DATA ANONYMIZATION

Devansh Kumar[1](20BIT0141), Ishaan Chhipa[2](20BIT0179), Aakanksha S. Bagal[3](20BIT0190), Vinay R. Maske[4](20BIT0128), Yashashvi Kala[5](20BIT0180)

BTech, IT

[1,2,3,4,5]**School of Information Technology (SITE), VIT University, Vellore, Tamil Nadu, India**

Abstract

In this digital age where data is valued at the highest, we need to ensure that the data collected by organizations does not contain personal data which may violate the privacy of the users. Data anonymization protects private/confidential information through encryption without affecting the actual structure of the data. But there is a drawback, when you clear data of identifiers, attackers can use de-anonymization methods to retrace the data anonymization process. Since data usually passes through multiple sources some available to the public de-anonymization techniques can cross-reference the sources and reveal personal information. That's one of the reasons why the European Union's General Data Protection Regulation (GDPR) demands the pseudonymization or anonymization of stored information of individuals living in the EU. Anonymized data sets are not classified as personal data, and so are not subject to the rules of GDPR. This permits organizations to use the information for broader purposes while remaining compliant and protecting the rights of the data subjects. We propose a 3-stage model to anonymize the input data.

Introduction

Data anonymization is the process of protecting private or sensitive information by erasing or encrypting identifiers that connect an individual to stored data, method of information sanitization which involves removing or encrypting personally identifiable data in a dataset. The goal is to ensure the privacy of the subject's information. It also maintains the structure of the data, enabling analytics post-anonymization. For example, you can run Personally Identifiable Information (PII) such as names, social security numbers, and addresses through a data anonymization process that retains the data but keeps the source anonymous.

Name—no matter what context this arises, the name is the most significant key identifier in a data set. A data set reduces a data source's list of variables. If this information is obtained by the cybercriminal, they can readily trace the source of a data set—even encoded data sets. Thus, names must be anonymized

Credit card details—this field deals with credit card numbers, other details like expiration date and CVV, and credit card tokens. They are regarded as highly personal, are unique to the individual, and can have financial implications for the individual if compromised. They must always be protected.

Mobile numbers—if a cybercriminal gains access to a mobile number they could also gain access to additional, more sensitive data about the individual. Thus, personal phone numbers should always be anonymized.

Photograph—photographs are the perfect means of identification. Often, photographs are collected to verify identity and to ensure security. A dataset containing photos of individuals must be safeguarded, and thus it is a strong candidate for anonymization.

Passwords—a cybercriminal could easily impersonate someone and gain access to private data by compromising their password. In any backend structure created to store passwords, you should encrypt and/or anonymize the data.

Security questions—such data sets are also key identifiers. Many software services and web applications use these questions as a step towards granting user access. Given this, it is important to encrypt them.

Literature Review

Domingo-Ferrer[1] et al. They have tackled this issue by means of a decentralized P2P network that avoids relying on a central data controller. This method offers formal privacy guarantees according to the k-anonymity privacy model. The resulting sets of anonymized trajectories effectively prevent unequivocal reidentification while keeping information loss reasonably low. The decentralized anonymization protocol has been designed so that anonymity and attribute non-disclosure are also ensured w.r.t. the other peers in the network. Adherence to the protocol rules by peers is incentivized by a punishment mechanism based on black lists.

Muhammad Ali[2] et al. ,a privacy-preserving data anonymization scheme for obfuscating the sensitive information in distribution networks is presented. The scheme accommodates the statistical distribution with the parameters estimated from the data provided. It involves two algorithms, a MLE for estimating the parameters from the data and a data anonymization procedure for generating anonymized datasets that are sufficiently realistic.

Tian[3] et al. ,They proposed a two-stage privacy-preserving method of graph neural networks in the social network domain. During the first stage, they designed a novel ∈-k anonymization method that can achieve ∈-local differential privacy (∈-LDP) and k-degree anonymity by incorporating the classical LDP and k-degree anonymization (k-DA) while retaining as much network community information as possible. At the second stage, they developed an adversarial training mechanism for GNNs to resist the disturbance from ∈-k anonymization and retain as much task performance as possible on anonymous social network data.

Girka[4] et al. ,This paper takes the problem of supervised machine learning with deep feedforward neural nets and provide an anonymization algorithm (based on the homeomorphic data space transformation), which guarantees privacy of the data and allows neural networks to learn successfully.

Wong[5] et al. ,studied a collaborative anonymization protocol that aims to increase the confidence of respondents during data collection. Unlike in existing works, our protocol does not reveal the complete set of quasi-identifier (QID) to the data collector (e.g., agency) before and after the data anonymization process. Because QID can be both sensitive values and identifying values, we allow the respondents to hide sensitive-QID attributes from other parties. Our protocol ensures that the desired protection level (i.e., k-anonymity) can be verified before the respondents submit their records to the agency. Furthermore, we allow honest respondents to indict a malicious agency if it modifies the intermediate results or not following the protocol faithfully.

Guo, Naixuan[6] et al. ,propose a new concept named natural equivalent class. It refers to the record set with the same quasi-identifier values naturally existing in the raw dataset. We theoretically prove that the natural equivalent class can effectively reduce the computational complexity of clustering algorithms as well as information loss. Then, we propose a novel clustering-based anonymization algorithm, which tries to cluster records without separating any natural equivalent class. Extensive experiments on real world datasets show that our approach outperforms the previous clustering-based anonymization algorithms in terms of efficiency and data utility.

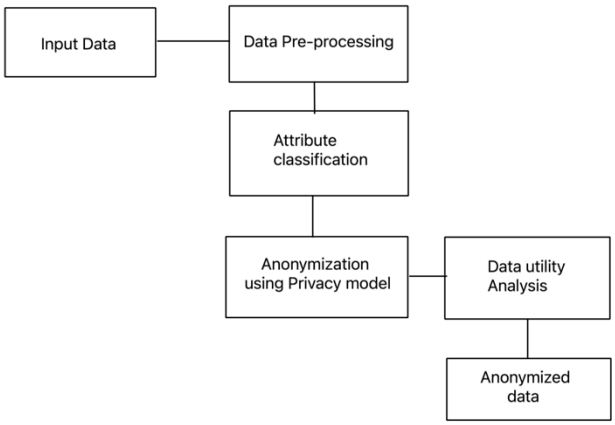
Deng, Xiaofeng[7], propose a framework that uses MapReduce to anonymize large-scale data before disseminating them to human workers. In order to guarantee the number and distribution of data records to be similar in all nodes, our framework first redistributes the original data to all participating nodes. Then a heuristic two-phase anonymization schema, which can be seamlessly integrated into the framework, is proposed. Experimental results show that with the same objective of privacy, our approach is scalable for large-scale data and can improve the average accuracy of human worker's analytic tasks.

Hassan, Fadi[8] et al. , propose a more general solution to text anonymization based on the notion of word embedding. The idea is to represent all the entities appearing in the document as word vectors that capture their semantic relationships. Then a particular entity can automatically be protected by removing the other entities co-occurring in the document whose vectors are similar to the particular entity's vector. Furthermore, these method does not require manually tagged training data and is language-agnostic. We empirically evaluated our proposal on a collection of biographies. Our results show a significant improvement of the detection recall in comparison with classical approaches to text anonymization based on named entity recognition

Proposed Methodology

The anonymization stage, the first stage among the three stages, starts with an attribute classification process. Once the attribute are categorized according to its sensitivity, the user is prompted to select a re- quired privacy model for anonymization. After the anonymization process, the utility of anonymized data needs to be evaluated to ensure that the data is still useful for the intended purpose. The last stage starts with the analysis of the data re-identification risk. We need to try various attacks an adversary can do to the resulting anonymized data to de-anonymize it. Once

those risks, if any, are handled properly, the data can be exported to a data format the user needs.



Stage one: Data Pre-processing

1. Outlier removal: Removal of outliers from the input dataset to make sure that the anonymization process can be performed effectively.
2. Datatype selection: Each attribute is given a certain datatype.

Datatype selection is also included in the pre-processing step, wherein each attribute is assigned a specific datatype. These can be strings, numbers, floating point numbers etc.

Stage two: Attribute Classification

1. Sensitive: Sensitive attributes are values which the users don't want to be publicly disclosed. Examples include salary details, disease information, political/religious preferences etc. These value are retained because the utility of the dataset may depend upon studying these attributes.
2. Insensitive: All attributes other than direct identifiers, quasi and sensitive at- tributes. Examples include height, weight, eye colour etc. These values are not collected in most cases, and even if collected, they can be retained or removed because they may not influence the utility of the dataset or privacy of the users.
3. Identifiers (PIIs): Direct identifiers can uniquely and directly identify an individual/user. Examples include name, Social Security Number, email, phone numbers etc. The initial values are removed.
4. Quasi-identifiers: Such attributes an be linked with auxiliary information to reveal someone's identity. Examples include age, gender, race, zip code etc. Such values are generalized.

Stage three: Privacy models

1. k-anonymity
2. l-diversity
3.t-closeness

Implementation and Results

After making sure that the dataset is parsed properly, we collect the attribute classification information, which tells the model which attributes are identifier, quasi-identifier, insensitive and sensitive.

```
Attribute : 'Email'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 1

Attribute : 'Age'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 2

Attribute : 'Education'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 4

Attribute : 'Marital-status'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 4

Attribute : 'Gender'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 3

Attribute : 'Income'

        1. Identifier
        2. Quasi-identifier
        3. Sensitive
        4. Insensitive
Q> Please select the attribute type: 3
```

On selecting, the model anonymizes the data accordingly

	A	B	C	D	E	F
1	Email	Age	Education	Marital-status	Gender	Income
2	DavidLewis@gmail.com	25	11th	Never-married	Male	<=50k
3	FranciscoMarkland@gmail.com	38	HS-grad	Married-civ-spouse	Male	<=50k
4	DustinBushey@gmail.com	28	Assoc-acdm	Married-civ-spouse	Male	>50k
5	DonaldCraft@gmail.com	44	Some-college	Married-civ-spouse	Male	>50k
6	BradleyFarley@gmail.com	34	10th	Never-married	Male	<=50k
7	MatthewLuke@gmail.com	63	Prof-school	Married-civ-spouse	Male	>50k
8	SherrySherwood@gmail.com	24	Some-college	Never-married	Female	<=50k
9	PaulMerlo@gmail.com	55	7th-8th	Married-civ-spouse	Male	<=50k
10	ChristopherLoudermilk@gmail.com	65	HS-grad	Married-civ-spouse	Male	>50k
11	JamesHudec@gmail.com	36	Bachelors	Married-civ-spouse	Male	<=50k
12	ClaireHernandez@gmail.com	26	HS-grad	Never-married	Female	<=50k
13	JohnDenn@gmail.com	48	HS-grad	Married-civ-spouse	Male	>50k
14	AnthonySmith@gmail.com	43	Masters	Married-civ-spouse	Male	>50k
15	PaulAguilar@gmail.com	20	Some-college	Never-married	Male	<=50k
16	MarshaHarper@gmail.com	43	HS-grad	Married-civ-spouse	Female	<=50k

Figure: sample dataset

	A	B	C	D	E	F
1	Email	Age	Education	Marital-status	Gender	Income
2	*	23-24	Some-college	Never-married	Female	<=50k
3	*	23-24	HS-grad	Separated	Male	<=50k
4	*	23-24	Bachelors	Never-married	Male	<=50k
5	*	23-24	Some-college	Married-civ-spouse	Male	<=50k
6	*	23-24	Bachelors	Never-married	Male	<=50k
7	*	23-24	Some-college	Separated	Male	<=50k
8	*	23-24	Some-college	Never-married	Male	<=50k
9	*	23-24	11th	Never-married	Female	<=50k
10	*	23-24	HS-grad	Never-married	Female	<=50k
11	*	23-24	10th	Married-civ-spouse	Male	<=50k
12	*	23-24	Assoc-voc	Married-civ-spouse	Male	<=50k
13	*	23-24	HS-grad	Never-married	Male	<=50k
14	*	23-24	Bachelors	Never-married	Female	<=50k
15	*	23-24	Bachelors	Never-married	Female	<=50k
16	*	23-24	Bachelors	Married-civ-spouse	Male	>50k

Figure: anonymized data

The results are compared according to utility measures – GenILoss (Generalized information Loss), Cavg score, DM score and Privacy models – k-anonymity, l-diversity, and t-closeness

```
--------- Utility Metrices ---------

DM score (lower is better):
  BEFORE: 9709430 || AFTER: 6898560 || True
CAVG score (near 1 is better):
  BEFORE: 4.584 || AFTER: 7.121 || False
GenILoss: [0: No transformation, 1: Full supression]
 Value: 0.006179846822754476
```

```
- The nominal value for k is : 222.5
- The l value should be within the range : [2, 1]
- The t value should be within the range : [0.67, 0.0)
```

Comparing our proposed model to previous works in this field, we conclude that our model is on par to the existing works and tools used for anonymization.

Conclusion

Various researches that has been proposed to anonymize data by upholding utility while preserving user's privacy from malevolent adversaries, namely privacy preserving data publishing (PPDP) techniques, are carefully considered and studied in this work. In recent years, there is an increase in focus on the rapid development of more practical anonymization solutions due to the significant rise in privacy breaches across the globe, and this area is attracting researchers' interests. Although increasing amount of data offers unprecedented opportunities for analytics, but it also introduces challenges in the area of user privacy.
The proposed tool also gives predictions on probable values of k and probable range of values for l and t. The statistical measures of the input dataset is anonymized using differential privacy by carefully adding noise into the values. Anonymized data utility is measured using GenILoss, Discernibility metric and Average Equivalent Class Size metric so that the data analyst can make carefully thread the balance of data utility and privacy.

References

[1] Josep Domingo-Ferrer, Sergio Martínez, David Sánchez. Decentralized *k*-anonymization of trajectories via privacy-preserving tit-for-tat. Elsevier 2022. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0140366422001153

[2] Muhammad Ali, Krishneel Prakash, Carlos Macana, Md Rabiul, Akhtar Hussain, Hemanshu Pota. Anonymization of distribution feeder data using statistical distribution and parameter estimation approach. Elsevier 2022. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S2213138822002041

[3] Tian, Hu, et al. "∈-k anonymization and adversarial training of graph neural networks for privacy preservation in social networks." *Electronic Commerce Research and Applications* 50 (2021): 101105. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1567422321000776

[4] Girka, Anastasiia, et al. "Anonymization as homeomorphic data space transformation for privacy-preserving deep learning." *Procedia Computer Science* 180 (2021): 867-876. https://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1877050921003914

[5] Wong, Kok-Seng, et al. "Privacy-preserving collaborative data anonymization with sensitive quasi-identifiers." *2019 12th CMI Conference on Cybersecurity and Privacy (CMI)*. IEEE, 2019. https://ieeexplore.ieee.org.egateway.vit.ac.in/document/8962140?arnumber=8962140

[6] Guo, Naixuan, et al. "Data anonymization based on natural equivalent class." *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2019. https://ieeexplore.ieee.org.egateway.vit.ac.in/document/8791905?arnumber=8791905

[7] Deng, Xiaofeng, Fan Zhang, and Hai Jin. "Data Anonymization for Big Crowdsourcing Data." *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2019. https://ieeexplore.ieee.org.egateway.vit.ac.in/stamp/stamp.jsp?tp=&arnumber=9093748

[8] Hassan, Fadi, et al. "Automatic anonymization of textual documents: detecting sensitive information via word embeddings." *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2019. https://ieeexplore.ieee.org.egateway.vit.ac.in/document/8887419?arnumber=8887419