrate the effectiveness of these approaches.

Pierangela Samarati and Latanya Sweeney [9] discuss the issue of releasing data related to person called person- specific while preserving the privacy of the individuals to which the data relate. Their strategy is based on the k- anonymity concept. If attempts to connect specifically identifying information to its contents ambiguously map the information to at least k individuals, the table provides k- anonymity. They demonstrate how generalization and suppression methods can be utilized to accomplish k- anonymity. The principle of minimal generalization is also implemented, which defines the property of a release process of not distorting the data more than is needed to accomplish k-anonymity. There are also suggested preference policies for choosing between various limited generalizations. Finally, the authors provided an algorithm as well as experimental results obtained when the algorithm was used to generate real-world medical data releases.

Sweeney et al [10] suggested the use of k-anonymization, which is one of the well-studied techniques of privacy- preserving data mining. When projected on a subset of public attributes, this approach recommends that the values of the public attributes be generalized in such a way that each published record becomes non distinguishable from at least k 1 other records. As a result, each person can be connected to a set of records in a published anonymized table with a size of at least k, providing some protection of privacy.

Meyerson and Williams [11] looked into the issues and conducted research on the assumption that the table entries should be left alone or removed entirely. The cost function to minimize in this setting is the total number of deleted entries in the table. They demonstrated that the k- dimensional perfect matching problem is NP-hard by demonstrating a reduction. They developed two approximation algorithms: one that runs in O (n2k) time and achieves an approximation ratio of O (k ln k); and another that runs in O (n2k) time and guarantees an approximation ratio of O (k ln k) (k ln n).

Li, N. et al. [12] aims to close the gap by examining k-anonymization approaches and demonstrating that they are incapable of providing again recognition attack. After the experiment is completed, study proposes an alternative approach to the production disturbance. Random sampling is used for achieving differential privacy at the start of the process and to fine-tune the effects so that they archive a high layer of privacy. Because of the random sampling, the analysis is vulnerable to a re-identification attack. The paper focuses solely on output disruption (i.e. added the random sampling is after query results). Since random sampling generally utilized to improve the algorithm's privacy, and can be extended to anything. The emphasis of the paper is solely on the disruption of production (i.e. added the random sampling after query results). Random sampling can be extended to the input disturbance as it is used to improve the privacy of the algorithm. The paper is concerned solely with the disruption of output. Random sampling is utilized for interrupting the inputs and is used to enhance the privacy of algorithms. In order to offer a sufficient protection of privacy when publishing micro data, the privacy gap δ must be minimal. Random sampling can be satisfied with the smooth application of author(s), (ϱ, δ). - differential privacy. Various other techniques available. Other methods that are not evaluated. Other methods which are not evaluated in this Article may be used for the purpose of satisfying the needs of the Member States (ϱ, δ).

F. K. Dankar et al. [13] look at some of the practical issues that come with employment of differential privacy to health data. They also suggested a model for using differential privacy to analyze health-care data. The study looks at a re- identification assault on personal information from a patient's health records. Although health-care data must also be disclosed, the authors argue that particular considerations must also be discussed. If this is completed, further methods for enforcing differential privacy would be fruitful.

Friedman, A., and Schuster [14] solve the problem by using the Data Mining-Decision Tree algorithm to simultaneously understand the privacy and nature of the algorithm. The enhanced algorithm achieves a higher degree of privacy and data accuracy on the fewer learning samples taken for testing. The data variance is relatively high after the completion of the experiment. The reviewer suggested other stopping laws, but they were not experimentally tested in this article. While the paper discusses an auxiliary assault, it does not go into detail on how it was avoided.

Table 4. Literature Review

Author	Description	What They Used	Drawback
Swagatika Devi et al	Focuses on retaining the diversity of sensitive attributes while maintaining a minimum group size of k	Introduced the l- diversity strategy	k- anonymity
A. Machanavajjhala et al	Used to protect the location's privacy	K- anonymity property	privacy
F. Song et al	To fix the issue of privacy security	Modern k- anonymous approach	Little information loss
Angiuli et al	High- achieving students have rare quasi- identifying characteristics, making high-grade records more likely to be suppressed.	Suppression - heavy anonymization approaches	
Kamra A et al	For preventing data loss and thus maintaining high data quality and to classify data records that are naturally identical into the same equivalence class	Clustering- based strategy	k-member clustering problem
LeFevre et al	It turn the k- anonymity problem into a partitioning problem	partition the d- dimensional space into at least k partitions	In certain cases involving categorical data with no meaningful order
Zhao FeiFei et al	Used to reduce information loss in the K- anonymity process	Two Times K- anonymity strategy	
Pierangela Samarati et al	generalization and suppression methods used to achieve k- anonymity	generalization and suppression methods	Person specific data preserving
Sweeney et al	Used in privacy- preserving data mining	k- anonymization	Less privacy protection
Meyerson et al	They achieves an approximation ratio	two approximation algorithms	
Li, N. et al	Random sampling is used for achieving differential privacy at the start of the process and to fine- tune the effects so that they archive a high layer of privacy	Examined k-anonymization approaches, used random sampling	incapable of providing again recognition attack
F. K. Dankar et al	Model is used to analyze health care data	Model for differential privacy.	Stated re- identification attack but don't provided solution to avoid it.

Friedman, A	achieves a higher degree of privacy and data accuracy on the fewer learning samples	Data Mining- Decision Tree algorithm	Discussed an auxiliary assault but not studied how to avoid it
Simi Ms et al	K– anonymization used to prevent such attacks	K-anonymization	privacy of a person on social site

According to their study, several organizations have released significant microdata. Simi, Ms., Nayaki, K., and Elayidom, M. [15] work in business and research-oriented fields, participating in data processing and cloud services involving quality data. It includes personal details such as DOB, Pin- code, sex and marital status, which may be combined with other public data for the recognition of an individual, but excludes explicit identification marks such as name and address. This type of involvement attack can be used to access certain personal information from the social media site, putting the privacy of a person at risk. By changing the microdata, K-anonymization is used to prevent such attacks. K-anonymization is used to avoid such attacks by changing microdata. It would be difficult to find an appropriate way to anonymized data with the potential for more data. Based on a series of studies and a systematic comparison, the author proposes three best algorithms along with their performance and efficacy. Researchers may use studies to assess the relationship between k values, anonymization levels, quasi- identifying selection, and time of execution.

Dataset Description

Adult dataset:

This was extracted from a dataset of the US Census Bureau Data Extraction System. It consists of 32,561 records over the 15 attributes: age, work class, final weight, education, education number, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and salary class.

III.CONCLUSION

In this paper we discussed about the privacy preserving data publishing and data anonymization. We also discussed about various anonymization techniques and mainly focused on k-anonymity which comprises of both generalization and suppression. Here several attacks on dataset are also discussed.

IV. REFERENCES

[1]. Swagatika Devi, K-ANONYMITY: The History of an IDEA International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2011.

[2]. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-Diversity: Privacy beyond k-anonymity. In ICDE, 2006.

[3]. F. Song, T. Ma, Y. Tian and M. Al-Rodhaan, "A New Method of Privacy Protection: Random k-Anonymous," in IEEE Access, vol. 7, pp. 75434-75445, 2019.

[4]. Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A.D., Seaton, D. T., Chuang, I. Privacy, anonymity, and big data in the social sciences. Communications of the ACM 57(9): 56-63,2014.

[5]. Angiuli, O., Blitzstein, J., and Waldo, J. How to de-identify your data. Queue 13, 8 Sept. 2015.

[6]. Byun JW., Kamra A., Bertino E., Li N. Efficient k-Anonymization Using Clustering Techniques. In: Kotagiri R., Krishna P.R., Mohania M., Nantajeewarawat E. (eds) Advances in Databases: Concepts, Systems and Applications. DASFAA 2007. Lecture Notes in Computer Science, vol 4443. Springer, Berlin, Heidelberg, 2007.

[7]. K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In International Conference on Data Engineering, 2006.

[8]. Zhao FeiFei, Dong LiFeng, Wang Kun, Li Yang, Study on Privacy Protection Algorithm Based on K-Anonymity, Physics Procedia, Volume 33, ISSN 1875-3892, 2012

[9]. Samarati, Pierangela; Sweeney, Latanya, “Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.” Carnegie Mellon University. Journal contribution, 2018.

[10]. Sweeney L, k-anonymity: a model for protecting privacy. Int J Uncertain Fuzzy Knowledge Based System 10 (5):557–570,2002.

[11]. Meyerson A, Williams R, On the complexity of optimal k-anonymity. In: PODS ’04: proceedings of the twenty-third ACM SIGMOD-SIGACT- SIGART symposium on principles of database systems, pp 223–228,2004.

[12]. Li, N., Qardaji, W. H., & Su, D. Provably private data anonymization: Or, k-anonymity meets differential privacy. 2011.

[13]. Dankar, F. K., & El Emam, K. (2012, March). The application of differential privacy to health data. In Proceedings of the 2012 Joint EDBT/ICDT Workshops (pp. 158-166). ACM.

[14]. Friedman, A., & Schuster, A. (2010, July). Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-502). ACM

[15]. Simi, Ms&Nayaki, K &Elayidom, An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity. IOP Conference Series: Materials Science and Engineering. 225. 012279. 10.1088/1757- 899X/225/1/012279,2017

[16]. Feng Bo, HaoWenning, Chen Gang, Jin Dawei, Zhao Shuining, “An Improved PAM Algorithm for Optimizing Initial Cluster Centre,” IEEE, 2012, 978-1-4673-2008- 5/12.

Cite this article as :

Kalpana Chavhan, Dr. Praveen S. Challagidad, "Anonymization Techniques for Privacy Preservation in Social Networks: A Review", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 4, pp. 14-21, July-August 2021. Available at

doi : https://doi.org/10.32628/IJSRST21842

Journal URL : https://ijsrst.com/IJSRST21842