

# MEDI-CHAL PRESENTATION

A Study of Privacy Metrics for Synthetic Data  
- Saloni Dash





# OVERVIEW

- WHAT IS PRIVACY?
- WHAT ARE SOME PRIVACY PRESERVING TECHNIQUES?
  - K ANONYMITY
  - $\epsilon$ -DIFFERENTIAL PRIVACY
  - I-DIVERSITY
  - t-CLOSENESS
- METRICS FOR QUANTIFYING PRIVACY
  - MINIMUM DISTANCE ACCUMULATION
  - CONFIDENCE LEVEL
  - AVERAGE CONDITIONAL ENTROPY
  - VARIANCE
  - HIDDEN FAILURE
  - K ANONYMITY, I-DIVERSITY, t-CLOSENESS,  $\epsilon$ -DIFFERENTIAL PRIVACY



# WHAT IS PRIVACY?



## DATA COLLECTION PRIVACY



- The assumption is that the entity collecting the data is not to be trusted.
- Raw data is randomised by adding noise.

## DATA PUBLISHING PRIVACY

- A trusted data custodian that has a database consisting of rows of data (records).
- The custodian needs to publish an anonymized version of the data, a version that preserves utility as well as privacy
- Privacy Models :
  - Generalization - replacement of a value for a more general one (parent)
  - Suppression - removal of some attribute values to prevent information disclosure
  - Anatomization - de-associates QIDs and sensitive attributes in two separate tables
  - Perturbation - replacement of the original data for synthetic values with identical statistical information

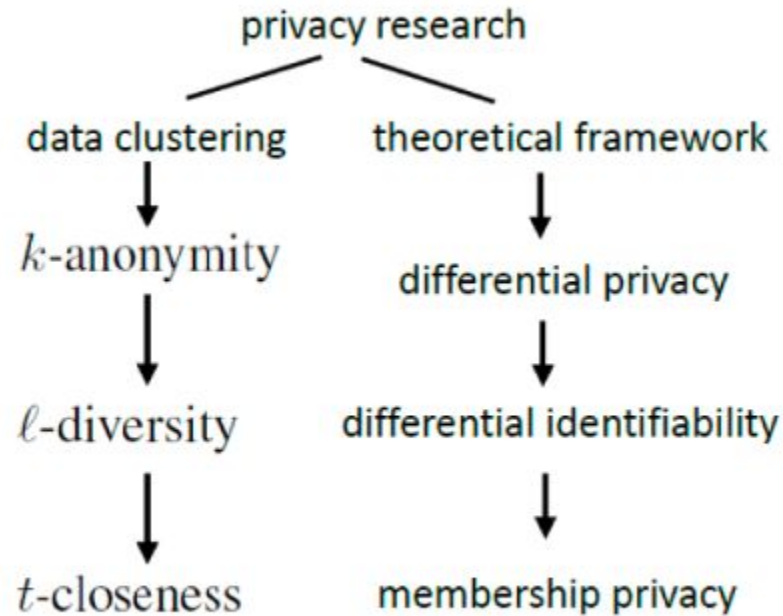


- Goals of Data Anonymization:
  - UNIQUE IDENTITY DISCLOSURE: If data is published then there should not be any record that can identify an individual.
  - SENSITIVE ATTRIBUTE DISCLOSURE: Attackers won't be able to learn about sensitive attribute of an individual via disclosed attributes.
- Classification of Attributes :
  - IDENTIFIERS: These are the attributes which can directly identify an individual. So these attributes are removed before publishing the data. For e.g. name, social security number, etc.
  - QUASI-IDENTIFIERS: These are the attributes which cannot identify an individual directly but if they are linked with publicly available data then they can identify an individual easily. For e.g. zip code, age, sex, etc. An Equivalence Class is a set of records that have same value for all the quasi-identifier attributes.
  - SENSITIVE ATTRIBUTES: These are the attributes which an individual wants to hide from others. For e.g. disease, wages. ( Depends on requirement)
  - NON-SENSITIVE ATTRIBUTES: Any attributes other than the above three mentioned are known as Non-Sensitive Attributes.



# **PRIVACY PRESERVING TECHNIQUES**






**FIGURE 2.** The categories of privacy study.



# K-ANONYMITY

- If the identifiable attributes of any database record are indistinguishable from at least other  $k-1$  records, then the dataset is said to be  $k$ -anonymous.
- An attacker could not identify the identity of single record since other  $k-1$  similar records exist.
  - Eg:- If you try to identify a man from a release, but the only information you have is his birth date and gender. There are  $k$  people meet the requirement. This is  $k$ -Anonymity.
- The set of  $k$  records is known as equivalence class.
- Note that “identifiable attributes” in the aforementioned definition refers to QIDs (Quasi Identifiers).
- In the  $k$ -anonymity model, the value  $k$  may be used as a measure of privacy: the higher the value of  $k$ , the harder it is to de-anonymize records. In theory, in an equivalence class, the probability of de-anonymizing a record is  $1/k$ .





	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of  $k$ -anonymity, where  $k=2$  and  $QI=\{Race, Birth, Gender, ZIP\}$



# ALGORITHMS FOR IMPLEMENTING K ANONYMITY

- **MONDRIAN ALGORITHM:**

- Mondrian is a multidimensional k-anonymity algorithm. This algorithm is very fast, scalable and produces better results.
- This algorithm uses strict partitioning and relaxed partitioning methods which further results in better data utility.
- A partition of the multidimensional space corresponds to unique anonymization result.
- If partitions are not intersecting or overlapping with one another then it is known as Strict Partitioning. And if the partitions are overlapping with one another then it is known as Relaxed partitioning.

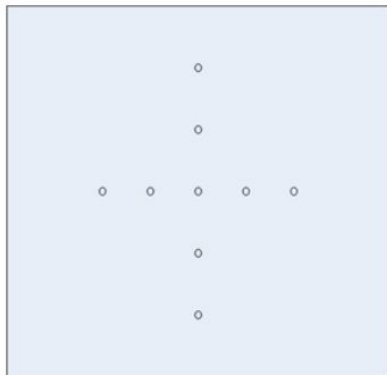


FIGURE 4: DATA IN 2-DIMENSION PLANE

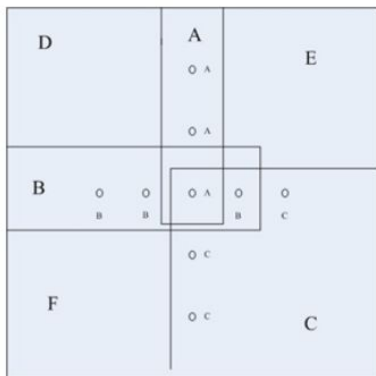


FIGURE 5: RELAXED PARTITION

Relaxed Partition is much better than Strict Partition.

Suppose we have to strictly partition table in figure 4 into at least two regions then at least there would be one region which will contain no more than two tuples.

Therefore, if 3-anonymity or more is required then there would be no strict partition based algorithm which can serve our purpose.



- OTHER ALGORITHMS:
  - DATAFLY ALGORITHM
  - $\mu$ -ARGUS
  - OPTIMAL K-ANONYMITY
  - INCOGNITO
  - BOTTOM-UP GENERALIZATION
  - TOP-DOWN SPECIALIZATION

## DISADVANTAGES OF K-ANONYMOUS MODEL :

TABLE 4: 2-ANONYMOUS TABLE

AGE	SEX	ZIP CODE	DISEASE
[20-40]	M	18***	HIV
[20-40]	M	18***	HIV
[41-50]	F	120**	CANCER
[41-50]	F	120**	HEART DISEASE

TABLE 5: EXTERNAL DATA AVAILABLE TO AN ATTACKER

NAME	AGE	SEX	ZIP CODE
ANDREW	31	M	18601
CLARKE	27	M	18555
ROSY	49	F	12001
ANA	42	F	12456

- Homogeneity Attack - if the sensitive attribute lacks diversity in values and attacker is only interested in knowing the value of sensitive attribute then the aim of is achieved.
- Background Attack - This model assumes that attacker has no additional background knowledge. Suppose, if attacker knows that Ana has low chance of Cancer, then after combining table 4 and 5, attacker can conclude that Ana is suffering from a heart disease.



# I-DIVERSITY

- It expands the k-anonymity model by requiring every equivalence class to abide by the I-diversity principle.
- An I-diverse equivalence class is a set of entries such that at least I “well-represented” values exist for the sensitive attributes. A table is I-diverse if all existing equivalence classes are I-diverse.
- If there are at least I distinct values in an equivalence class then it is known as distinct I-diversity.
- A stronger notion of I-diversity is the definition of entropy I-diverse, defined as follows. An equivalence class is entropy I-diverse if the entropy of its sensitive attribute value distribution is at least  $\log(I)$ .
- It suffers from skewness and similarity attacks.

TABLE 8: EXAMPLE TABLE FOR L-DIVERSITY

AGE	ZIP CODE	SALARY	DISEASE
24	12889	2K	GASTRIC ULCER
26	12110	3K	GASTRITIS
28	12005	4K	STOMACH CANCER
31	15601	6K	FLU
33	15666	7K	BRONCHITIS
37	15689	9K	CANCER
43	19123	11K	HEART DISEASE
45	19765	12K	GASTRITIS
49	19303	14K	PNEUMONIA

TABLE 10: TABLE TO EXPLAIN SKEWNESS ATTACK

AGE	ZIP CODE	SALARY	DISEASE
[11-20]	12***	2K	NEGATIVE
[11-20]	12***	3K	NEGATIVE
[21-30]	156**	4K	NEGATIVE
[21-30]	156**	6K	NEGATIVE
[31-40]	19***	7K	NEGATIVE
[31-40]	19***	9K	POSITIVE
[41-50]	170**	11K	NEGATIVE
...	...	...	...
[81-90]	170**	12K	NEGATIVE



## t-CLOSENESS

- This model requires the distribution of the sensitive values in each equivalence class to be “close” to the corresponding distribution in the original table, where close is upper bounded by the threshold  $t$ .
- The distance between the distribution of a sensitive attribute in the original table and the distribution of the same attribute in any equivalence class is less or equal to  $t$  (t-closeness principle).
- The three most common functions are the variational distance, the Kullback-Leibler (KL) distance and the Earth Mover's distance (EMD).



# ε-DIFFERENTIAL PRIVACY

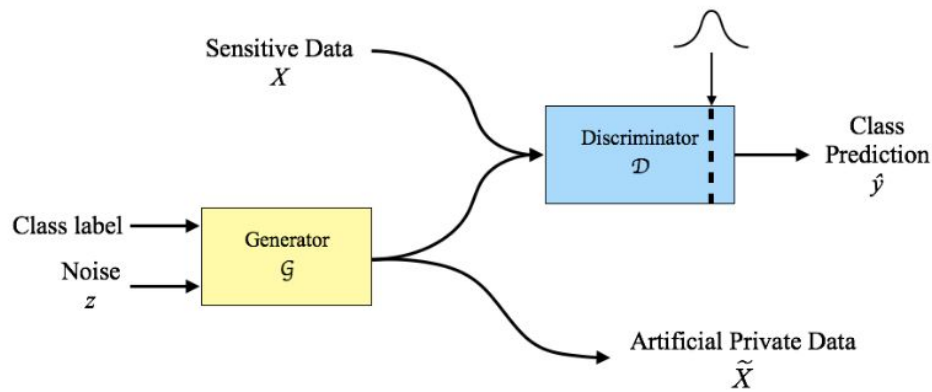
- The notion of differential privacy is to measure the difference on individual privacy disclosure between the presence and the absence of the individual's record.
- The ε-differential privacy model ensures that a single record does not considerably affect the outcome of the analysis over the dataset. In this sense, a person's privacy will not be affected by participating in the data collection since it will not make much difference in the final outcome.
- Let  $K(\cdot)$  be a randomized function, and  $D_1$  and  $D_2$  two databases differing at most on one record, then:

$$Pr[K(D_1) \in S] \leq e^\epsilon \times Pr[K(D_2) \in S].$$

where  $S \in \text{Range}(K)$



- Interactive Methods
- Non-Interactive Methods
  - Laplacian Noise
  - Introduce a Gaussian noise layer in the discriminator network of GAN
    - make the output and the gradients differentially private with respect to the training data, and then use the generator component to synthesise privacy-preserving artificial dataset.



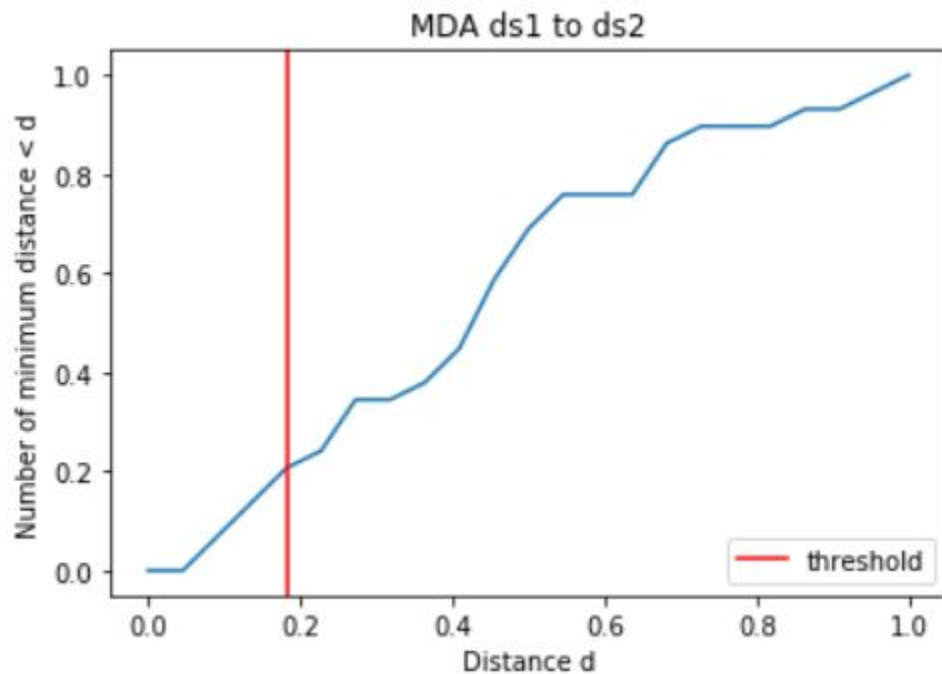


# PRIVACY METRICS





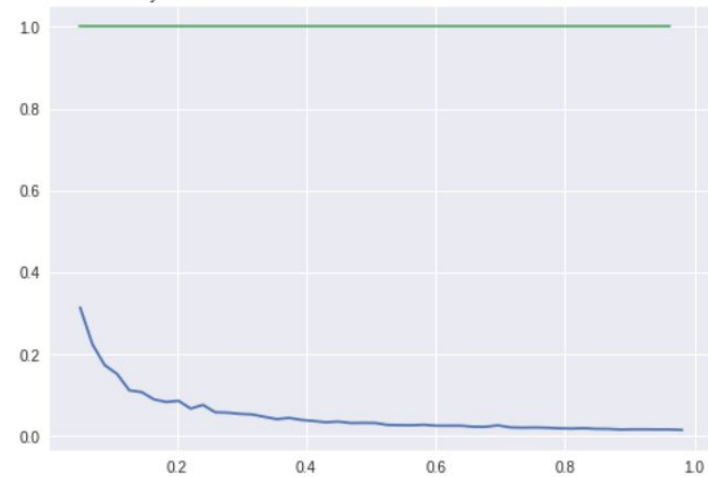
# MINIMUM DISTANCE ACCUMULATION



Privacy and Resemblance Variance with Threshold for manhattan metric



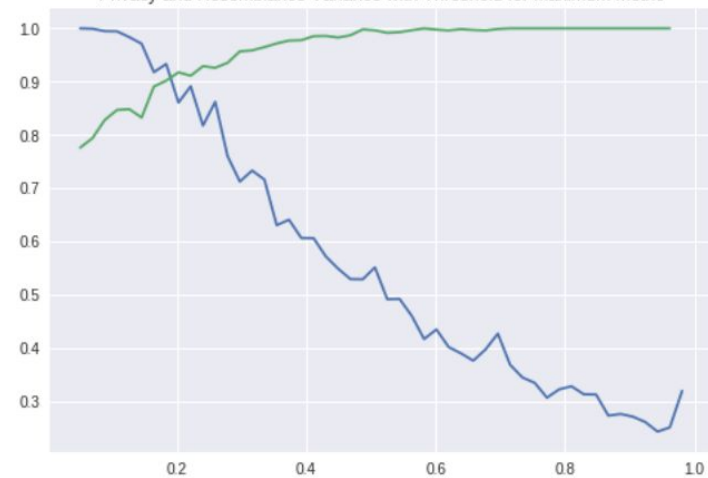
Privacy and Resemblance Variance with Threshold for minimum metric



Privacy and Resemblance Variance with Threshold for euclidean metric



Privacy and Resemblance Variance with Threshold for maximum metric







## CONFIDENCE LEVEL :

- It measures how well the original values may be estimated from the randomized data. If an original value may be estimated to lie in an interval  $[x_1, x_2]$  with  $c\%$  confidence, then the interval  $(x_2 - x_1)$  is the amount of privacy at  $c\%$  confidence.
- The problem with this metric is that it does not take into account the distribution of the original data, therefore making it possible to localize the original distribution in a smaller interval than  $[x_1, x_2]$ , with the same  $c\%$  confidence.

## AVERAGE CONDITIONAL ENTROPY :

- It is based on the concept of information entropy. Given two random variables  $X, Z$  the average conditional privacy of  $X$  given  $Z$  is  $H(X|Z) = 2^{h(X|Z)}$  where  $h(X|Z)$  is given by

$$h(X|Z) = - \int f_{X,Z}(x, z) \log_2 f_{X|Z=z}(x) dx dz$$



## VARIANCE :

- In multiplicative noise randomization, privacy may be measured using the variance between the original and the perturbed data. Let  $x$  be a single original attribute value, and  $z$  the respective distorted value,  $\text{Var}(\mathbf{x}-\mathbf{z})/\text{Var}(\mathbf{x})$  expresses how closely one can estimate the original values, using the perturbed data .

## HIDDEN FAILURE :

- It is used to measure the balance between privacy and knowledge discovery. The hidden failure may be defined as the ratio between the sensitive patterns that were hidden with the privacy-preserving method, and the sensitive patterns found in the original data.

$$HF = \frac{\#R_P(D')}{\#R_P(D)}$$

- If  $HF = 0$ , all sensitive patterns are successfully hidden, however, it is possible that more non-sensitive information will be lost in the way.



- RELATIVE MEASUREMENT

- Kullback-Leibler Distance

$$D(p, q) = \sum_{x \in \mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)},$$

- Correntropy

$$\hat{V}_{m,\sigma}(A, B) = \frac{1}{m} \sum_{j=1}^m k_{\sigma}(A_j - B_j),$$

- Kolmogorov-Smirnov Distance

$$d = \sup_{x \in \mathcal{X}} |F(x) - \hat{F}_Z(x)|$$

- INFORMATION THEORETIC MEASUREMENT

$$L = \max (H(V|S) - H(V|S, E)) ,$$





# CONCLUDING REMARKS

- Differential Privacy is a widely accepted standard in the research community.
  - It can be easily implemented for GANs (by adding noise in the discriminator)
  - For other generators you can add Laplacian noise
- K-anonymity could be used as a baseline/one of the benchmarks
  - Can be implemented using Mondrian Algorithm
- To measure privacy, relative distance measurements / entropy can be used. X



# REFERENCES

1. Aggarwal, C. C. (2015). Data mining: the textbook (1st ed.). New York, NY, USA: Data Mining: The Textbook.
2. Agrawal, D. & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. Proc. 20th ACM SIGMOD- SIGACT-SIGART Symp. Principles Database Sys, 247–255.
3. Agrawal, R. & Srikant, R. (2000). Privacy-preserving data mining. ACM SIGMOD Rec, 29 (2), 439–450.
4. Dankar, F. K. & Emam, K. E. (2013). Practicing differential privacy in health care: a review. TRANSACTIONS ON DATA PRIVACY.
5. Dwork, C. (2006). Differential privacy. Automata, Languages and Programming, 4052,1–12.
6. E. Bertino, D. L. & Jiang, W. (2008). A survey of quantification of privacy preserving data mining algorithms. Privacy-Preserving Data Mining, 183–205.
7. MENDES, R. & VILELA, J. P. (2017). Privacy-preserving data mining: methods, metrics, and applications. IEEE Access.



8. N. Li, T. L. & Venkatasubramanian, S. (2007). T-closeness: privacy beyond k-anonymity and l-diversity. Proc. IEEE 23rd Int. Conf. Data Eng. (ICDE), 106–115.
9. Oliveira, S. R. M. & Zaiane, O. R. (2010). Privacy preserving clustering by data transformation. J. Inf. Data Manag, 1 (1), 37.
10. Samarati, P. & Sweeney, L. (1998a). Generalizing data to provide anonymity when disclosing information. Proc. PODS, 188.
11. Samarati, P. & Sweeney, L. (1998b). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. IEEE Symp. Res. Secur. Privacy, 384–393.
12. T. S. Gal, Z. C. & Gangopadhyay, A. (2008). A privacy protection model for patient data with multiple sensitive attributes. Int. J. Inf. Secur. Privacy, 2 (3), 28.
13. Yu, S. (2016). Big privacy: challenges and opportunities of privacy study in the age of big data. IEEE Access, 4, 2751–2763.

**THANK YOU!**

