A Comparative study on Privacy Preservation techniques using Fisher Yates, Mondrian and Datafly algorithms

1. **Prof. T.Bhasker Reddy**
Professor, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapuramu, A.P. India.

2. **Deepu J.J.Lazarus**
Research Scholar, Department of Computer Science & Technology, Sri Krishnadevaraya University, Anantapuramu, A.P. India.

3. **Dr. M.Suresh Babu**
Professor, Department of Computer Science & Engineering, K.L. University off Campus, Hyderabad.

Abstract

Privacy preservation assumes an indispensable job in forestalling individual private information protected from the imploring eyes. Anonymization strategies empower production of data which license investigation and assurance security of touchy data in information against assortment of assaults. It cleans the data. It can likewise keep the individual unknown utilizing encryption method. There are different anonymization techniques, data masking, data shuffling techniques and calculations which are examined in this paper. Paper centers around Generalization and Suppression procedures and depicts Fisher Yates, Datafly and Mondrian calculations and furthermore talks about their correlation.

Keywords : 1. Privacy preservation 2. Encryption technique 3. Generalization and Suppression 4. Shuffling.

1.0 Introduction

Privacy preservation is achieved for streaming data by using one of the anonymization techniques called 'shuffling' with Big data concept. K- anonymity, t-closeness, L-diversity are usually used techniques for privacy concern in a data. But in all these techniques information loss and data utility are not preserved very well. Dynamically Anonymizing Data Shuffling technique is used to overcome this information loss and also to improve data utility in streaming data. From vast variety of streaming data from medical records and trends are changing continuously based on the condition of the patient. So to search a key word try to generate many results in short time. The ultimate aim of the paper is to provide researchers with the ability to

IJSER

analyze data from across the globe. Hence a successful data protection mechanism must ensure that the anonymized data provides results that are similar to that using the original data.

Algorithm

Begin

Step1: Pick a Record O_R from the Original Table**(T_O)**

- **Step2:**Generate a Random value R using the Fisher Yates Algorithm
- **Step3:** The R condition is 1 <=R <= Count of Rows of Disease Master Table **(T_{DM})**
- **Step4:** If the problem of R^{th} Record of **T_{DM}**matches with the problem of **T_O** then go to Step2 otherwise go to next step
- **Step5:** Modify Name, Age and PIN Codevalues of O_R with the values of Name,Age and PIN Code of the R^{th}record and store into Modified table (T_{Mod})
- **Step6:** If the value exceeds the count of rows then make a cycle
- **Step7:**Modify the Problem of the O_R with the problem of R from Master Table and store into Modified table(T_{Mod})
- **Step8:** Store the R value in Key table (T_K) of private cloud

End

***** The problems / issues in Table T_o are not all diseases, they are symptoms of a disease.

The T_o table can be converted into a frequency table with the number of people having a disease based on the symptom.

A naive bayes algorithm can be applied to confirm the disease based on the symptoms.

IJSER

2.0 Fisher Yates Algorithm

SD - Streaming data
StD - Structured data
AS - Attribute selection data
AN - Anonymized data

1. Input the credential details to access the stream data (SD)
2. Using Apache Flume Source, grab Streaming Data to processing place
3. Redirect SD to store in database
4. If Streaming Data is unstructured or semi structured Convert to structured data (SD)
5. else Streaming Data = Structured data
6. Apply attribute selection
7. Find sensitive attribute, quasi identifier, insensitive attribute
8. If Structured data is sensitive attribute discard that.
9. Else if Structured Data is insensitive attribute discard that
10. Else Structured data = Attribute selection data
11. Apply shuffle method

For (i=AS length -1, i > 0, i-)
A = AS(index)
AS(index) = AN[i]

AN[i] = A
Return Anonymized data (AN)

2.1. Mondrian for L-Diversity

Mondrian is a Top-down greedy data anonymization algorithm for relational dataset, proposed by Kristen LeFevre. To our knowledge, Mondrian is the fastest local recording algorithm, which preserve good data utility at the same time. Although LeFevre gave the pseudocode in his papers, the original source code is not available. We can find the Java implementation in Anonymization Toolbox. Mondrian for L-diversity is based on InfoGain Mondrian , but more simple.

We used both adult and INFORMS dataset in this implementation. For clarification, we transform NCP to percentage. This NCP percentage is computed by dividing NCP value with the number of values in dataset. The range of NCP percentage is from 0 to 1, where 0 means no information loss, 1 means loses all information (more meaningful than raw NCP, which is sensitive to size of dataset).

The Final NCP of Mondrian on adult dataset is about 79.04%, while 11.01% on INFORMS data (with L=5).

Implementation is based on Python. We can run Mondrian in following steps:

1. Download (or clone) the whole project.
2. Run "anonymized.py" in root dir with CLI.

Parameters:

run Mondrian with adult data and default l(l=5)

python anonymizer.py

run Mondrian with adult data l=10

python anonymized.py a 10

a: adult dataset, i: INFORMS dataset

l: varying l, qi: varying qi numbers, data: varying size of dataset, one: run only once

2.2. Datafly Algorithm

Datafly is a greedy heuristic algorithm which is used to anonymize a table in order to satisfy k-anonymity. Currently supports the CSV format.

Usage

datafly.py [-h] --private_table PRIVATE_TABLE --quasi_identifier
QUASI_IDENTIFIER [QUASI_IDENTIFIER ...]

IJSER

--domain_gen_hierarchies DOMAIN_GEN_HIERARCHIES
[DOMAIN_GEN_HIERARCHIES ...] -k K --output OUTPUT
Python implementation of the Datafly algorithm. Finds a k-anonymous representation of a table.

optional arguments

--private_table PRIVATE_TABLE, -pt PRIVATE_TABLE
Path to the CSV table to K-anonymize.

--quasi_identifier QUASI_IDENTIFIER [QUASI_IDENTIFIER ...], -qi QUASI_IDENTIFIER [QUASI_IDENTIFIER ...]
Names of the attributes which are Quasi Identifiers.

--domain_gen_hierarchies DOMAIN_GEN_HIERARCHIES [DOMAIN_GEN_HIERARCHIES ...], -dgh DOMAIN_GEN_HIERARCHIES [DOMAIN_GEN_HIERARCHIES ...]
Paths to the generalization files (must have same order as the QI name list.

-k K Value of K.

output

OUTPUT, -o OUTPUT
Path to the output file.

3.0 Domain Generalization Hierarchy file format

For each Quasi Identifier attribute it must be specified a corresponding Domain Generalization Hierarchy, which is used to generalize the attribute values. Each DGH is specified through a DGH file, which in each line specifies the hierarchy relationship of a value for that attribute. For example, for an attribute age, the file could be in this format:

...
42,30-45,30-60,1-60,1-120
43,30-45,30-60,1-60,1-120
44,30-45,30-60,1-60,1-120
45,30-45,30-60,1-60,1-120
46,45-60,30-60,1-60,1-120 ...

As shown above each line specifies for a value not generalized (generalization level 0) its hierarchy relationship in the form level 0,level 1,level 2,...,level n (from not-generalized to most generic value).

Original Table (T_O)

ID	Name	Age	PIN Code	Problem / Issue
P00001	Raju	30	123456	Cancer
P00002	Gopal	45	654321	Chest Pain
P00003	Mahi	60	324561	Head Ache
P00004	Abhi	23	234561	Kidney Problem
P00005	Salman	54	345216	ILD
P00006	Rajesh	40	432165	Dengue Fever
P00007	Anvesh	60	122546	Paralysis

IJSER

P00008	Sumati	55	213456	Arthritis
P00009	Raji	38	432155	Viral Fever
P00010	Laxmi	48	234546	Heart Defect
P00011	Suren	36	543122	BP
P00012	Teja	23	236541	Dengue Fever
P00013	Manu	21	132654	Malaria
P00014	Ramya	30	654321	Jaundice
P00015	Laxman	22	432651	Hepatitis
P00016	Anand	21	235644	Influenza
P00017	Valli	32	443215	Malnutrition
P00018	Varun	30	336542	Pneumonia
P00019	Jai	38	634215	Obesity and Genetics
P00020	Ramani	38	453216	Sinus
P00021	Sumant	26	345221	Swine Influenza
P00022	Nagesh	46	122334	Sugar
P00023	Nayesh	42	211335	Tuberculosis
P00024	Srinivas	43	346782	Typhoid Fever
P00025	Raghu	15	515643	Viral Fever
P00026	SriSri	18	505473	Calcium Defect
P00027	Veeren	20	500018	Bronchitis
P00028	Varun	33	500116	Measles
P00029	Preeti	22	230011	Cervical Cancer
P00030	Praveen	32	432765	Ulcer
:	:	:	:	:
:	:	:	:	:
P05000	Rangvi	43	634829	Malaria

Disease Master Table (T_{DM})

DCode	Disease
0001	Allergie
0002	Asthma.
0003	Cancer.
0004	Celiac Disease.
0005	Crohn's & Colitis.
0006	Heart Disease.
0007	Infectious Disease.
0008	Liver Disease.
0009	Lupus
0010	Multiple Sclerosis
0011	Relapsing Polychondritis
0012	Rheumatoid Arthritis
0013	Scleroderma
0014	Type 1 Diabetes
0015	smallpox.
0016	the common cold and different types of flu.
0017	measles, mumps, rubella, chicken pox, and shingles.
0018	hepatitis.
0019	herpes and cold sores.
0020	polio.
0021	rabies.
0022	Ebola and Hanta fever.
0023	HIV
0024	Severe acute respiratory
	syndrome (SARS)
0025	dengue fever, Zika, and Epstein-Barr
0026	Unintentional injuries. ...
0027	Chronic lower respiratory disease. ...
0028	Stroke and cerebrovascular diseases. ...
0029	Alzheimer's disease.
0030	Diabetes.
0031	Influenza and pneumonia.
0032	Kidney disease.
0033	Salmonella.
0034	Norovirus (Norwalk Virus)
0035	Campylobacter.
0036	E. coli.
0037	Listeria.
0038	Clostridium perfringens.
0039	Tuberculosis.
0040	Leprosy.
0041	Chronic Empyema.
0042	AIDS.
0043	II Neoplasms.
0044	Malignant Diseases.
0045	III Endocrine, Nutritional and Metabolic Disorders.
0046	Diabetes Mellitus
0047	Chondrosarcoma
0048	Ewing's sarcoma
0049	Malignant fibrous histiocytoma of bone/osteosarcoma
0050	Osteosarcoma
0051	Rhabdomyosarcoma
0052	Heart cancer
0053	Astrocytoma
0054	Brainstem glioma
0055	Pilocytic astrocytoma
0056	Ependymoma
0057	Primitive neuroectodermaltumor
0058	Cerebellar astrocytoma
0059	Cerebral astrocytoma
0060	Glioma
0061	Medulloblastoma
0062	Neuroblastoma
0063	Oligodendroglioma
0064	Pineal astrocytoma
0065	Pituitary adenoma
0066	Visual pathway and hypothalamic glioma
0067	Breast cancer
0068	Invasive lobular carcinoma
0069	Tubular carcinoma

IJSER

0070	Invasive cribriform carcinoma
0071	Medullary carcinoma
0072	Male breast cancer
0073	Phyllodestumor
0074	Inflammatory Breast Cancer
0075	Adrenocortical carcinoma
0076	Islet cell carcinoma (endocrine pancreas)
0077	Multiple endocrine neoplasia syndrome
0078	Parathyroid cancer
0079	Pheochromocytoma
0080	Thyroid cancer
0081	Merkel cell carcinoma
0082	Uveal melanoma
0083	Retinoblastoma
0084	Anal cancer
0085	Appendix cancer
0086	cholangiocarcinoma
0087	Carcinoid tumor, gastrointestinal
0088	Colon cancer
0089	Extrahepatic bile duct cancer
0090	Gallbladder cancer
0091	Gastric (stomach) cancer
0092	Gastrointestinal carcinoid tumor
0093	Gastrointestinal stromal tumor (GIST)
0094	Hepatocellular cancer
0095	Pancreatic cancer, islet cell
0096	Rectal cancer
0097	Bladder cancer
0098	Cervical cancer
0099	Endometrial cancer
0100	Extragonadal germ cell tumor
0101	Ovarian cancer
0102	Ovarian epithelial cancer (surface epithelial-stromal tumor)
0103	Ovarian germ cell tumor
0104	Penile cancer
0105	Renal cell carcinoma
0106	Renal pelvis and ureter, transitional cell cancer
0107	Prostate cancer
0108	Testicular cancer
0109	Gestational trophoblastic tumor
0106	Ureter and renal pelvis, transitional cell cancer

0107	Urethral cancer
0108	Uterine sarcoma
0109	Vaginal cancer
0110	Vulvar cancer
0111	Wilms tumor
0112	Esophageal cancer
0113	Head and neck cancer
0114	Nasopharyngeal carcinoma
0115	Oral cancer
0116	Oropharyngeal cancer
0117	Paranasal sinus and nasal cavity cancer
0118	Pharyngeal cancer
0119	Salivary gland cancer
0120	Hypopharyngeal cancer
0121	Acute biphenotypicleukemia
0122	Acute eosinophilic leukemia
0123	Acute lymphoblastic leukemia
0124	Acute myeloid leukemia
0125	Acute myeloid dendritic cell leukemia
0126	AIDS-related lymphoma
0127	Anaplastic large cell lymphoma
0128	Angioimmunoblastic T-cell lymphoma
0129	B-cell prolymphocytic leukemia
0130	Burkitt's lymphoma
0131	Chronic lymphocytic leukemia
0132	Chronic myelogenous leukemia
0133	Cutaneous T-cell lymphoma
0134	Diffuse large B-cell lymphoma
0135	Follicular lymphoma
0136	Hairy cell leukemia
0137	Hepatosplenic T-cell lymphoma
0138	Hodgkin's lymphoma
0139	Hairy cell leukemia
0140	Intravascular large B-cell lymphoma
0141	Large granular lymphocytic leukemia
0142	Lymphoplasmacytic lymphoma
0143	Lymphomatoid granulomatosis
0144	Mantle cell lymphoma
0145	Marginal zone B-cell lymphoma

0146	Mast cell leukemia
0147	Mediastinal large B cell lymphoma
0148	Multiple myeloma/plasma cell neoplasm
0149	Myelodysplastic syndromes
0150	Mucosa-associated lymphoid tissue lymphoma
0151	Mycosis fungoides
0152	Nodal marginal zone B cell lymphoma
0153	Non-Hodgkin lymphoma
0154	Precursor B lymphoblastic leukemia
0155	Primary central nervous system lymphoma
0156	Primary cutaneous follicular lymphoma
0157	Primary cutaneous immunocytoma
0158	Primary effusion lymphoma
0159	Plasmablastic lymphoma
0160	Sézary syndrome
0161	Splenic marginal zone lymphoma
0162	T-cell prolymphocytic leukemia
0163	Basal-cell carcinoma
0164	Melanoma
0165	Skin cancer (non-melanoma)
0166	Bronchial adenomas/carcinoids
0167	Small cell lung cancer
0168	Mesothelioma
0169	Non-small cell lung cancer
0170	Pleuropulmonaryblastoma
0171	Laryngeal cancer
0172	Thymoma and thymic carcinoma
0173	AIDS-related cancers
0174	Kaposi sarcoma
0175	Arrhythmias
0176	Deep venous thrombosis
0177	Heart failure
0178	Hypertension
0179	Inflammatory/infectious disease
0180	Endocarditis
0181	Myocarditis
0182	Pericardial disease Ischemic disease
0183	Peripheral vascular disease
0184	Shock

IJSER

0185	Syncope
0186	Valvular disease
0187	Acute/chronic respiratory failure
0189	Asthma
0190	Assisted ventilation
0191	Chronic obstructive pulmonary disease
0192	Interstitial lung disease
0193	Neoplasms/Malignancies
0194	Pleural disease
0195	Pneumonia
0196	Pneumothorax
0197	Pulmonary embolism
0198	Tuberculosis
0199	Bleeding
0200	Gastroparesis
0201	Lower gastrointestinal disorders
0202	Bleeding
0203	Constipation
0204	Diverticulosis
0205	Inflammatory bowel disease
0206	Infections
0207	Ischemia
0208	Obstruction
0209	Hepatobiliary disorders
0210	Gallbladder disease
0211	Liver disease
0212	Pancreatitis

0213	Malnutrition
0214	Neoplasms/Malignancies
0215	Endocrinology
0216	Adrenal gland disease
0217	Diabetes mellitus
0218	Diabetic ketoacidosis
0219	Hyperosmolar nonketotic coma
0220	Hypoglycemia/hyperglycemia
0221	Neoplasms/Malignancies
0222	Thyroid
0223	parathyroid gland disease
0224	Urology Acid-base disorders
0225	Acute kidney injury
0226	Chronic kidney disease
0227	End-stage renal disease
0228	Electrolyte disorders
0229	Diabetes insipidus
0230	Syndrome of inappropriate antidiuretic hormone secretion (SIADH)
0231	Neoplasms
0232	Malignancies
0233	Nephrolithiasis Urinary tract infections
0234	Anemias Bone marrow disorders
0235	Coagulopathies
0236	Hematologic malignancies
0237	Hematologic/oncologic emergencies
0238	Neutropenic fever
0239	Solid organ tumors
0240	Infectious Diseases
0241	Fungal infections
0242	HIV/AIDS
0243	Hospital-acquired infections
0244	Infection in

	immunocompromised host
0245	Sepsis syndrome
0246	Skin and soft-tissue infections
0247	Travel-related illnesses Viral infections
0248	Allergy
0249	Immunology/Rheumatology
0250	Anaphylaxis
0251	Autoimmune disorders
0252	Drug reactions
0253	Inflammatory/infectious arthropathy
0254	Certificate of Added
0255	Qualifications Examination in Hospital Medicine
0256	Disease and Disorder
0257	Organ transplantations
0258	Kidney Failure
0259	Sarcoidosis
0260	Neurology Delirium
0261	Dementia
0262	Encephalopathy Meningitis
0263	Encephalitis Myopathy Neoplasm
0264	Malignancies Paraplegia
0265	Quadriplegia Peripheral neuropathy
0266	Seizure disorders
0267	Stroke
0268	Ischemic
0269	Hemorrhagic
0270	Psychiatry Mood disorders
0271	Psychosis Substance use disorders
0272	Suicide/Overdose

IJSER

Modified Table (T_{MoD})

ID	Name	Age	PIN Code	Problem / Issue
P00001	Sumati	55	213456	Liver Disease
P00002	Jhon	20	546321	Gastroparesis
P00003	Sanvi	53	264531	Non-Hodgkin lymphoma
P00004	Abhi	23	234561	Kidney Problem
P00005	Virini	45	416532	Cervical cancer
P00006	Ramani	24	365421	Breast cancer
P00007	Rajesh	35	524612	Brainstem glioma
P00008	Venu	76	213456	Kidney disease
P00009	Nayesh	42	211335	HIV
P00010	Vishwesh	15	542364	II Neoplasms
:	:	:	:	:
:	:	:	:	:
P05000	Laxmi	48	234546	Multiple Sclerosis

Key Table(T_K)

ID	Key
P0001	8
P0002	20
P0003	200
P0004	153
P0005	98

P0006	67
P0007	54
P0008	32
P0009	23
P0010	43
:	:
:	:
P05000	10

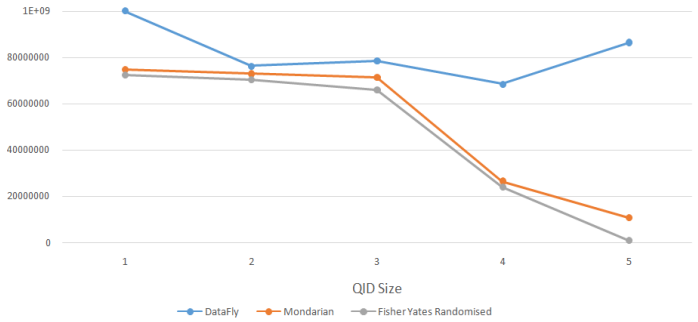

Figure 1: Data sets and QID Size

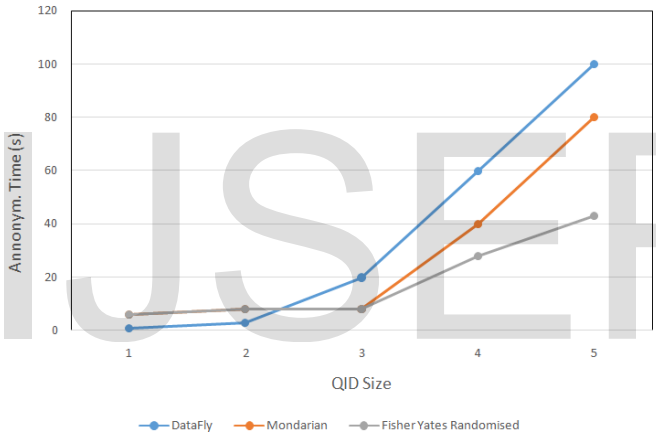

Figure 2 : Anonym Time and QID Size

4.0 Conclusion

Big Data is an emerging field and as big data contains individual specific information privacy is a major security concern hence there is a need to protect knowledge hidden in big data. Privacy Preservation and why Privacy Preservation in Big Data. There are various Privacy Preservation techniques such as Anonymization, Perturbation, Randomized Response, Condensation, and Cryptography. A detailed study of Anonymization Techniques used in Privacy Preservation in Big Data is done. k-anonymity and its various k-anonymity operators are explained in detail. Main focus of the study is Privacy Preservation using Anonymization Technique and a detailed study three Anonymization Algorithms are explained – Fisher Yates algorithm, Datafly Algorithm and Mondrian Algorithm. Fisher Yates algorithm is suitable for data streaming, Datafly algorithm is more suitable for synthetic dataset while Mondrian algorithm is more

IJSER

suitable for real dataset. Datafly Algorithm performs better when dataset is small. We have done a detailed study of thhree algorithms and achieved a detailed comparison of these three algorithms based on six parameters.

5.0 References

[1] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on. IEEE, 2006.

[2] Workload-aware Anonymization Techniques for Large-scale Datasets ACM Trans. Database Syst., ACM, 2008, 33, 17:1-17:47

[3] P. P. de Wolf, J.M.Gouweleeuw, P. Kooiman, L. Wil-lenborg, Reflections on PRAM. Statistical data protection, proceedings of the conference, Lisbon, 1998.

[4] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal kAnonymization. In Proc. of ICDE-2005, 2005 [13] Stephen Lee Hansen and Sumitra Mukherjee. A Polynomial Algorithm for Optimal Microaggregation

[5] Matthias Schmid1 and Hans Schneeweiss, 2005, The Effect of Microaggregation Procedures on the Estimation of Linear Models: A Simulation Study

[6] X. Zhang, C. Liu, S. Nepal, C. Yang, J. Chen, ”Privacy Preservation over Big Data in Cloud Systems,” Security, Privacy and Trust in Cloud Systems, pp 239-257, Springer.

[7] J. Sedayao, Enhancing cloud security using data anonymization, White Paper, Intel Coporation.

[8] Top Ten Big Data Security and Privacy Challenges, Technical report, Cloud Security Alliance, November 2012

[9] J. Salido, ”Differential privacy for everyone,” White Paper, Microsoft Coproration, 2012.

[10] Big Data Privacy Preservation, Ericsson Labs, http://labs.ericsson.com/blog/privacy-preservation-in-big-data-analytics

[11] O. Heffetz and K. Ligett, Privacy and data-based research, NBER Working Paper, September 2013.

[12] M. V. Dijk, A. Juels, ”On the impossibility of cryptog-raphy alone for privacy-preserving cloud computing,” Proceed-ings of the 5th USENIX conference on Hot topics in security, August 10, 2010, pp.1-8.

[13] F. H. Cate, V. M. Schnberger, ”Notice and Consent in a World of Big Data,” Microsoft Global Privacy Summit Summary Report and Outcomes, Nov 2012.

IJSER

[14] P. Samarati and L. Sweeney, ”Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and cell suppression,” Technical report, SRI International, 1998.

IJSER