# Linkage Attack Analysis K-anonymity, L-Diversity and T- closeness

**Course Name** - Data Privacy & Security

**Student Name :** Rithvij Pasupuleti

**Professor –** Dr . Han Wang

## Introduction:

Various institutions, companies and organizations often need to publish sensitive data for research and other purposes. While publishing the sensitive data online, it is very crucial to secure the information of the individuals from being identified uniquely in a database. The released data might provide many valuable insights for various users of it but there is always potential identity disclosure in these scenarios. K-anonymity and L-diversity are implemented in this context to limit the disclosure and potential linkage of data given the adversaries background knowledge. T-closeness has been proposed later to further improve the level of anonymity by considering the distribution of the sensitive attribute in an equivalence class. Linkage attacks are potential threats given the prior knowledge of an adversary to link attribute values in an anonymized dataset to extract sensitive information of individuals. Hence the linkage attack scenarios have been simulated and analyzed for various values of 'k', 'l' and 't' and further studied to gain valuable insights on the level of data anonymization to achieve desirable level of privacy and security

## Methodology and Algorithm Behind Implementation:

### Mondrian's Algorithm:

For K-anonymity model:

- It supports multidimensional models, which is ideal for our particular dataset.
- We tested with five different k values: 3, 10, 20, 50, and 100.
- Using the hierarchical clustering approach, the Mondrian divides the dataset into smaller groups.

1. Using the kd-tree, divide the input dataset into k-groups. The term "k-groups" denotes that each group has at least k records.

   Partition(region, k)

   a. Choose the optimal dimension that produces a k-anonymous partition.

   b. If possible, divide the region into R1 and R2 based on that dimension.

c. Return Partition(R1, k) U Partition(R2, k)

d. If this is not feasible, return

2. Generalization of each k-group so that each group has the same quasi-identifier.

## For L-Diversity Model:

Due to the constraints of the k-anonymity model, the l-diversity model was proposed.

The l-diversity model is an extension of the k-anonymity model in that it applies the l-diversity principle to each equivalence class.

The l-Diversity Principle states that According to the l-diversity principle, "at least l 'well represented values' exist for the sensitive attributes in each equivalence class." A dataset is said to be l-diverse if all of its equivalence classes exhibit the l-diversity property.

Mondrian's partitions: The partitions are separated to ensure that a dataset has at least l different values.

## For T-closeness model:

- t-closeness is an improvement of l-diversity group-based anonymization that is used to protect privacy in data sets by reducing data representation granularity.
- This decrease is a trade-off that results in some loss of efficacy of data management or mining algorithms in order to gain some privacy.
- The t-closeness model enhances the l-diversity model by treating attribute values differently by accounting for the distribution of data values for that attribute.
- According to t-closeness, each k-anonymous group's sensitive attribute values must have a statistical distribution that is "close" to the dataset's total distribution of that attribute.

Mondrian’s partition : The partition is divided in such a way that the difference in distances between the probability distribution of sensitive attributes in the whole data set when compared to anonymity data set is at least close to t

## For Linkage Attack :

- The goal is to link potential matches between anonymized data and attackers' known data.
- The quasi identifiers chosen for analysis are “age”, “education”, “marital-status” and “race” and sensitive attribute chosen is “occupation”
- In order to achieve this potential one-hot encoding has been applied to both the anonymized data and attacker’s known data.
- A simple K-dimensional tree has been constructed using anonymized data and the built

KD-tree has been queried with the attacker's knowledge dataset to find potential distances and indices of K-Nearest Neighbours.

- The above data has been stored as a data frame containing the indices and distances of the potential nearest rows for the attacker's data.
- Rows with "-1" would potentially indicate that no matching record has been found

```
class RecordLinkage:
    def __init__(self, df, knowledge):
        self.df = df
        self.knowledge = knowledge

        categories = (df.dtypes == "object").keys().to_list()
        self.enc = ce.OneHotEncoder(cols=categories, drop_invariant=False)
        df_concat = pd.concat([self.df, self.knowledge], ignore_index=True)
        self.enc.fit(df_concat)

    def execute(self, k=3):
        enc_df = self.enc.transform(self.df).astype("float64").values
        enc_knowledge = self.enc.transform(self.knowledge).astype("float64").values

        tree = KDTree(enc_df)
        dist, index = tree.query(enc_knowledge, k=k)
        return dist, index


def attack(df, knowledge):
    k = 3
    a = RecordLinkage(df, knowledge)
    ab=[]
    dist, index = a.execute(k)

    di = pd.DataFrame(np.hstack((index, dist)))
    #print(di)

    di.loc[di[3] > di[3].median(), :] = -1
    #print(di[3].median())
    # Display the top three
    ab.append(di.iloc[:, 0:k].astype(int))
    return di.iloc[:, 0:k].astype(int),ab
```

| | 0 | 1 | 2 |
|---|---|---|---|
| 0 | -1 | -1 | -1 |
| 1 | 403 | 449 | 402 |
| 2 | 827 | 824 | 825 |
| 3 | -1 | -1 | -1 |
| 4 | 79 | 77 | 78 |
| 5 | -1 | -1 | -1 |
| 6 | 235 | 234 | 233 |
| 7 | 1726 | 1725 | 1724 |
| 8 | -1 | -1 | -1 |
| 9 | 235 | 234 | 233 |
| 10 | 403 | 449 | 402 |

- Obtaining '-1' would be perfect scenario but the nearest neighbor indices that we obtained can still be validated to get more granular information about how effectively the linkage attack has affected the anonymized data
- In order to get even more granular details about the linkage happening from the obtained indices , respective counts are calculated along each feature column while simulating for various 'k' 'l' and 't' values.

```
link_cnt=0
for i in range(11):
  for j in range(3):
    x=ab[0][j][i]
    for k in range(4):
      if((dfad_cnt_k3.loc[x][feature_columns[k]])== knowledge.loc[i][feature_columns[k]]):
        #print(dfad_cnt_k3.loc[x])
        #print(knowledge.loc[i])
        link_cnt+=1;

print(link_cnt)
```

```
       0     1     2
0     -1    -1    -1
1    403   449   402
2    827   824   825
3     -1    -1    -1
4     79    77    78
5     -1    -1    -1
6    235   234   233
7   1726  1725  1724
8     -1    -1    -1
9    235   234   233
10   403   449   402
21
```

## Analysis On Linkage Attack:

### Attack Scenarios:

- We considered two linkage attack scenarios on the anonymized data -
  - One where the attacker has all the attributes known ( ideal scenario)
  - Another where the attacker only has some attributes known

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | age | workclass | fnlwgt | education | education- | marital-sta | occupatio | relationshi | race | sex | capital-gai | capital-los | hours-per- | native-cou | income |
| 2 | 80 | Federal-g | 76845 | 9th | 30 | Married-c | Farming-fi | Husband | Black | Male | 0 | 0 | 40 | United-Sta | <=50K |
| 3 | 25 | Private | 105598 | Bachelors | 40 | Divorced | Tech-supp | Not-in-fa | White | Male | 0 | 0 | 58 | United-Sta | <=50K |
| 4 | 28 | Self-emp- | 191681 | Associate- | 10 | Married-c | Exec-man | Husband | White | Male | 0 | 0 | 50 | United-Sta | >50K |
| 5 | 77 | Private | 309974 | Bachelors | 58 | Separated | Sales | Own-child | Black | Female | 0 | 0 | 40 | United-Sta | <=50K |
| 6 | 45 | Local-gov | 125927 | HS-grad | 20 | Married-c | Farming-fi | Husband | White | Male | 0 | 0 | 40 | United-Sta | <=50K |
| 7 | 84 | Private | 160647 | Bachelors | 13 | Married-c | Adm-cleri | Wife | White | Female | 0 | 0 | 40 | United-Sta | >50K |
| 8 | 68 | Private | 81534 | Some-coll | 40 | Divorced | Sales | Husband | Asian-Pac | Male | 0 | 0 | 40 | United-Sta | >50K |
| 9 | 36 | Private | 336367 | Assoc-acd | 12 | Never-ma | Exec-man | Unmarrie | White | Male | 0 | 0 | 50 | United-Sta | <=50K |
| 10 | 80 | Federal-g | 76845 | 7th-8th | 30 | Married-c | Farming-fi | Husband | Black | Male | 0 | 0 | 40 | United-Sta | <=50K |
| 11 | 68 | Private | 81534 | 10th | 35 | Divorced | other-serv | Husband | Asian-Pac | Male | 0 | 0 | 40 | United-Sta | >50K |
| 12 | 25 | Private | 105598 | 11th | 57 | Divorced | other-serv | Not-in-fa | White | Male | 0 | 0 | 58 | United-Sta | <=50K |

| age | workclass | education | education-num | marital-status | race | sex | native-country |
|---|---|---|---|---|---|---|---|
| | Federal-gov | 9th | 30 | Married-civ-spouse | | Male | |
| 25 | Private | | 40 | | | Male | United-States |
| 28 | | Associate-voc | 10 | Married-civ-spouse | | Male | |
| 77 | Private | Bachelors | 58 | | | Female | United-States |
| 45 | Local-gov | | 20 | Married-civ-spouse | | Male | United-States |
| 84 | | Bachelors | 13 | | | | United-States |
| | Private | Some-college | 40 | Divorced | | Male | |
| | Private | Assoc-acdm | 12 | Never-married | | Male | United-States |
| 80 | Federal-gov | | 30 | | | Male | United-States |
| | Private | HS-grad | 55 | Seperated | | Male | |
| | | 10th | 65 | Divorced | | | |

**Results Obtained for Attack Scenario-1(complete data known to adversary)**

| Simulations done for various K,L and T values | Counts |
|---|---|
| K=3 | 45 |
| K=3;L=2 | 30 |
| K=3;t=0.15 | 21 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=10 | 30 |
| K=10;L=2 | 30 |
| K=10;L=4 | 30 |
| K=10;T=0.15 | 21 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=20 | 30 |
| K=20;L=2 | 30 |
| K=20;L=4 | 30 |
| K=20;L=12 | 21 |
| K=20;T=0.15 | 21 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=50 | 27 |
| K=50;L=2 | 27 |
| K=50;L=4 | 27 |
| K=50;L=12 | 21 |
| K=50;T=0.15 | 21 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=100 | 21 |
| K=100;L=2 | 21 |
| K=100;L=4 | 21 |
| K=100;L=12 | 21 |
| K=100;T=0.15 | 18 |

## Attack Scenario-2 (Only partial data available with adversary)

| Simulations done for various K,L and T values | Counts |
|---|---|
| K=3 | 27 |
| K=3;L=2 | 27 |
| K=3;t=0.15 | 15 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=10 | 18 |
| K=10;L=2 | 18 |
| K=10;L=4 | 18 |
| K=10;T=0.15 | 15 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=20 | 18 |
| K=20;L=2 | 18 |
| K=20;L=4 | 18 |
| K=20;L=12 | 12 |
| K=20;T=0.15 | 15 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=50 | 18 |
| K=50;L=2 | 18 |
| K=50;L=4 | 18 |
| K=50;L=12 | 12 |
| K=50;T=0.15 | 15 |

| Simulation done for various K,L and T values | Counts |
|---|---|
| K=100 | 15 |
| K=100;L=2 | 15 |
| K=100;L=4 | 15 |
| K=100;L=12 | 12 |
| K=100;T=0.15 | 15 |

The above implementations have been done in below google colab notebooks and relevant results have been obtained-

https://colab.research.google.com/drive/1DHXAGoMc9YEMXamURot4Ovom-LfuSobP?usp=sharing
https://colab.research.google.com/drive/1v_D8drSnNPNDZM5h6hdoAqIVNL7889Ev?usp=sharing

## Future Work:

- Taking the attribute values into context, rigorous analysis can be done for various 'k', 'l' and 't' values for a larger number of attack samples on anonymized data.
- Wide range of 't' values can be analyzed for different real-world datasets and proper metric can be proposed to choose best 't' for different datasets with different attribute distributions

- Choosing Quasi-identifiers again is crucial in any anonymization process and better feature selection algorithms can be implemented for choosing best number of quasi-identifiers in context to the sensitive attributes in the dataset

- Machine learning models can be proposed to give well-predicted analysis on any given dataset of any domain for the user to choose desirable ‘k’, ‘l’ and ‘t’ values depending upon the data privacy and security the user wants to achieve for his dataset.

## Conclusion:

- Anonymization has been achieved for various ‘k’, ‘l’ and ‘t’ values on Adult dataset using Mondrian’s algorithm

- Linkage attack has been simulated for various anonymized datasets for two attack scenarios, one where the adversary has all the attribute data available and one where the adversary has partial data available.

- Respective counts have been calculated and analyzed for when correct linkage has been done.

- Larger values of ‘k’, ‘l’ and ‘t’ will make the anonymized data more redundant and saturation can be observed at one point.

- Although increases in ‘k’ value and ‘l’ value result in better privacy, it is not always the case that the anonymized data that we obtain in this process is always secure and resistant to attack. Sometimes even smaller values of ‘k’ also might provide better security against an attack.

- Attack-scenario-2 which is a more possible adversary scenario will have a lesser number of correct linkages as the anonymity level increases when compared to attack-scenario-1.

## Potential Research Done:

Below papers have been referred to determine suitable and relevant ‘t’ values for the dataset. In both the papers, Multiple analysis techniques have been analyzed to obtain and determine suitable ‘t’ value for given quasi-identifiers.

https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf

**$t$-Closeness: Privacy Beyond $k$-Anonymity and $\ell$-Diversity**

Ninghui Li, Tiancheng Li
Department of Computer Science, Purdue University
{ninghui, li83}@cs.purdue.edu

Suresh Venkatasubramanian
AT&T Labs – Research
suresh@research.att.com

**Abstract**

*The $k$-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of* closed. Two types of information disclosure have been identified in the literature [4, 8]: *identity disclosure* and *attribute disclosure*. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute

https://www.researchgate.net/publication/272864425_Determining_t_in_t-closeness_using_Multiple_Sensitive_Attributes

**Determining *t* in *t-closeness* using Multiple Sensitive Attributes**

Debaditya Roy
Department of Computer Science and Engineering
NIT Rourkela

Sanjay Kumar Jena
Department of Computer Science and Engineering
NIT Rourkela

Further, to apply Mondrian;s algorithm, we had to go through below paper at adopt the algorithm for K-anonymity, L-diversity and t-closeness-

K. LeFevre, D. J. DeWitt and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 25-25, doi: 10.1109/ICDE.2006.101.

A. Machanavajjhala, J. Gehrke, D. Kifer and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," 22nd International Conference on Data Engineering (ICDE'06), Atlanta, GA, USA, 2006, pp. 24-24, doi: 10.1109/ICDE.2006.1.

J. -H. Weng and P. -W. Chi, "Multi-Level Privacy Preserving K-Anonymity," 2021 16th Asia Joint Conference on Information Security (AsiaJCIS), Seoul, Korea, Republic of, 2021, pp. 61-67, doi: 10.1109/AsiaJCIS53848.2021.00019.