DATA PRIVACY

USING

K-ANONYMITY, L-DIVERSITY AND T-CLOSENESS

Submitted by:

Nishit Gupta IIT2014502

Sahil Prakash IIT2014504

Sacheendra Mohan Singh IIT2014506

Project Supervisor

Dr. K.P. Singh

MOTIVATION

- Various organizations often need to publish micro data (e.g. medical data, census data, employee data) for research, study and other purposes.
- This field-structured and person specific data is prone to attacks and misuse.
- The identity reveal can be a threat to a person's job, family and most importantly life.
- The data holders need to share a version of the data to the requesting authority.
- So, there is a requirement that the data must be released in a version which has the scientific guarantee that the individuals who are a subject of the data can not be re-identified while the data remain practically useful.

OBJECTIVE

- To implement various data privacy techniques such as
  - **K-anonymity**
  - **L-diversity**
  - **T-closeness.**
- Analyze the **flaws and advantages** of each mentioned technique.
- Finding out whether **Machine Learning** (training the system) methods could be incorporated for identity disclosure.
- Applying various **Machine Learning algorithms** for feature extraction and verifying the same.

LITERATURE SURVEY

S No.	Author	Paper Title	Year	Crux	Venue
1.	Latanya Sweeney	*K-Anonymity: A Model for protecting privacy.*[1]	2002	*This Paper introduces a formal protection model called k-anonymity and a set of accompanying policies for deployment . A released table provides k-anonymity protection if information for each person contained cannot be distinguished from at least k-1 individuals whose information is also in the dataset.*	*International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems October 5,2002*
2.	A.Machanavajjhala J. Gehrke D.Kifer M.Venkitasubramaniam	*l-Diversity: Privacy Beyond K- Anonymity.*[2]	2006	*In this paper it is shown that k-anonymity model does not guarantee privacy against attackers using two simple attacks, first, lack of diversity in sensitive attributes, second, using background knowledge . It proposes a powerful privacy definition called l - diversity which requires that each equivalent class has at least l well represented values for sensitive attribute.*	Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on 3-7 April 2006
3.	Ninghui Li Tiancheng Li Suresh Venkatasubramanian	*T-closeness: Privacy Beyond K-Anonymity and l-Diversity.*[3]	2007	*This Paper shows limitations of l-diversity and then proposes another privacy notion called t-closeness , which requires that the distribution of sensitive attribute in any equivalence class is close to the distribution of the attribute in overall table. EMD is proposed as a distance measure, to evaluate the closeness.*	Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 15-20 April 2007

TARGETS ACHIEVED TILL NOW

- Implemented all the three data privacy techniques – **K-anonymity, L-diversity and T-closeness.**
- Step by Step analyzed the **advantages and flaws** of each technique and realized the need of one technique over the other.
- **Verified whether identity could be revealed** with the anonymized data by
  - Running the same Query on Original Data and Anonymized Data.
  - Doing a Survey of the output obtained from both the queries.

DATA-SET[4]

Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
39	Male	White	Bachelors	Never-married	United-States	State-gov	Adm-clerical	<=50K
50	Male	White	Bachelors	Married-civ-spouse	United-States	Self-emp-not-inc	Exec-managerial	<=50K
38	Male	White	HS-grad	Divorced	United-States	Private	Handlers-cleaners	<=50K
53	Male	Black	11th	Married-civ-spouse	United-States	Private	Handlers-cleaners	<=50K
28	Female	Black	Bachelors	Married-civ-spouse	Cuba	Private	Prof-specialty	<=50K
37	Female	White	Masters	Married-civ-spouse	United-States	Private	Exec-managerial	<=50K

METHODOLOGICAL STEPS

K-Anonymity

Let $T(A_1, A_2, \ldots\ldots, A_n)$ be a table and QI_T be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if each sequence of values in $T[QI_T]$ appears with **at least** k-occurrences in $T[QI_T]$.

- To implement **k-anonymity,** the following generalizations are applied:
- Age :
  - 1 -> 1-30
  - 2 -> 31-60
  - 3 -> 61-120
- Gender (Suppressed) :
  - Any -> {Male. Female}
- Race (Suppressed) :
  - Any -> {Black, White}
- Education :
  - School -> {1st-4th, 5th-6th,7th-8th,9th, 10th, 11th, 12th}
  - Undergraduate -> {Some-college, Bachelors}
  - Graduate -> {HS-grad, Assoc-voc, Assoc-acdm, Prof-school, Doctorate, Masters}
- Marital Status :
  - spouse present -> {Married-civ-spouse, Married-AF-spouse}
  - spouse not present -> {Divorced, Never-married, Separated, Widowed, Married-spouse-absent}
- Work Class :
  - Government -> {Federal-gov, Local-gov, State-gov}
  - Non-Government -> {Private, Priv-house-serv, Self-emp-not-inc, Self-emp-inc}
  - Unemployed -> {Without-pay, Never-worked}
- Country :
  - A*a -> [Asia, America, Africa] ={United-States, Cambodia, Puerto-Rico, Canada, Outlying-US(Guam-USVI-etc), India, Japan, South China, … }
  - Europe -> {Holand-Netherlands, England, Germany, Greece, Italy, Poland, Portugal, Ireland, France, … }

K-Anonymity Result

Queried Data

Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
48	Male	White	7th-8th	Married-civ-spouse	Italy	Self-emp-not-inc	Craft-repair	<=50K

- For the record shown above, after applying the **k-anonymity** approach on the data-set, we got the following equivalent class as the result :

Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K
2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K
2	Any	Person	School	spouse present	Europe	Non-Government	Craft-repair	<=50K

- It can be seen from above result, that we are still able to identify the Occupation of the record as 'Craft-repair', as all the records in the equivalent contain the same value, leading to Attribute-disclosure.
- To improve upon this, the approach of l-diversity is used.

L-Diversity

An equivalence class is l-diverse if it contains at least 'l' well represented values for sensitive attribute S. A table is l-diverse if every equivalence class is l-diverse.

Distinct l- Diversity

An equivalence class has distinct l-diversity if it has at least 'l' well - defined sensitive values. When each equivalence class in the table has distinct l-diversity, the table is said to be having distinct l-diversity.

- To implement **l-diversity** :
- The Occupations were divided into 3 groups so that each group had almost equal frequencies in the original data set and the l-diversity principle was then applied on each group :
  - Group 1 = {Tech-support, Craft-repair, Prof-specialty, Machine-op-inspct}
  - Group 2 = {Sales, Exec-managerial, Handlers-cleaners}
  - Group 3 = {Other-service, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv,… }

Generalizations (with addition/changes to K-anonymity)

- Work Class :
  - Employed -> {Federal-gov, Local-gov, State-gov, Private, Priv-house-serv, … }
  - Unemployed -> {Without-pay, Never-worked}
- Education :
  - School -> {1^{st}-4^{th}, 5^{th}-6^{th},7^{th}-8^{th},9^{th}, 10^{th}, 11^{th}, 12^{th}}
  - Graduate -> {Some-college, Bachelors, HS-grad, Prof-school, Doctorate, Masters}
- For some records, the following generalizations were applied on Age :
  - 4 -> {31-120}
  - 5 -> {1-30} OR {61-120}

L-Diversity Result

Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
2	Any	Person	Any	Any	Europe	Employed	Craft-repair	<=50K
2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
2	Any	Person	Any	Any	Europe	Employed	Exec-managerial	<=50K
2	Any	Person	Any	Any	Europe	Employed	Sales	<=50K

- After implementing L-diversity, we get the above output for the query which improves the flaw of K-anonymity.
- But for the following equivalence class

Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
5	Any	Person	School	spouse present	A*a	Employed	Craft-repair	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Transport-moving	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K
5	Any	Person	School	spouse present	A*a	Employed	Exec-managerial	<=50K

- We have 5 records out of 7 which have Occupation as 'Exec-managerial' (71.4 %). In the original dataset, there were 663 out of 5000 records with Occupation as Exec-managerial (13.26 %).
- This leads to skewness attack, where we can infer more details about a particular class as compared to the original dataset.
- The quality of data is also reduced.

T-Closeness

An equivalence class has t-closeness, if the distance between distribution of sensitive attribute in that class and the distribution of that attribute in the overall table is not more than a threshold 't'.

Metric for calculating the distance: **Earth Mover Distance (EMD)**

EMD : minimal amount of work needed to transform one distribution to another. Let one distribution be assumed as mass of earth lying in space. The second distribution is assumed as holes in space. We need to fill masses in holes. This is EMD.

EMD between distributions P = $(p_1, p_2, p_3, \ldots)$ and Q = $(q_1, q_2, q_3, \ldots)$

$f_{i,j}$ = flow of mass from element i of P to element j of Q

$d_{i,j}$ = ground distance between element i of P and element j of Q

EMD : $D[P,Q] = \sum \sum d_{i,j} f_{i,j}$

- To implement **t-closeness** :

All generalizations for K-anonymity were used. However some changes were done which are as follows:

- Education :
  - School -> {1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th, 12th}
  - Undergraduate -> {Some-college, Bachelors}
  - Graduate -> {HS-grad, Assoc-voc, Assoc-acdm, Prof-school, Doctorate, Masters}
- For some records, the following generalizations were applied on the Education attribute :
  - School/Grad -> {1st-4th, 5th-6th, 7th-8th, 9th, Prof-school, Doctorate, Masters}
  - School/UG -> {1st-4th, 5th-6th, 7th-8th, 9th, Some-college, Bachelors}
- For some records, the following generalizations were applied on the Age:
  - 4 -> {1-60}
  - 5 -> {1-120}
- For some records, the following generalizations were applied on the Salary:
  - Any -> <=50K or >50K

T-Closeness Result

- After implementing **t-closeness** approach, we get the following class for the same record :

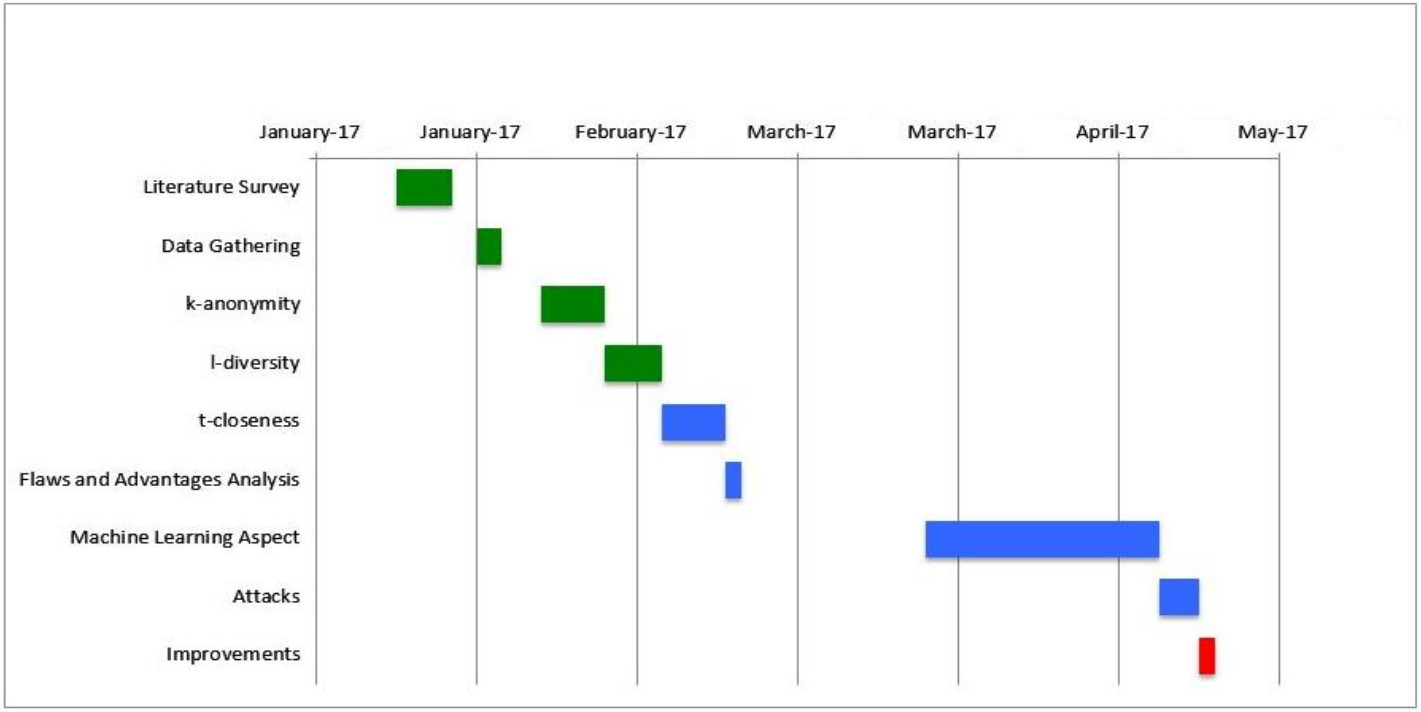
Age	Sex	Race	Education	MaritalStatus	Country	WorkClass	Occupation	SalaryClass
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Other-service	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Transport-moving	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Prof-specialty	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Craft-repair	Any
4	Any	Person	School/Grad	spouse present	Europe	Non-Government	Exec-managerial	Any

- We observe that we overcome the limitations of both the k-anonymity approach (attribute disclosure) as well as l-diversity approach (skewness attack).

✓The advantages of data security comes at the price of reduced quality of the released data-set.

✓To better incorporate the results of the approaches, it is beneficial to implement k-anonymity and t-closeness together.

Activity Time-Chart



WORK TO BE DONE

- Machine Learning Algorithms to be implemented for feature extraction.
  - Sampling, Clustering
    - By picking some arbitrary records as training tuples from different equivalence classes

    or
    - By picking some clusters of data tuples as training set

REFERENCES

- [1]. L. Sweeney, "K-Anonymity: A model for protecting privacy", in International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002
- [2]. Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer and Muthuramakrishnan Venkitasubramaniam, "L-Diversity : Privacy beyond K-Anonymity", in Data Engineering, 2006. ICDE '06. Proceedings of the 22nd International Conference on 3-7 April 2006
- [3]. Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity", in Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on 15-20 April 2007
- [4]. https://archive.ics.uci.edu/ml/datasets/Adult

THANK YOU!