

About the Book :

The main goal of this research is to secure data in cloud using data anonymization technique called K-Anonymity. The central objective of the thesis is to protect the identity of individuals from datasets and keep the data utility for further studies. Specifically, the main objectives of this thesis work are as follows: To protect the confidentiality of the individuals and to ensure that the right data is getting to the right people in the right format. To secure the incremental data in cloud using appropriate anonymization method depending upon the nature and purpose of data analysis. Protecting de-identifiability of individuals from dataset by applying data protection techniques. Equal balance between privacy and utility. Privacy preservation is hard requirement that must be satisfied and utility is the measure to be optimized. Time taken to update the records is compared with existing system.

About the Author :

Dr. GANDRAKOTA KATHYAYANI is a distinguished figure in the field of academia, holding a prestigious position as a Professor in the Department of Applied Mathematics at Yogi Vemana university, KADAPA(Andhra Pradesh), With an M.Sc degree, PGDCA, and Ph.D degree, her qualifications are impressive. Her illustrious teaching career spans over 25 years, starting from June 1998 and continuing to the present. Her expertise spans Fluid Dynamics, Heat Transfer Flows, Mathematical Modeling, Boundary Layer Flows, Magneto Hydrodynamics, and Peristaltic Flows, with substantial contributions in research and academia. Guiding academic pursuits, she has mentored 1 M.Phil and 3 Ph.D students to completion, while currently supervising the research of 6 Ph.D students. Her scholarly publications include 27 articles in national and international journals, along with authoring 2 influential books. Engaging in continuous learning, she has participated in 45 workshops, presented 72 papers at conferences, and contributed insights to 233 conferences and seminars. Recognized for her contributions, she has received 3 prestigious awards from esteemed academic bodies. In administrative roles, she has chaired the Board of Studies (B.O.S) in 15 universities and served on editorial boards of UGC CARE listed journals. An active member of professional mathematics bodies and societies, she fosters a culture of continuous learning. Her organizational skills shine through in coordinating 4 national conferences, 1 international symposium, and 36 workshops, conferences, seminars, and webinars. Beyond academia, her commitment is evident in co-curricular activities, NSS initiatives, extension activities, and holistic student development. Contributing to academic and administrative committees, she shapes the academic landscape through diligent efforts. In conclusion, Prof. G. Kathyayani's distinguished 25-year teaching and research career, accomplishments, impactful mentorship, and leadership roles make her a valuable asset to Yogi Vemana University and the academic community.



Rs.600 /-



Research Culture Society and Publication
An International ISBN Books Publisher
www.researchculturesociety.org



A SYSTEMATIC COMPARISON AND EVALUATION OF PRIVACY PRESERVATION TECHNIQUES

A SYSTEMATIC COMPARISON AND EVALUATION OF PRIVACY PRESERVATION TECHNIQUES



Prof. G. KATHYAYANI

Research Culture Society and Publication
www.researchculturesociety.org



A SYSTEMATIC COMPARISON AND EVALUATION OF PRIVACY PRESERVATION TECHNIQUES

Dr. G. KATHYAYANI

Professor

ISBN : 978-93-92504-33-4

Published by :



Research Culture Society and Publication

www.researchculturesociety.org

A SYSTEMATIC COMPARISON AND EVALUATION OF PRIVACY PRESERVATION TECHNIQUES

- Dr. G. KATHYAYANI

Copyright: © The research work, information compiled as a theory with other contents are subject to copyright taken by author(s) / editor(s) / contributors of this book. The author(s) / editor(s)/ contributors has/have transferred rights to publish book(s) to 'Research Culture Society and Publication'.

Imprint:

Any product name, brand name or other such mark name in this book are subjected to trademark or brand, or patent protection or registered trademark of their respective holder. The use of product name, brand name, trademark name, common name and product details and distractions etc., even without a particular marking in this work is no way to be constructed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Disclaimer:

The author (s), contributors and editor(s) are solely responsible for the content, images, theory, and datasets of the papers compiled in this book. The opinions expressed in our published works are those of the author(s)/contributors and do not reflect our publication house, publishers and editors, the publisher does not take responsibility for any copyright claim and/or damage of property and/or any third parties claim in any matter. The publication house and/or publisher is not responsible for any kind of typo-error, errors, omissions, or claims for damages, including exemplary damages, arising out of use, inability to use, or with regard to the accuracy or sufficiency of the information in the published work.

Published and Printed at : (First Edition – August, 2023)

Research Culture Society and Publication / Research Culture Society

(Reg. International ISBN Books and ISSN Journals Publisher)

India : C – 1, Radha Raman Soc, At & Po - Padra, Dis - Vadodara, Gujarat, India – 391440.

USA : 7886, Delrosa Avenue, Sanbernardino, CA 92410.

Canada : Loutit Road, Fort McMurray, Alberta, T9k0a2.

Greece : Mourkoussi Str, Zografou, Athens, 15773

Email: RCSPBOOKS@gmail.com / editor@ijrcs.org

www.researchculturesociety.org / www.ijrcs.org

MRP : Rs. 600 /-

ISBN : 978-93-92504-33-4




Research Culture Society and Publication

(Reg. International ISBN Books and ISSN Journals Publisher)

Email: RCSPBOOKS@gmail.com / editor@ijrcs.org

WWW.RESEARCHCULTURESOCIETY.ORG / WWW.IJRCS.ORG

Conference, Seminar, Symposium organization in association/collaboration with different Institutions.
Conference, Seminar, Symposium Publication with ISSN Journals and ISBN Books (Print / Online).






RESEARCH CULTURE SOCIETY & PUBLICATION
International Book and Journals Publisher
With ISBN and ISSN approval
We publish all subject books in all Categories

Book Publication (Print & Online) with ISBN or ISSN
Fiction, Non-Fiction, Collection of Poem - Stories, Critical Theories, Science Fiction,
Biographies, Autobiography, Fantasy etc. (Visit our web for more details)
Thesis / Dissertation converted in to Book, Conference / Seminar Edited Book

Author Guidelines & Support : Quality Publication : Nominal Publication fee

Send Book Manuscript soft copy to :- RCSPBOOKS@gmail.com
OR
Submit online on :- <https://ijrcs.org/book-publication/>

  +91 9033767725 www.ijrcs.org



CALL FOR PAPERS



International
ISSN Journals and
ISBN Books Publisher



International
Peer-Reviewed
Refereed
Indexed
ISSN Approved
High Impact Factor
Journals with
Quality Publication

Research Culture Society Journals
IJIRMF, JRCS, JSHE, JEDI, Shikshan Sanshodhan

Research Study Fields

Research Publication in all subjects / topics of the following study fields :
Science, Engineering, Healthcare Sciences,
Agriculture, Pharmacy, Medicine, Nursing
Commerce, Management, Social Sciences,
Law, Humanities, Education, Life Skills

Free e-Certificates
Digital Object Identification
Nominal Processing Fee



Submit papers to
editor@ijrcs.org
Or
editor@ijirmf.com



TOGETHER WE REACH THE GOAL
www.IJindex.org

<http://jshe.researchculturesociety.org/>
<http://shikshansanshodhan.researchculturesociety.org/>
<http://jedi.researchculturesociety.org/>

WWW.IJRCS.ORG
WWW.IJIRMF.COM

*Dedicated to My
Beloved
Sringeri Sharadamba....*



Acknowledgements

My profound thanks to my Research Gurus **Prof. D.V. Krishna** and **Prof. C. Sankaraiah** for giving me Research knowledge.

I express my reverential thanks, deep sense of everlasting gratitude and earnest admiration to my esteemed Well-wisher, Philosopher **Prof. M.Suresh Babu**, Head, Department of Information Technology, Teegala Krishna Reddy Engineering College (UGC-Autonomous Institution), Meerpet Medbowli, Hyderabad, for his intellectual advice, ingenious suggestions, scholarly guidance, invaluable discussions, invariable counsel, unremitting encouragement, meticulous care and unforgettable help at different stages of my research Project.

I would like to express my gratitude to **Prof.S.Srinivas**, Department of Mathematics, School of Advanced Science, Vellore Institute of Technology (VIT-AP, Amaravathi) Vijayawada and **Prof. K.Marthu Prasad**, Department of Mathematics School of Science, GITAM (Deemed to be University), Hyderabad, Telangana, India, and **Prof. M.Suresh Babu**, Head, Department of Information Technology, Teegala Krishna Reddy Engineering College (UGC-Autonomous Institution), Meerpet Medbowli, Hyderabad to give permission to learning and Analyzing Preservation Techniques and Recent Software Tools in my throughout project and for the opportunity to engage in such a valuable learning experience. The fieldwork has significantly contributed to my understanding of the practical aspects of my field of study.

I was indebted to express my sincere thanks to **Prof. M.Surya Kalavathi** former vice chancellor, **Prof. D.Vvijaya Raghava Prasad**, Former Registrar and Former director of IQAC and **All IQAC team** to encourage and support to do **Seed money Project**.

I convey my thanks to our department *Colleges* and **Prof.K.Krishna Reddy** Dean of Sciences and also I express my gratitude to Honorable **Prof.C.Sudhakar**, Vice chancellor, **Prof.Venkata subbaiah** Registrar, **Prof.Raghunath Reddy**, Principal **Coordinators of Research Cell, Academic Section** and **Controller of Examination Members** of Yogi Vemana University, Kadapa, who helped me in various ways.

I am thankful to the **All authorities** of **Yogi Vemana University**, Kadapa, for providing me necessary facilities for the successful completion of my research project.

I am grateful to acknowledge Non-Teaching Staff, *Mrs.Sarala, Mrs. Nagalakshmi, Mrs. Sumalatha, Mrs. Lakshmi Kesamma*, YOGI VEMAN UNIVERSITY, Kadapa for helping me in pursuing my project work.

Finally, I take the opportunity to thank one and all who has directly or indirectly helped me in completing this task.

I am grateful to the *almighty* for supporting me in all aspects of my research work.

My heartiest thanks to my beloved husband *Mr. N. V. Raghavendran* for his valuable and immeasurable assistance and cooperation. My heartiest thanks to my son *N. Vishnu Bharadwaj*, Parent-in-laws *Late Mr.N. Subbarao, Late Mrs.N. Saradamma*, Parents *Late Mr.G. Nagendra Sarma, Mrs. G. Kalavathamma* and well-wishes of their co-operation during the course of my research work.



KATHYAYANI GANDRAKOTA

TABLE OF CONTENTS

Sr.No	CONTENTS	Page No.
a.	Acknowledgement	6-7
b.	Table of Contents	8
-	CHAPTERS	-
1	INTRODUCTION AND BASIC CONCEPTS	9-27
2	EVALUATION BY USING ANONYMIZATION METHOD OF PRIVACY PRESERVATION TECHNIQUES WITH AN EFFICIENT APPROACH AND A SYSTEMATIC COMPARISON	28-42
3	SECURE DATA DEDUPLICATION SYSTEM FOR INTEGRATED CLOUD AND SEAMLESS CONNECTIVITY WITH PRIVACY AND SECURITY AWARENESS	43-56
4	PRIVACY PRESERVATION IN DYNAMIC DATA THROUGH SYNONYMOUS LINKAGE ON MICROAGGREGATION	57-69

Chapter-1

Introduction

Privacy Preservation in Heterogeneous Database

How to protect patient data & privacy:

Safeguarding patient information is an Enormous errand for medical services associations, as securities should be set up for inside and outer dangers. What's more, HIPAA guidelines include a layer of required boundaries that medical care associations should have set up to be consistent and not face punishments. HIPAA has 2 sorts of rules to safeguard patient data that should be observed: the Protection Rule, and the Security Rule. The Protection Rule safeguards what is known as by and by recognizable data, or PII, and who might approach it, while the Security Rule safeguards generally private wellbeing data (PHI) a covered substance makes, gets, keeps up with, or sends in electronic structure, known as ePHI, and guarantees that main approved clients approach that data. The greatest distinction is the Protection Rule likewise safeguards composed or oral correspondence of PII, while the Security Rule doesn't. The electronic frameworks inside your medical services association hold the most important data, so consistence with the Security Rule is a key stage in how to safeguard patient information. Lead a full gamble examination of how you as of now safeguard patient information.

The initial step of safeguarding patient information is directing a full gamble examination to figure out what frameworks your association has, and which ones should be the most secured. Not all frameworks contain delicate data, and they probably won't require similar shields as something like your EMR framework. With a gamble examination spread out, you can begin investigating what cycles ought to be carried out and the framework shields that should be set up for every framework. For instance, patient information is one of the most pursued kinds of data, so safeguarding against unseemly admittance to your association's EMR framework is an important defense. Step-by-step instructions to safeguard patient information with patient protection observing.

Review controls are expected under the Specialized Protections inside the HIPAA Security Rule. As it expresses, "a covered element should execute equipment, programming, or potentially procedural components to keep and look at access and other movement in data frameworks that contain or utilize ePHI". The greatest framework that utilizes this data is the Electronic Clinical Record framework (EMR) that your office is using. While the standard doesn't express that product should be the strategy utilized, going the course of physically

examining is basically outside the realm of possibilities for consistency groups to deal with. More than 1,000,000 gets are made into the EMR every day, making it difficult to review all get in a sensible measure of time and inclined to blunder. Using a patient protection observing framework that accomplishes the work for you can assist with smoothing out this cycle and guarantee any dubious gets to patient data are hailed and surveyed as quickly as possible.

Guarantee suitable access for outsiders and business partners. The HIPAA Security Rule was laid out for medical services associations, yet additionally for any party that contacts or cooperates with PHI. This incorporates business partners - outer outsiders like cases processors, bill gatherers, clinical transcriptionists, advisors, or bookkeeping firms. Medical care suppliers are expected to show an elevated degree of perceivability and control around business partner exercises to stay in consistence with obligatory principles on safeguarding patient information. This incorporates guaranteeing that business relates just access the patient information they need and that's it.

Here are a few different ways you can keep up with perceivability and command over the entrance your business partners need to EMR and PHI:

- Try to have a Business Partner Understanding (BAA) set up. In consistence with HIPAA, every outsider or business partners are expected to give recorded as a hard copy that they will defend the data.
- Utilize least restricted admittance for business partner access freedoms, so they are just getting to data that is totally basic to their business.
- Execute multifaceted verification to rapidly and proficiently confirm client access.
- Lead an expected level of investment expected by HIPAA, for example, documentation and checking of business partner movement and hazard evaluations.

Fortunately, there are arrangements that can assist with smoothing out these cycles. Remote access instruments worked for medical services associations can normalize and limit access while likewise inspecting business partner action so IT groups aren't stalled by access demands and assembling documentation. These frameworks likewise give medical care associations more inward feeling of harmony about the "who, what, when, why, and how" of business partners getting to EMR and patient documents.

Teach staff on the best way to safeguard patient information:

The most ideal way for staff to safeguard patient information is through ceaseless training. HIPAA schooling is required, yet proceeding with training on accepted procedures for being watchful in different regions, like email, can additionally guarantee consistence and

assurance of patient information. Cycles ought to be set up for fresh recruits, as well as a proceeding with training plan for current representatives. Assisting staff with monitoring what outside dangers resemble and teaching them on why unseemly admittance to clinical records is a serious infringement of HIPAA that can bring about outrageous outcomes can assist with keeping representatives honest about remaining consistent, and ensuring their colleagues stay consistent too. Safeguarding patient information is a prerequisite and a need. Not exclusively can medical services information breaks bring about HIPAA punishments, however it can cost your association more cash after the break to take care of the expense of remaking the trust from patients. Leading a gamble examination of the frameworks inside your association to get a comprehension of what shields should be set up is stage one. From that point, begin executing innovation to shield patient information from insider and untouchable dangers through evaluating worker admittance to patient records, and guaranteeing perceivability and command over business partners by just allowing admittance to the data they need. Worker schooling can guarantee everybody on staff figures out the standards of HIPAA consistence and are cautious with regards to untouchable dangers. Consistence officials have a weighty obligation, yet with an arrangement and innovation, safeguarding patient information can be a piece simpler.

Patient data breaches can strike anywhere:

Ensure you're ready for them all

Ongoing medical services information breaks demonstrate that dangers can emerge from any place. From seller access dangers to insiders sneaking around, the security scene requirements to change.

The dangers of ransom ware and cybercrime strike dread into the hearts of medical services associations across the globe. Also, understandably. The Hive ransom ware bunch, alone, designated north of 1,500 casualties in more than 80 nations before they were penetrated by the FBI. In any case, information breaks don't simply begin with criminal associations looking for millions through misrepresentation and coercion.

That implies you really want to comprehend - and safeguard against - the large number of information penetrates that could influence your association. The following are a couple of kinds of information breaks, and how might have been forestalled them.

Outsider information breaks:

Today's almost difficult to carry on with work without connecting something like one outsider specialist co-op. Furthermore, giving admittance to any outsider seller makes a place of weakness that should be gotten.

Be that as it may, what happens when an outsider seller is hit with a cyberattack? Broward Wellbeing, a medical services framework in Florida, as of late figured out that it can make immense downstream impacts.

Broward Wellbeing encountered an information break when a troublemaker acquired unapproved admittance to their organization through an outsider clinical supplier. The by-and-by recognizable data (PII) uncovered included names, dates of birth, monetary data, telephone numbers, email addresses, Federal retirement aide numbers, protection data, driver's permit numbers, and clinical records - including clinical chronicles, conclusions, and therapies.

The medical services framework examined and doesn't really accept that the information was abused. Notwithstanding, they executed a venture-wide secret key reset and improved safety efforts that included multifaceted confirmation for all clients. They likewise offered free wholesale fraud administrations to the 1,357,879 people influenced.

This information break highlights the requirement for a powerful seller-restricted admittance to the board (VPAM) arrangement. Without one, associations keep themselves open to the dangers of not securing outsider-restricted admittance. It's at this point not protected to furnish your sellers with wide, restricted admittance in light of trust. All things being equal, you really want to step up your own security system to safeguard your association's most vulnerable assault vector.

Insider sneaking around:

Having a VPAM arrangement is vital, yet it's just essential for the image. With information breaks, now and again the call is coming from inside the house. That is the reason the capacity to recognize insider dangers is so significant. Whether because of carelessness or pernicious aim, insider sneaking around can cause a ton of harm.

During a customary security review, DCH Wellbeing found a medical clinic representative had gotten to electronic patient records without approval. Further examination showed that this wasn't the initial time. Between September 2021 and December 9, 2022, the worker got to and saw around 2,530 patient records.

While an information break recuperation master tracked down no abuse of the data, and DCH Wellbeing gave free fraud security administrations to the impacted patients, the representative actually had improper admittance to delicate information - and was gotten past the point of no return.

Anyway, what might have been finished here? In the event that the worker was somebody with approved admittance to patient information - like a medical caretaker - who was abusing those freedoms, a patient protection checking arrangement would resolve the

issue. In the event that the worker wasn't in the clinical group, then better access administration matched with patient security observation would've forestalled - or halted - the unseemly access.

One way or another, when delicate patient information is in question, it is principal to safeguard it.

Hindering breaks before they occur, regardless of where they start:

The shifted beginning stages of information breaks imply that you really want a genuinely strong security system that covers something beyond the "conventional" thought of breaks, including seller access and insider sneaking around.

Manual observing has esteem yet isn't sufficient to stay aware of the quantity of gets to an EHR, or the changed cyberthreats to the present medical care associations. In a perfect world, observing for patient protection ought to be computerized with frameworks that utilize man-made reasoning, AI, and social examination, as well as human skill.

Similarly, risk examination ought to accomplish more than get or discourage cybercriminals who focus on your association. They should likewise address security gambles from workers and outsider merchants.

Patient protection arrangements guarantee HIPAA consistence, yet their motivation is considerably more prominent. On top of forestalling punishments and criminal arraignment, decreasing gamble with social checking and investigation assists work with trust. A climate of trust and security advances patient maintenance, energizes dynamic associations with treatment plans, and supports patient-focused care.

AI-POWERED SOLUTION INCREASES VISIBILITY AND STREAMLINES INVESTIGATIONS

Protecting Patients Privacy:

TGH is quite possibly of the biggest emergency clinic in Florida, with in excess of 1,000 beds, serving a populace of north of 4 million individuals. In excess of 10,000 clients, including clinicians, occupants, volunteers, understudies, project workers, and others all enter the clinic's Legendary EHR frameworks. Checking patient protection in such an enormous and different climate can be an overwhelming recommendation. The emergency clinic has been a client of Imprivata FairWarning beginning around 2015 and has had the option to help with working on the framework's usefulness to foster a more productive and powerful consistency observing arrangement.

Man-made consciousness through rules-based and AI smooth out examinations TGH utilizes Imprivata FairWarning to consequently distinguish known marks of PHI burglary and misuse in light of authoritatively characterized arrangements. "We have rules set up to produce alarms when certain surely known occasions happen, as on the off chance that we see surprisingly high admittance to segment data or expired patient records," calls attention to Maceira. Rules-Based Upheld Approaches spread the word about it simply for TGH to hail action normally connected with normal security infringement like collaborator sneaking around or celebrity sneaking around. The Imprivata arrangement additionally utilizes conduct investigation, contrasting live occasions with authentic information, to recognize bizarre movement and distinguish already obscure danger pointers. The TGH expert constantly gives criticism to assist with preparing the AI model and adjusting its investigation calculations. Imprivata FairWarning likewise utilizes artificial intelligence to distill crude alarms into significant and noteworthy bits of knowledge. The arrangement sifts through misleading up-sides and removes irrelevant cautions, assisting agents with focusing on the issues that imply the best likely danger. Overall, just a single out of each and every 200 cautions prompts an examination. Subsequently, TGH conducts 30% less examinations each month than industry peers, as indicated by Imprivata benchmarks.

Overseen administration improves on tasks and opens up inner assets

TGH exploits Imprivata FairWarning Oversight Security Administrations to speed up chance-to-esteem and smooth out continuous tasks. The Imprivata oversight administrations group goes about as the primary line of examination, proactively investigating and overseeing cautions for TGH, assisting the TGH consistence with joining to save much additional time and exertion. "By far most of alarms are inspected and gotten out by the Imprivata group, which makes our life a lot more straightforward," makes sense of Maceira. "We just engage with surprising cautions that require further examination.

Around 17% of our dubious alerts are simulated intelligence created. Without conduct investigation, these would all go undetected. "Nancy O'Neill, Sr. Head of Corporate Consistence and Security Boss Consistence and Protection Official, Tampa General Medical Clinic" further developed consistence, diminished risk, expanded patient trust Imprivata Fair Warning assists Tampa With generating Emergency clinic safeguard patient information, work on examinations, and further develop HIPAA consistence. The arrangement assists TGH with mechanizing tasks, stay away from expensive fines and claims, and work on persistent trust.

Patient Privacy Versus Patient Confidentiality:

Privacy relates to how a person's information or data is collected Patient protection is the right of people to hold their data back from being openly unveiled (either orally or composed), and the assumption that individual wellbeing data is shared exclusively among themselves and medical services experts.

Secrecy connects with a supplier's commitment to keep up with the patient's security concerning how the data or information is overseen - Patient privacy ensures that the data or information that an individual has revealed, won't be imparted to others without the patient's express consent.

Confidentiality relates to a provider's obligation to maintain the patient's privacy with regard to how the information or data is managed – Patient confidentiality guarantees that the information or data that an individual has disclosed, will not be shared with others without the patient's express permission.

By law, patients are always in control of their information and who has access to it, and the patient always decides who, when, and where to share their health information. With the passage of HIPAA in 1996, protecting patient privacy and ensuring confidentiality of information is of the utmost importance to providers.

Are you Compromising Your Patient's Privacy?

Even the most careful dedicated health care workers can unintentionally breach a patient's privacy or disclose confidential information. For example, a study done by Cornell University students in the Fall of 2011, found that acoustic and visual privacy at Cayuga Medical Urgent Care Center was an issue.

Acoustic Privacy: When patients enter the Urgent Care and queue up to check-in at front registration, they can hear the person in front of them checking in and register. Moreover, there was of auditory_ privacy at the back registration area, as two back registration stations were located right next to each other. Acoustic quality throughout the waiting and staff area is also problematic, as there is absolutely no acoustic privacy; people in the waiting area can hear everything that is going on in alcove area. On the other hand, interestingly people inside the staff area are completely isolated and unable to hear things outside glass area. This resulting in non-responsive staff, patients had to knock the counter couple times to call the staff. The receptionist who is inside the nurse nook area could not hear the patients until moments later. The isolation and not being able to hear outside also makes nurses, doctors, and staffs not aware that their voices can be heard from outside.

Visual Privacy: Glass windows at the reception area allow patients in waiting room to see in. Since the windows are fixed and cannot be closed, patients can hear nurses/receptionists talk. Furthermore, patients can see “Providers” aka doctor board from reception area. Patients sitting at first back station can see across nurses station to front registration computer screen.

The Role of Health Providers in Protecting Patient Privacy:

Clearly, health providers want to do all that they can to ensure that they are protecting patient privacy. For example, as stated in the ANA Code of Ethics, “The nurse advocates for an environment that provides for sufficient physical privacy, including auditory privacy for discussions of a personal nature and policies and practices that protect the confidentiality of information.” That being said, what are some relatively easy, inexpensive, and immediate ways you protect your patients’ privacy?

- Maintain the privacy of the patient’s personal information by creating an environment conducive to a private conversation
- Have personal patient information protected from public view or earshot
- Restrict access to medical records and any patient information that is displayed openly in waiting or treatment areas
- Avoid discussing patient care anywhere it can be overheard by those not directly involved in the care of the patient, or by the public.

Easy Solutions for Protecting Patient Privacy:

Often, because of structural or budgetary constraints, these privacy fixes are usually not as easy as they seem. However, using Screenflex fabric or vinyl covered portable dividing walls, instantaneously allows facilities to provide added levels of privacy. Movable walls can configure quickly and configure easily into any number of shapes, and because they come in 36 different heights and lengths, there’s sure to be one to fit within your institution’s existing space. For added protection, an antimicrobial coating can be applied during the manufacturing process.

Screenflex partitions are ideal for:

- Protecting patient privacy during the intake, triage or immunization process
- Creating confidential consultation areas
- Reducing the spread of infectious diseases with the use of fabric panel surfaces treated with antimicrobial agents
- Providing additional display surfaces for hanging important health notices, information on medical services, and facility hours

- Shield confidential information from public view.
- Allowing facilities the ability to position the portable walls to create temporary rooms within an existing floor plan
- Stores compactly when not in use.

The Goal of Privacy preservation in medical centers is the workforce be able to articulate their duties and responsibilities with regards to:

- Patient Confidentiality
- Patient Privacy
- Secure Computing
- Breach Responsibilities

Confidentiality everyone in the organization is responsible for patient confidentiality:

- Board members
- Executive leadership
- Clinical staff
- Physicians and nurses
- Administrative and clerical staff
- Students and interns
- Volunteers

Confidentiality The following is a list of patient information that must remain confidential:

- Identity (e.g. name, address, social security #, date of birth, etc.)
- Physical condition
- Emotional condition
- Financial information

Confidentiality Guiding Principles:

- Access patient information only if there is a ‘Need to Know’
- Discard confidential information appropriately – (e.g. Locked Trash Bins or Shredders)
- Forward requests for medical records to the Health Information Management Department.
- Do not discuss confidential matters where others might over hear. – (e.g. Cafeteria, Elevator, Buses, or Restaurants)
- Do not leave patients charts or files unattended
- Report suspicious activities that may compromise patient confidentiality to the BMC Privacy Officer

Privacy State & Federal Laws that Protect Patient Privacy:

- Health Insurance Portability & Accountability Act of 1996 (HIPAA) & American Recovery

and Reinvestment Act of 2009 (ARRA) – HITECT breach notification provisions

- Massachusetts regulations and statues – Patient Bill of Rights – 201 CMR 17.00 Standards for the Protection of Personal Information

- The Privacy Act of 1974 Privacy

What is the purpose of HIPAA?

Improve the efficiency and effectiveness of the health care system • Encourage the development of an electronic health record • Establish national standards for electronic transmission of certain health information • Establish national standards to protect health information Ensure patient confidentiality • Protect patient privacy • Build loyalty and trust • Provide exceptional customer service

What is PHI?

PHI stands for Protected Health Information and includes demographic information that identifies an individual and – Is created or received by a health care provider, health plan, employer, or health care clearinghouse. – Relates to the past, present, or future physical or mental health or condition of an individual. – Describes the past, present or future payment for the provision of health care to an individual.

Who has to follow HIPAA?

Anyone who:

- Currently works directly with patients
- Currently sees, uses, or shares PHI as a part of their job
- Currently access any hospital systems, records, tools, and information that may contain PHI

Privacy HIPAA Defines these 18 Elements PHI Identifiers

1. Name
2. Full face photo
3. Finger or voice print
4. Telephone number
5. Address/zip code
6. E-mail address
7. Fax number
8. Internet Protocol (IP) address
9. Uniform Resource Locator (URL)
10. Social security number
11. Medical record number
12. Insurance number

13. Account number
14. All elements of dates
15. Vehicle identifier
16. Certificate/license
17. Device ID/serial number
18. Any unique identifying number, characteristics or code

Where is PHI Found?

- Medical records
- Patient information systems
- Billing information (bills, receipts, EOBs, etc.)
- Test results
- X-rays
- Clinic lists
- Labels on IV bags
- Patient menus
- Conversations
- Telephone notes (in certain situations)
- Patient information on a mobile device

Permitted Uses and Disclosures of PHI Include:

- Treatment of the patient
- Direct patient care
- Coordination of care
- Consultations
- Referrals to other health care providers
- Payment of healthcare bills
- Operations related to healthcare
- Research when approved by an Institutional Review Board (IRB)
- Required by law (e.g. subpoena, court order, etc.) Need-to-know Employees should only use/access the “minimum necessary” information to perform their jobs.

Patient Rights:

- Right to Access
- Any information contained in their medical and billing record
- Right to Amend • Patients may request in writing, an amendment to their medical records if

they feel it contains incorrect or incomplete information

- Right to an Account of Un-Authorized Disclosures • Patients have the right to receive a list of disclosures (information released outside of BMC), other than for treatment, payment, or operations
- Right to Request Special Communications • Patients may ask BMC to contact them via an alternative phone number or address
- Right to Request Restrictions • Patients may request not to be included (opt-out) in the directory. Patient information should not be shared with clergy, friends, or anyone
- Right to Receive a Notice of Privacy Practices • BMC is required to provide a written notice of how we will use and disclose our patients health information
- Right to File a Complaint • Patients have the right to file a complaint without fear of retaliation

Security:

- When we protect patient data, we help build trust between patients and providers.
- Ensure Protected Health Information (PHI) is not disclosed to unauthorized persons.
- Do not send email containing Protected Health Information (PHI) unless it is encrypted.

Breach Awareness:

A breach may have occurred if there has been an unauthorized acquisition, access, use or disclosure of PHI (written, oral, or electronic), that poses a significant risk of financial, reputational, or other harm to a patient.

Employees viewing their own and their minor children's medical record:

- Leaving patient identifiable information in public areas (by reception desk, visible computer screens, copiers) • An employee checking a co-worker's record to look up a birthday or address
- Discussing PHI in a public place where it could be overheard by others
- Inappropriately accessing or disclosing patient information
- Lost, stolen or misplaced laptops and flash drives containing unsecured PHI.

Breach Consequences:

Members of the workforce who fail to follow and uphold Boston Medical Centers privacy and security policies, will be subject to appropriate disciplinary action, up to and including termination.

Tips to Protect Patient Confidentiality, Privacy, and Security

Think before you Act!!

- Never look at a patient's record out of curiosity even with good intentions
- Follow the minimum necessary standard

- Double check names and phone numbers before sending PHI by fax or email
- Log out of your computer if you have to leave your workstation.
- Never share passwords
- Familiarize yourself with the organizations Notice of Privacy Practices

Patient Confidentiality, Privacy, and Security Awareness

Privacy Preserving Data Mining: Models and Algorithms proposes a number of techniques to perform the data mining tasks in a privacy-preserving way. These techniques generally fall into the following categories: data modification techniques, cryptographic methods and protocols for data sharing, statistical techniques for disclosure and inference control, query auditing methods, randomization and perturbation-based techniques. This edited volume also contains surveys by distinguished researchers in the privacy field. Each survey includes the key research content as well as future research directions of a particular topic in privacy.

Privacy Preserving Data Mining: Models and Algorithms is designed for researchers, professors, and advanced-level students in computer science. The main aim is, to develop efficient methodologies to find privacy preserving association rule mining in centralized as well as in distributed database environment without violating any privacy constraints. The issue of privacy constraints for centralized database environment is entirely different from distributed database environment. So the approaches for privacy preserving association rule mining in centralized are also different from approaches in distributed environment. Achieving privacy preserving goals in both these environments in the process of mining is a challenging task.

The goal of achieving privacy in centralized database environment is, to obtain a distorted database which hides the sensitive item sets. When mining task is performed on distorted database all the sensitive rules should be hidden without any side effects. In distributed environment, the goal is to find privacy preserving association rule mining without revealing any individuals private data or information when the database is distributed among n number of sites.

Three methodologies are proposed based on heuristic and exact approach related to centralized database environment to find privacy preserving association rule mining by hiding sensitive item sets with minimum side effects.

Horizontally partitioned data: It divides database into a number of non-overlapping horizontal partitions. In this scenario different places have different record about same entities

or people which are used for mining purposes. Many of these methods use specialized versions of the general approaches discussed for various problems.

Vertically partitioned data: In Vertically partitioned data sets; each site has different number of attributes with same number of transaction. The approach of vertically partitioned mining has been extended to a variation of data mining applications such as decision trees, SVM Classification, Navie Bayes Classifier, and k-means clustering.

Based on heuristic approach, a new methodology is proposed by incorporating suggested Criteria1 and Criteria2 to identify the victim item and selecting suitable supporting transactions efficiently for sanitization purpose to hide the sensitive item sets. The main aim of suggesting Criteria1 and Criteria2 is to minimize the side effects in the process of hiding sensitive item sets. Especially in case of overlapping patterns, suggested two criteria are helpful to speed up the sanitization process by reducing number of modifications.

As another problem, a modified Inline approach is proposed in order to hide the sensitive item sets. When more number of constraints exists in the formed constraint satisfaction problem (CSP), finding an optimum solution is a complex task. To reduce this complexity, a divide and conquer strategy is applied on the constraints based on the dependency of the variables to divide the constraint satisfaction problem into sub constraint satisfaction problems. Parallelism concept is also adopted in the proposed methodology to solve each sub CSP in parallel and individually. In this way, the efficiency of the proposed methodology is increased.

Another exact based approach, partition based hybrid hiding methodology is proposed especially to hide the sensitive item sets for large databases. Large databases produces many constraints with many unknown variables in CSP which makes it difficult to get the optimum solution using Binary Integer Programming (BIP) and sometimes it may lead to unsolvable problem. To avoid this situation, divide and conquer strategy is applied over the large database to partition the database into small size databases. This partitioning can be viewed as a binary tree structure. All the partitioned databases which are at leaf nodes of the tree are solved in parallel and individually by applying BIP. The solutions of these partitioned databases (children) are merged to get the solution of its parent database. This process is repeated at each level of the tree to get the solution for each parent database and finally the solution of the original database which is at root node can be obtained by simply merging the solution of its children databases.

In distributed database environment, three partitioned strategies such as horizontal, vertical and mixed are considered. For each partitioned strategy, methodologies are proposed to find global association rules by preserving individual's privacy constraints.

As a first case in horizontally partitioned model, a methodology which utilizes the sign based secure sum cryptography technique is proposed to find global association rules when trusted party exist. The performance analysis of this model proved that it is efficient in terms of privacy and communication cost.

As a second case of horizontally partitioned database model, in which no party can be treated as Trusted Party when the data is distributed among n number of sites is considered. To enhance the privacy, this model adopted hash based secure sum cryptography technique in the process of finding global association rules. The performance of this model is analyzed in terms of privacy and communication.

In case of vertically partitioned databases, a methodology is proposed in this work to find global association rules without revealing individual's privacy. This methodology utilized scalar product cryptography technique. The representation of each partition database in T-ID form helps to find the scalar product efficiently.

A combination of horizontal and vertical partitioning that is mixed partition is also considered. In the first proposed mixed partitioned model, initially the database is horizontally partitioned into two or more databases which are further partitioned into two or more vertical partitioned databases. In the second model, initially the database is vertically partitioned into two or more databases and then each partitioned databases is further horizontally partitioned into two or more databases.

The methodologies are proposed for these two mixed models. The performance of the proposed methodologies is analyzed in terms of privacy and communication.

Apart from analyzing each proposed methodology in centralized as well as in distributed environment, experiments are conducted with synthetic dataset for each methodology. A comparison analysis is performed.

This study has been carried out to develop methodologies in centralized as well as in distributed environment to find privacy preserving association rule mining without revealing any private data or information.

Three methodologies are proposed to hide the sensitive item sets in centralized database environment. The first methodology is related to heuristic based approach which utilizes two suggested criteria to efficiently find the victim item and its supporting transactions. The

proposed methodology efficiently performs sanitization process especially when overlapping patterns exist in the sensitive item sets.

The proposed modified Inline algorithm efficiently hides the sensitive item sets with minimum side effects by dividing the constraint satisfaction problem into sub constraint satisfaction problems based on the dependency of variables in the constraints with the help of divide and conquer strategy. By solving each sub problem in parallel and individually, the solution for original constraint satisfaction problem is obtained. Experimental results proved that the proposed modified Inline algorithm outperforms heuristic based proposed algorithm in terms of side effects. The modified Inline algorithm not only hides the sensitive item sets but also takes care of not hiding the non sensitive frequent item sets and also avoids generation of new frequent item sets as border revision concepts in the formation of constraint satisfaction problem is considered.

The proposed partition based hybrid hiding methodology efficiently finds the solution for large size database by applying partitioning strategy and parallel concept. Experimental results proved that this methodology outperforms modified Inline and heuristic based approaches in terms of side effects.

The proposed methodology in horizontally partitioned databases with Trusted Party efficiently finds privacy preserving association rule mining by using the adopted sign based secure sum cryptography technique. Another methodology is proposed to handle a situation when no party can be treated as Trusted Party in the process of finding privacy preserving association rule mining with the help of hash based secure sum cryptography technique. The performance analysis of these two methodologies proved that these are efficient in terms of privacy and communication. The experimental results also proved that they are efficient than the existing considered algorithm.

In case of vertically partitioned databases, the proposed methodology which adopted a scalar product technique, efficiently finds global association rules without revealing individual private data or information. The proposed methodologies in case of two mixed partitioned models efficiently finds global association rules by incorporating tree structure for partitioning. Each proposed methodology in centralized as well as in distributed environment is analyzed individually but experiments are also conducted on synthetic dataset for comparison analysis purpose.

The proposed methodologies works well for any number of transactions, any number of attributes and for any number of sensitive item sets in case of centralized database environment. The proposed methodologies in distributed database environment efficiently

finds global association rules for any number of sites, any size of database with any number of attributes.

The work carried out is extended in different ways as follows:

- The methodologies related to heuristic based approach such as randomization, blocking and border based can also be considered as a future work to find privacy preserving association rule mining in centralized database environment.
- The other two cryptography techniques such as secure set union, secure size of set intersection can also be considered to find global association rules in distributed environment by satisfying privacy constraints.

REFERENCES

- [1] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, “From Data Mining to Knowledge Discovery in Databases”, American Association for Artificial Intelligence, pp. 37-54,1996.
- [2] J. Han and M.Kamber, Data Mining Concepts and Techniques, Elsevier 2001.
- [3] R. Agrawal and R. Srikant, “Mining Sequential patterns”, Proc.1995 International Conference on Data Engineering (ICDE’95), pp 3-14, Taipei, Taiwan, March 1995.
- [4] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. “State-of-the-art in privacy preserving data mining”. SIGMOD Record, 33(1):50–57,2004.
- [5] Ahmed HajYasien, “Preserving Privacy In Association Rule Mining”, Ph D.,thesis, Griffith University, June 2007.
- [6] Ming-Syan Chen, Jiawei Han,Yu, P.S., “Data mining: an overview from a database perspective”, IEEE Transactions on Knowledge and Data Engineering, Vol. 8 No. 6, pp 866 – 883,1996.
- [7] Yongjian Fu, “Data mining: Tasks, techniques and Applications”, Department of Computer Science”, University of Missouri- Rolla,1997
- [8] Michael Goebel , Le Gruenwald, “A Survey Of Data Mining And Knowledge Discovery Software Tools”, SIGKDD Explorations, ACM SIGKDD, Vol:1, Issue 1, pp 20- 33, June 1999.
- [9] Thair Nu Phyu ,”Survey of Classification Techniques in Data Mining”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2009, Vol I,IMECS-2009, Hong Kong, 2009.
- [10] Yongjian Fu, Distributed data mining: Overview, University of Missouri-Rolla, 2001.

- [11] R Agarwal, T Imielinski and A Swamy, “Mining Association Rules between Sets of Items in Large Databases”, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, page 207-210, 1993.
- [12] R. Agrawal and R. Srikant. “Fast, algorithms for mining association rules in large databases”, Proceedings of the 20th VLDB Conference Santiago, Chile, pp 487-499, 1994.
- [13] R. Srikant and R. Agrawal, “Mining Generalized Association Rules”, Proc. 21st VLDB Conference, Zurich , Swizerland.,1995.
- [14] Mohammed J. Zaki, “Parallel and Distributed Data Mining: An Introduction”, Large-Scale Parallel Data Mining Lecture Notes In Computer Science, Vol. 1759, 2000.
- [15] Qinghua Zou, Wesley Chu , Johnson, D. , Chiu, H. “A pattern decomposition (PD) algorithm for finding all frequent patterns in large datasets”, International Conference on Data Mining, ICDM 2001, Proceedings IEEE, 673 – 674, 2001.
- [16] Sotiris Kotsiantis, Dimitris Kanellopoulos , “Association Rules Mining: A Recent Overview”, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82
- [17] Pei-ji Wang, Lin Shi, Jin-niu Bai, Yu-lin Zhao ,”Mining Association Rules Based on Apriori Algorithm and Application”, International Forum on Computer Science-Technology and Applications, IFCSTA Dec. 2009,
- [18] C. Clifton and D. Marks. “Security and privacy implications of data mining. In Workshop on Data Mining and Knowledge Discovery”, pp 15–19, Montreal, Canada, University of British Columbia, Department of Computer Science, February 1996.
- [19] C. Clifton. “Protecting against data mining through samples”. In V. Atluri and J. Hale, editors, Proceedings of the 13th International Conference on Database Security (IFIP WG 11.3): Research Advances in Database and Information Systems Security, Vol. 171,pp 193–207,2002.
- [20] C. Clifton, M. Kantarcioglu, and J. Vaidya. “Defining Privacy For Data Mining”. In Proc. of the National Science Foundation Workshop on Next Generation Data Mining, pp 126-133, Baltimore, MD, USA, November 2002.
- [21] Oliveira and Zaïane Oliveira, S. R. M., Zaïane, O. R., “Towards Standardization in Privacy-Preserving Data”, Mining, Edmonton, 2004.
- [22] Vassilios S. Verykios¹, Elisa Bertino², Igor Nai Fovino², “State-of-the-art in Privacy Preserving Data Mining”, SIGMOD Record, Vol. 33, No. 1, March 2004.

- [23] Jaideep Vaidya, Chris Clifton, “Privacy-Preserving Data Mining: why, how and when”, IEEE Security & Privacy, , 2004
- [24] Verykios, V.S., Bertino, E., Nai Fovino, I., Parasiliti, L., Saygin, Y., and Theodoridis, Y. “State-of-the-art in privacy preserving data mining”, SIGMOD Record, SIGMOD Record, Vol. 33, No. 1, 50–57, March 2004.
- [25] Martin Meints and Jan Möller , “Privacy Preserving Data Mining: A Process Centric View from a European Perspective”,2008.
- [26] Elisa Bertino , Igor Nai Fovino Loredana Parasiliti Provenza ,”A Framework for Evaluating Privacy Preserving Data Mining Algorithms”, Data Mining and Knowledge sDiscovery, Vol.: 11, Issue: 2, pp 121–154, 2005.
<https://yumstories.com/index.php?board=63.0>

Chapter-2

Evaluation By Using Anonymization Method of Privacy Preservation Techniques With An Efficient Approach And A Systematic Comparison

Abstract

Data is ceaselessly being gathered because of the inescapability of consistently associated gadgets and the pervasiveness of Internet of Things advancements in individuals' lives. IoT gives the interconnection between different heterogeneous gadgets and sensors that can screen and accumulate a wide range of information about machines and human public activity. Notwithstanding the advantages that can emerge out of gathering information, clients are uncovering touchy and confidential data with perhaps deceitful elements. The tremendous measure of information being gathered about people has brought new difficulties in protecting their security when this information is scattered. Therefore, Protection Saving Information has turned into a functioning examination region, wherein numerous anonymization calculations have been proposed. Notwithstanding, given the huge number of calculations accessible and restricted data with respect to their presentation, it is challenging to distinguish and choose the most proper calculation given a specific distributing situation, particularly for specialists. Here, we play out a precise correlation of notable k-anonymization calculations to gauge their proficiency and their viability. These substances can process, take apart and mine data to remove important information, yet furthermore sell or possibly share the assembled data with pariahs, using it harmfully. With the creating number of maltreatment of data and data breaks, insurance has been another subject and serious security concerns have been animated. To resolve these issues, various Privacy-Preserving Mechanisms (PPMs) and instruments have been proposed.

Keywords : 1. Pervasiveness 2. Privacy preserving 3. Anonymization

1.0 Introduction:

Progress in logical exploration relies upon the accessibility and sharing of data and thoughts. Yet, the scientists are zeroing in on saving the protection of people. This issue prompts an arising research region, security safeguarding Enormous Information. Many creators proposed numerous methodologies[7] for security safeguarding Enormous Information for brought together as well concerning disseminated data set. Security safeguarding has started as a significant worry regarding the outcome of the Huge Information.

Security saving Huge Information (PPBD) manages safeguarding the protection of individual information or touchy information without forfeiting the utility of the information. Individuals have become very much aware of the security interruptions on their own information and are exceptionally hesitant to share their delicate data. This might prompt the incidental consequences of the Enormous Data[8]. Inside the limitations of security, a few techniques have been proposed yet this part of exploration is in its earliest stages. The outcome of protection saving Enormous Information calculations is estimated as far as its exhibition, information utility, level of vulnerability or protection from Huge Information calculations and so forth. Anyway no protection safeguarding calculation exists that beats all others on every conceivable model. Rather, a calculation might perform better compared to one more on one explicit rule. Thus, the point of this paper is to introduce flow situation of security safeguarding Enormous Information apparatuses and procedures and propose some future examination headings. Progresses in equipment technology[10] have expanded the ability to store and record individual information about buyers and people. This has caused worries that individual information might be utilized for different nosy or vindictive purposes.

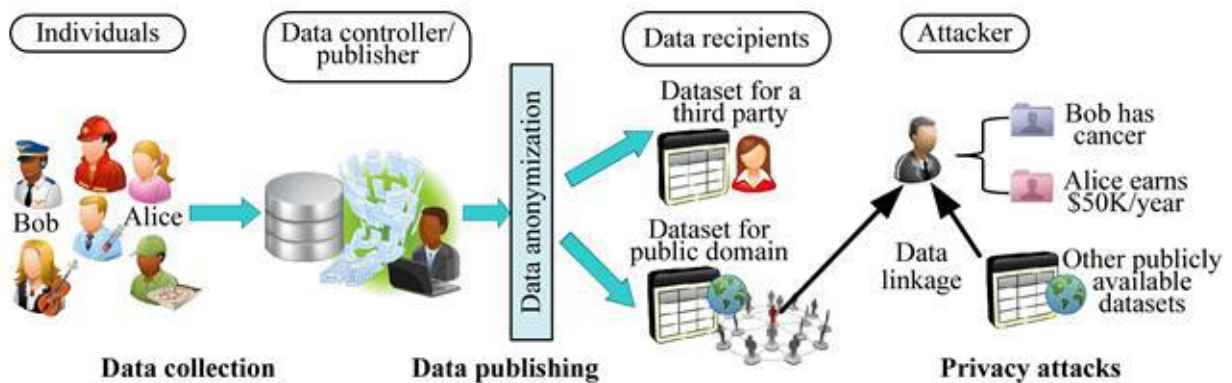


Figure 1: Overview of Privacy Preservation in Data Publishing

2.0 Methodology:

Despite the fact that PPMs mean to protect clients' security, this can come to the detriment of a debased utility of information. In this way, the choice of a PPMs ought to consider the clients' evenhanded as well as the compromise between the protection level and the utility of information, which are ordinarily application-explicit. Taking into account the heterogeneity of the gathered information, choosing and designing the legitimate PPMs is very difficult. To automatize this interaction and to give a coherent and methodical design of the fundamental parts and ideas of protection, a few instruments were created. These devices were

proposed to work with the design of PPMs and the examination of results. Nonetheless, choosing the legitimate PPMs as per the qualities of the information stays as a test.

To more readily comprehend how to distinguish PPMs as indicated by the information attributes, this overview presents a modern and exhaustive survey on heterogeneous data types and material PPMs. Recently, a couple of general outlines have focused in on PPMs for Huge Information and how they can gauge up to the extent that cultivated security level, data utility, unpredictability, or possibly application fields. Other more unambiguous surveys look at PPMs for a specific data type or a breaking point social event of data types, also utilization of PPMs for express spaces. Our outline shifts from past composition by proposing a security logical order for heterogeneous data types that spreads out an association between different data types and PPMs.

In this survey, PPMs are gathered by the overall arrangements of data they can be applied to (coordinated, semi-coordinated and unstructured), as well as their sensibility for consistent or detached application. The essential responsibility of this review is the detail of a logical grouping of data types for each characterization of data that is sensible for the ID of looking at PPMs, to allow the peruser to suitably grasp the secret guidelines of the addressed PPMs and their substantiality to the data types in the logical order inside. This concentrate further contributes by presenting and standing out existing security gadgets from concession with the data types and PPMs made available, likewise the insurance and utility appraisal features of such tools[9].

2.1 Privacy Preserving in Big Data:

Models and Calculations are intended for analysts, teachers, and high level understudies in software engineering. The fundamental point is, to foster proficient techniques to find security saving affiliation rule mining in concentrated as well as in appropriated data set climate without abusing any protection imperatives. The issue of protection requirements for brought together information base climate is altogether not quite the same as appropriated data set climate. So the approaches for privacy preserving association rule mining in centralized are also different from approaches in distributed environment. Achieving privacy preserving goals in both these environments in the process of mining is a challenging task.

2.2 Algorithms and techniques :

Algorithms and strategies proposes various procedures to play out the Large Information undertakings in a security protecting way. These strategies by and large fall into

the accompanying classes: information change procedures, cryptographic strategies and conventions for information sharing, factual procedures for revelation and derivation control, question examining strategies, randomization and irritation based methods. This altered volume additionally contains studies by recognized scientists in the security field. Each overview incorporates the key examination content as well as future exploration bearings of a specific point in protection.

The objective of accomplishing security in concentrated data set climate is, to get a mutilated data set which conceals the delicate thing sets. While mining task is performed on contorted information base every one of the delicate standards ought to be concealed with no aftereffects. In conveyed climate, the objective is to find protection safeguarding affiliation rule mining without uncovering any people private information or data when the data set is disseminated among n-number of locales. Three strategies are proposed in light of heuristic and careful methodology connected with concentrated data set climate to find protection saving affiliation rule mining by concealing delicate thing sets with least aftereffects.

3.0 Literature Review :

Fundamental point of Privacy protection in information mining is to discover such arrangement which will limit the danger of abuse of the information. There are number of viable strategies for security safeguarding information mining which have been proposed by various authors. Here for this situation it is essential to keep up the advantages of protection safeguarding even after the changed dataset is made accessible for mining. Characterized such strategies are as per the following.

3.1.1 Randomization:

This is one of the well known methods in protection safeguarding information mining examines. By adding commotion to the first information, the estimations of the records are made. In this strategy the individual estimations of the records can never again be recuperated as clamor added to unique information is enormous enough to keep up the protection. Randomization procedures accomplish both, protection conservation and information disclosure with the assistance of arbitrary clamor based annoyance and Randomized Response conspire. In spite of the fact that this procedure causes high data misfortune, it is a progressively effective strategy.

3.1.2 Anonymization:

Anonymization utilizing speculation and concealment anonymization method[11], makes undefined records among gathering of records. k-anonymity is known as delegate

anonymization system. To distinguish records exceptionally it considers semi identifiers which can be utilized related to open records. There are such a large number of procedures which has been proposed, for example, k-anonymity, L-diversity, t-closeness, Personalized secrecy. Here anonymization strategy brings about loss of data somewhat yet guarantees the creativity of information.

The aim of the PPBD algorithms is the extraction of relevant knowledge from large amount of data, while protecting sensitive data or information. The several existing Big Data techniques, incorporating privacy protection mechanisms such as association rule mining, classification and clustering techniques are discussed in. An important aspect is discussed in determining suitable algorithms for various Big Data techniques[5,6] to protect sensitive data or information by doing modifications to the original database before releasing it to the intended parties and they also presented comprehensive set of criteria with respect to existing PPBD algorithms which helps the designer to determine which algorithm meets specific requirements. The authors in, proposed classification of PPBD techniques based on different dimensions such as data distribution, data modification,

Big Data algorithm, data or rule hiding, privacy preservation. They also discussed various methods exist in each classification of methodology[10] based on the dimension. The existing methodologies are discussed for different privacy preserving Big Data techniques such as classification, association rule mining and clustering in various dimensions. They evaluated the algorithms related to heuristic-based techniques, cryptography-based techniques, and reconstruction-based techniques for different Big Data techniques. The problem of protecting sensitive knowledge in large databases is addressed. The authors in [14] addressed PPBD technique and presented national security applications where privacy is the main concern.

They viewed the privacy problem as a form of inference problem and introduced the notion of privacy constraints. They described an approach for privacy constraint processing. The authors surveyed the current state of the art in Statistical Disclosure Control methods for protecting individual data (micro data) [15]. A classification of micro data protection methods such as perturbative masking methods, nonperturbative masking methods and synthetic micro data generation are presented. The authors in [13] emphasized the important aspects such as identification of suitable evaluation criteria and the development of related benchmarks required in the design of privacy preserving Big Data algorithms. In this article, they also discussed issues related to recent research in the privacy preserving Big Data to balance the trade-off between the right to privacy and the need of knowledge discovery.

4.0 Algorithms Evaluated:

The goal of PPBD methods is to change the information by making it less unambiguous, so that the people's security is safeguarded; while planning to hold the convenience of the anonymized information. The substance of PPBD is to produce datasets that have great utility for various errands, as usually, all the potential usage situations for the information are obscure at the hour of distribution. For instance, under open information drives [2] and [1], distinguishing every one of the information recipients is unimaginable. In this way, any information regulator engaged with the sharing - save in mechanisms [3].

Be that as it may, this is definitely not a minor undertaking, as specialists are not really specialists in that frame of mind of information protection. In addition, it is much of the time the case that no techniques exist that guarantees that anonymization is directed successfully in an association. This can lead professionals to utilize basic techniques for de-ID (e.g., eliminating all immediate identifiers, for example, names and government backed retirement numbers), before delivering information.

Regardless, it has been demonstrated that this approach alone isn't sufficient to safeguard protection. This issue happens in light of the fact that it is as yet conceivable to consolidate different datasets or have foundation information about people, to make surmisings about their character. The re-distinguishing proof of an individual is accomplished by connecting credits, referred to as semi identifiers (QIDs), such as orientation, date of birth or PIN code.

As result, various anonymization calculations have being proposed in the space of PPBD. In any case, the determination of the most suitable calculation for a given distributing situation is trying for experts: Not just there is a plenty of anonymization calculations from which one can pick, however every recently presented calculation guarantees a specific predominance over the others. The first trial assessments of these calculations are typically restricted, as they are for the most part centered around showing the advantages of the proposed calculation over a portion of the recently proposed ones. This present circumstance of ten limits the extent of their assessment concerning the exploratory setups utilized (e.g., utilizing a solitary examination metric, excluding scenarios). Additionally, in the situations where the creators present another measurement, the measurement will in general incline toward the proposed calculation because of the specific perspectives estimated by the measurement.

The previously mentioned circumstances could confound experts, driving them

to erroneously expect to be that assuming a calculation out performs others for a specific measurement, this calculation can be viewed as the best no matter what the information boundaries. Be that as it may, this conduct isn't ensured, as the presentation of a calculation can shift when over again input informational collection with various qualities is anonymized, or the arrangement used to test the calculation changes. Taking into account these difficulties, we accept that there is serious areas of strength for a to broaden the current assessments of these anonymization calculations to cover an additional comprehensives to trial designs. As a result, the goal of this work is to give specialists more nitty gritty clarifications of the explanations for the presentation varieties of the calculations. Some model situations are found in[12], and include: a clinic giving data about tolerant confirmations, a school sharing understudy instruction information, a retailer sharing information about clients, and so on. In our review, we center around k-anonymity [4]because it is a major guideline of security and, as opposed to different models which are too prohibitive to ever be viable or challenging to be accomplished for certain situations, its reasonable effortlessness has made it broadly examined and embraced in various spaces, for example, medical care, information mining and factual exposure control. The k-namelessness model is a premise to propose safer models and significance in security conservation. Recently acquainted models of upgrades with k-anonymity[5], so they can't remain solitary and should be joined by k-obscurity; hence, they can't supplant k-namelessness. Moreover, the improvement of new calculations in light of k-obscurity is as yet embraced by analysts. Essentially, k-secrecy fills in as base for new anonymization methods on various settings.

Generalization and Suppression: Suppression comprises in supplanting a portion of the first information with an extraordinary worth (e.g., "*") to demonstrate that this information isn't uncovered. Speculation (likewise called recoding) comprises in supplanting the upsides of a characteristic with less unambiguous yet predictable qualities; frequently utilizing a worth speculation order (VGH). The qualities at the least level (right) are in the ground space of the property, which relate to the most unambiguous qualities (unique qualities). The most elevated level (left) showing the"*"value, compares to the greatest speculation or full concealment of the worth. Recoding can be acted in a worldwide (full-space speculation) or nearby plan. Nearby recoding can apply various guidelines to similar occurrences of properties, so that a few examples stay with the particular qualities, while others are summed up. Going against the norm, worldwide

recoding comprises in applying similar speculation to all cases of a characteristic, to such an extent that all values are summed up to a similar level of the VGH. Worldwide recoding is additionally ordered in two sorts: single-layered, which treats each quality in the QID bunch freely; and complex, which recodes a space of n-vectors that are the cross result of the spaces of the individual QID credits

4.1 Challenges:

The Ultimatum is to limit the risk of re-identification by anonymizing big data in cloud computing environment. One of the major challenging tasks is to manage Big Data which is large and complex so it is difficult to store, analyze, capture, search, share, transfer, visualize information privacy. Energy performance and system size is a challenging problem to tackle in cloud computing[6]. Issues of low trust on cloud computing is an obstacle, in case of critical storage. In our similar review, we have chosen three k-anonymization calculations utilizing speculation and concealment. We have picked these in view of the accompanying reasons: (1) these calculations have been broadly referred to in the writing, (2) these calculations utilize various procedures of anonymization permitting a more thorough assessment, (3) a public execution of these calculations is accessible and (4) these calculations can be assessed inside a similar structure, considering an all the more fair correlation. In the accompanying segment, we depict the calculations applicable to the extent of this work.

4.2 Datafly:

Data fly is a greedy heuristic algorithm that performs single – dimensional full-domain generalization. It counts the frequency over the QID set and if k- anonymity is not yet satisfied, it generalizes the attribute having the most distinct values until k-anonymity is satisfied. Whereas this algorithm guarantees a k-anonymous transformation, it does not provide the minimal generalization.

4.3 Incognito:

In-cognito is a solitary layered full-space speculation calculation that forms a speculation cross section and navigate by utilizing a base up broadness first pursuit. The quantity of substantial speculations for a property is characterized by the profundity of its VGH. This would be: Two for conjugal status(M) and age(A); four for PIN code(P). An illustration of the speculation cross section made for that QID set should be visible in Table. Every hub in the grid addresses an anonymization arrangement. For instance, the hub <M1,A1,P1> implies that the three QIDs have been summed up once (one step up in their VGH), which is the

arrangement displayed in Table. To anonymize information, In secret purposes prescient labeling to lessen the pursuit space. This truly intends that, while crossing the grid, assuming a tribute is found to fulfill k-obscurity, its immediate speculations can be all pruned as it is ensured that they likewise fulfill k-secrecy. Dissimilar to Information fly, Undercover produces an ideal arrangement. This implies that the anonymized arrangement contains the maximal amount of data as per picked data metric.

4.4 Mondrian:

Mondrian is an insatiable calculation that parcels the area space recursively into multi locales, every one of which contains essentially k records. It begins with the most un-explicit (most generalized)value of the characteristics in the QID set and practices as segments are performed on the information. To pick the aspect (i.e., property) on which to play out the segment, Mondrian utilizes the quality with the largest (standardized) scope of values. Assuming numerous aspects have a similar width, the first that empowers a reasonable cut (i.e., the cut doesn't cause an infringement of k-secrecy) is chosen. When the aspect is chosen, Mondrian utilizes a middle parceling way to deal with pick the split worth, the worth at which the segment will be performed. To find the middle for a characteristic, a recurrence sets approach is utilized. This is, the information is examined including the frequencies for every one of the special qualities in the property until the middle position is found. The worth at which the middle is found turns into the split worth.

5.0 Performance and Comparison of various Anonymization algorithms:-

Playing out a fair correlation of anonymization calculations is intrinsically a difficult errand, as each proposed calculation utilizes various measurements and settings. The exhibition of the calculations could change among various blends of datasets and input boundaries (e.g.,an calculation might function admirably in a few exploratory designs and perform ineffectively in others).As an outcome, it is vital to survey the calculations by characterizing a typical setup that mirrors the boundaries utilized in existing assessments. Moreover, a correlation requires the utilization of rules that can be generally material to gauge various parts of the calculations (e.g., effectiveness and information utility). In PPBD all the potential usage scenarios are often un known.

5.1 Efficiency:

A calculation ought to be assessed by the assets expected to do the anonymization. This is a significant aspect, as the anonymization cycle can be escalated in asset utilization. At the point when assets are restricted, they address an imperative in the determination of a calculation. In any event, when a calculation accomplishes a decent degree of utility in the anonymized information, on the off chance that it isn't productive as far as memory utilization, computer chip use or execution time, then, at that point, it probably won't be down to earth for use. To gauge the execution time, one could screen the slipped by time for the various phases of the anonymization cycle: The information transfer, the actual anonymization and the information yield. As the information transfer and result steps don't change among calculations, we will zero in on estimating the anonymization time as it were. To break down the exhibition of the calculations concerning anonymization time, we look at the connection between the anonymization time and the expense of the accompanying practical qualities of the calculations: for Information fly, the quantity of speculation activities performed; for Undercover, the quantity of the hubs or speculation states assessed in the made cross section; and for Mondrian, the quantity of parcels performed on the information.

Original Table (To)

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
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
:	:	:	:	:
P05000	Rangvi	43	634829	Malaria

Disease Master Table (T_{DM})

DCode	Disease
0001	Allergie
0002	Asthma.
0003	Cancer
0004	Chest Pain
0005	Head Ache
0006	Kidney Problem
0007	ILD
0008	Dengue Fever
0009	Paralysis
0010	Arthritis
0011	Viral Fever
0012	Heart Defect
0013	BP
0014	Dengue fever
0015	Malaria
0016	Jaundice
0017	Hepatitis
0018	Influenza
0019	Malnutrition
0020	Pneumonia
0021	Obesity and Genetics
0022	Sinus
0023	Cancer
0024	Chest pain
0025	Orthritis

.....
0300	Diabetes

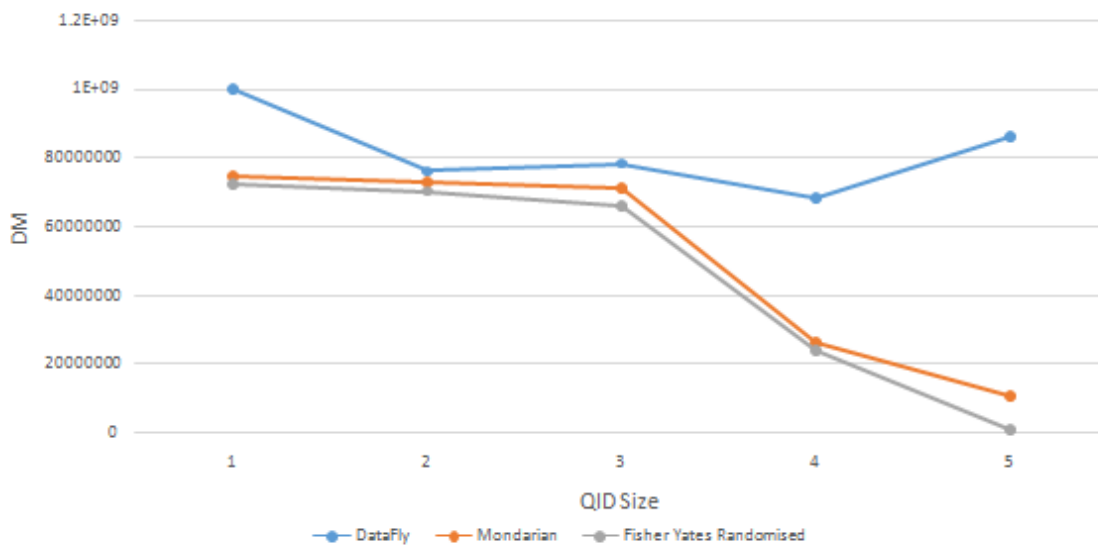
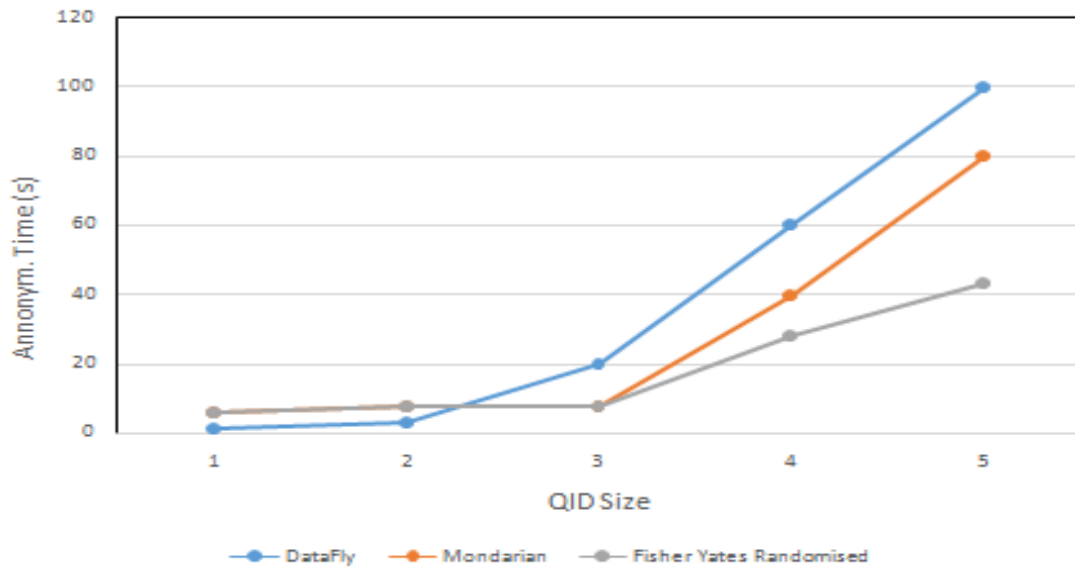
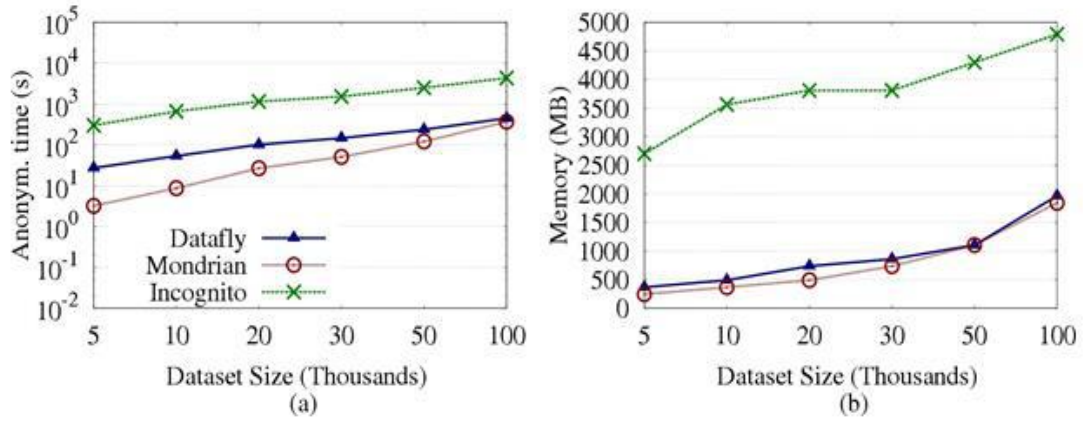
Modified Table (T_{Mod})

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

6.0 Results:



7.0 Conclusion:

We have assessed a deliberate assessment, concerning reasonableness, effectiveness, throughput and information utility, of three of the k-anonymization calculations. Utilizing freely accessible executions of the calculations under a typical casing work, we distinguish the situations where the calculations performed well or inadequately, regarding the measurement of interest and give inside and out conversation of the purposes for these ways of behaving. The outcomes exhibited that there is no best anonymization calculation for all situations, yet the best performing calculation in a given circumstance is impacted by various variables. In view of our examination, we gave bits of knowledge about the elements to consider while choosing an anonymization calculation and examined the highlights of a bunch of broadly useful utility measurements. In addition, we rouse the significance of thinking about the requirements of professionals (who may not be information security specialists), to offer rules that work with them with the reception of protection saving strategies. Basic investigation, strength and shortcoming of the calculations are inspected.

8.0 References:

- [1] CentralStatisticsOfficeDatabases.<http://www.cso.ie/en/databases/>.
- [2] UTDAAnonymizationToolBox.<http://cs.utdallas.edu/dspl/cgi-bin/toolbox/>.
- [3] C. C. Aggarwal. On k-Anonymity and the Curse of Dimensionality. In Proceedings of the 31stInternational Conference on Very Large Data Bases, VLDB '05, pages 901–909. VLDB Endowment,2005.
- [4] M.R.S.Aghdam and N.Sonehara. On Enhancing Data Utilityin k-anonymization for Data without Hierarchical Taxonomies. International Journal of Cyber-Security and Digital Forensics,2(2):12–22,2013.
- [5] D. Agrawal and C. C. Aggarwal. On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS'01,pages247–255,2001.
- [6] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. Synthetic Data Generation using Benefactor Tool. Technical report, University College Dublin,UCD-CSI-2013-03,2013.
- [7] K. S. Babu, N. Reddy, N. Kumar, M. Elliot, and S. K. Jena. Achieving k-anonymity Using Im-proved Greedy Heuristics for Very Large Relational Databases. Transactions on Data Privacy,6(1):1–17,2013.

- [8] K. Bache and M. Lichman. UCIMachineLearningRepository, 2013.
- [9] M. Barbaro and T. Zeller. A Face Is Exposed for AOL Searcher No. 4417749, 2006.
- [10] R. J. Bayardo and R. Agrawal. Data Privacy Through Optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering, ICDE '05, pages 217–228, 2005.
- [11] V. Bergmann. DataBenerator Tool. <http://databene.org/databene-benerator/>.
- [12] E. Bertino, I. N. Fovino, and L. P. Provenza. A Framework for Evaluating Privacy preserving Data Mining Algorithms. Data Mining and Knowledge Discovery, 11(2):121–154, 2005.
- [13] S. M. Blackburn, P. Cheng, and K. S. McKinley. Myths and Realities: The Performance Impact of Garbage Collection. SIGMETRICS Performance Evaluation Review, 32(1):25–36, 2004.
- [14] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-Preserving Data Publishing. Foundations and Trends in Databases, 2(1–2):1–167, 2009.
- [15] Open Data websites. <http://www.data.gov/>, <http://data.gov.uk/>.

Chapter-3

Secure data deduplication system for integrated cloud and seamless connectivity with Privacy and Security Awareness

Abstract :

The paper proposes a new method for implementing data deduplication (DD) over encrypted data in integrated cloud-fog storage and computing environments. The proposed technique aims to reduce data redundancy and ensure the security of the data by using Convergent Encryption (CE) and Modified Elliptic Curve Cryptography (MECC) algorithms. The proposed method encrypts the file using CE and then re-encrypts it using MECC, which provides a robust encryption approach to ensure the security of the data. The method also recognizes data redundancy at the block level, which helps to reduce the redundancy of data more effectively with Privacy and Security Awareness. The testing results show that the proposed approach can outperform some state-of-the-art methods in terms of computational efficiency and security levels. This is an important contribution, as efficient and secure data deduplication techniques are essential for optimizing storage in integrated cloud-fog storage and computing environments. The proposed method appears to be a promising approach for implementing secure data deduplication in integrated cloud-fog storage and computing environments. By combining CE and MECC algorithms, the method can effectively reduce data redundancy while ensuring the security of the data. However, further testing and evaluation may be required to validate the effectiveness and performance of the proposed method in real-world scenarios.

Keywords:

Convergent encryption (CE), Modified elliptic curve cryptography (MECC), Edge computing, Integrated cloud and fog networks, Hash tree. Secure hash algorithm (SHA)

1.0 Introduction:

Data deduplication is a technique used to eliminate duplicate copies of data, which can reduce storage requirements and improve overall system efficiency. In integrated cloud-edge networks, where data is shared between multiple nodes, a secure data deduplication system can be highly beneficial. However, such a system must also ensure the security and privacy of the data. One approach to building a secure data deduplication system for integrated cloud-edge networks is to use a hybrid architecture that combines the benefits of both centralized and

decentralized systems. In this architecture, data is stored in a centralized location, such as a cloud server, but the deduplication process is performed locally on edge devices. This approach reduces the amount of data that needs to be transferred over the network, while also keeping the data secure and private. To ensure the security of the data, the system can use encryption techniques to protect the data during transmission and storage. This can be achieved through the use of secure protocols such as HTTPS or SSL/TLS. The system can also use secure hashing algorithms to detect and eliminate duplicate data, while maintaining the integrity of the original data. Another important aspect of a secure data deduplication system is the authentication and access control mechanisms. Access to the data must be restricted only to authorized users, and the system must be able to track and log all access attempts. This can be achieved through the use of access control lists and user authentication protocols such as OAuth or OpenID Connect. Overall, a secure data deduplication system for integrated cloud-edge networks can provide significant benefits in terms of storage efficiency, system performance, and data security. By using a hybrid architecture and incorporating encryption and access control mechanisms, the system can ensure the privacy and integrity of the data while improving overall system efficiency.

This paper discusses the need for cloud computing, fog networks, and edge computing to process and store large amounts of data generated by the Internet of Things (IoT) and other applications. The advantages and disadvantages of each of these technologies are discussed, including issues related to latency, energy efficiency, reliability, security, and privacy. The passage also describes the importance of data deduplication in optimizing storage and security and proposes a secure data deduplication system that uses convergent and MECC algorithms over the integrated cloud-fog-edge environment. The limitations of convergent encryption are discussed, and a solution is proposed to address the possibility of dictionary attacks. The passage highlights the need for efficient resource allocation methods to improve performance and reliability in cloud and fog environments.

The significant contributions of the paper presents a new method for constructing a secure deduplication system that utilizes Convergent and Modified Elliptic Curve Cryptography algorithms in both cloud and fog/edge environments. The paper also evaluates the proposed technique's performance in terms of computational efficiency and security level. Additionally, the paper validates the proposed deduplication technique's ability to identify data dedundancy at the block level, which can effectively reduce data redundancy and minimize storage space in the cloud environment.

2.0 Related works:

- Make sure to have a Business Associate Agreement (BAA) in place. In **compliance with HIPAA**, all third parties or business associates are required to provide in writing that they will safeguard the information.
- Use **least privileged access** for business associate access rights, so they are only accessing information that's absolutely critical to their business.
- Implement **multi-factor authentication** to quickly and efficiently authenticate user access.
- Conduct due diligence required by HIPAA, such as documentation and **monitoring** of business associate activity and risk assessments.

Thankfully, there are **solutions** that can help streamline these processes. **Remote access tools** built for healthcare organizations can standardize and restrict access while also auditing business associate activity so IT teams aren't bogged down by access requests and gathering documentation. These systems also give healthcare organizations more peace of mind about the "who, what, when, why, and how" of business associates

Computationally intensive, which may impact their efficiency. As a result, researchers have proposed several optimization techniques to improve the efficiency of public-key-based deduplication schemes. For instance, Chen et al. proposed a scheme called S-PDP, which is a secure and efficient privacy-preserving data deduplication scheme based on public-key cryptography. The scheme utilizes a novel hash-based approach to ensure data privacy and authenticity while reducing the computational overhead of signature verification. Similarly, Wang et al. proposed a scheme called PIR-TS, which uses homomorphic encryption and a probabilistic index to achieve efficient privacy-preserving data deduplication.

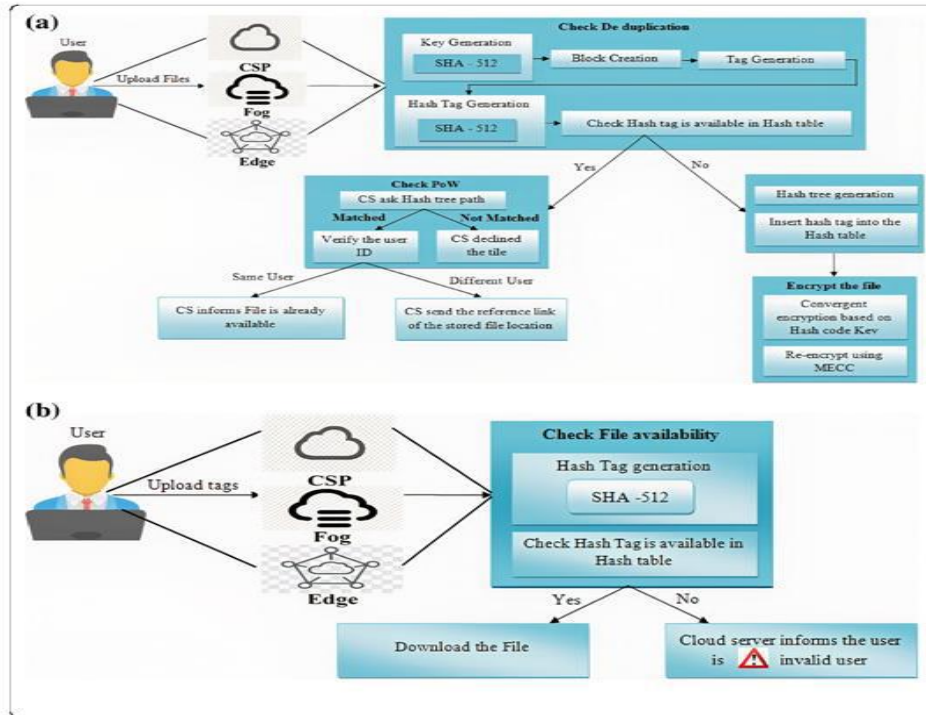
In addition to these optimization techniques, researchers have also explored the use of emerging technologies, such as blockchain, to enhance the security and privacy of deduplication systems. For instance, Wang et al. proposed a blockchain-based deduplication scheme that enables secure and efficient data sharing among multiple parties. The scheme utilizes smart contracts and decentralized consensus mechanisms to ensure the integrity and authenticity of the reduplicated data.

The field of secure deduplication is a rapidly evolving area of research, with many innovative solutions being proposed to address the security and privacy challenges associated with data deduplication in cloud and remote storage environments. As the volume and complexity of data continue to grow, it is likely that we will see further advances in this area,

with new techniques and technologies being developed to ensure the secure and efficient management of data in the cloud.

3.0 Proposed secure deduplication approach:

The paper proposes a secure data deduplication system for cloud computing environments that utilizes convergent and MECC algorithms. The system aims to overcome the security challenges associated with data deduplication, which is a technique used to optimize storage space in cloud environments. The proposed methodology is analyzed in four scenarios. The first scenario is when a new user tries to upload a new file. In this scenario, the system generates a unique identifier for the file using a convergent encryption algorithm. This identifier is used to check if the file already exists in the cloud storage, and if it does, the system avoids uploading the file again. The second scenario is when the same user tries to upload the same file. In this case, the system uses the same identifier generated in the first scenario to determine if the file already exists in the cloud storage. If the file exists, the system avoids uploading it again. The third scenario is when different users try to upload the same file to the cloud server. In this scenario, the system uses a multi-authority MECC algorithm to securely compare the identifiers generated by different users. If the identifiers match, the system avoids uploading the file again. Finally, the fourth scenario is when users try to download the file. In this case, the system uses the identifier to locate the file in the cloud storage and allows the user to download it securely. The proposed methodology is illustrated using a block diagram in Fig. 1a and b. Overall, the system provides a secure and efficient way of managing data in cloud environments using data deduplication.



Convergent encryption (CE) is a method used to encrypt data in data deduplication, which helps to improve data confidentiality. In this method, a convergence key (CK) is generated using the hash code (HC) value of the file. This CK is then used to encrypt all the blocks of data copy. The purpose of using a CK is to make sure that the same file with the same content is always encrypted in the same way. This allows the system to detect duplicate files easily, as the same tags will be provided for the same files.

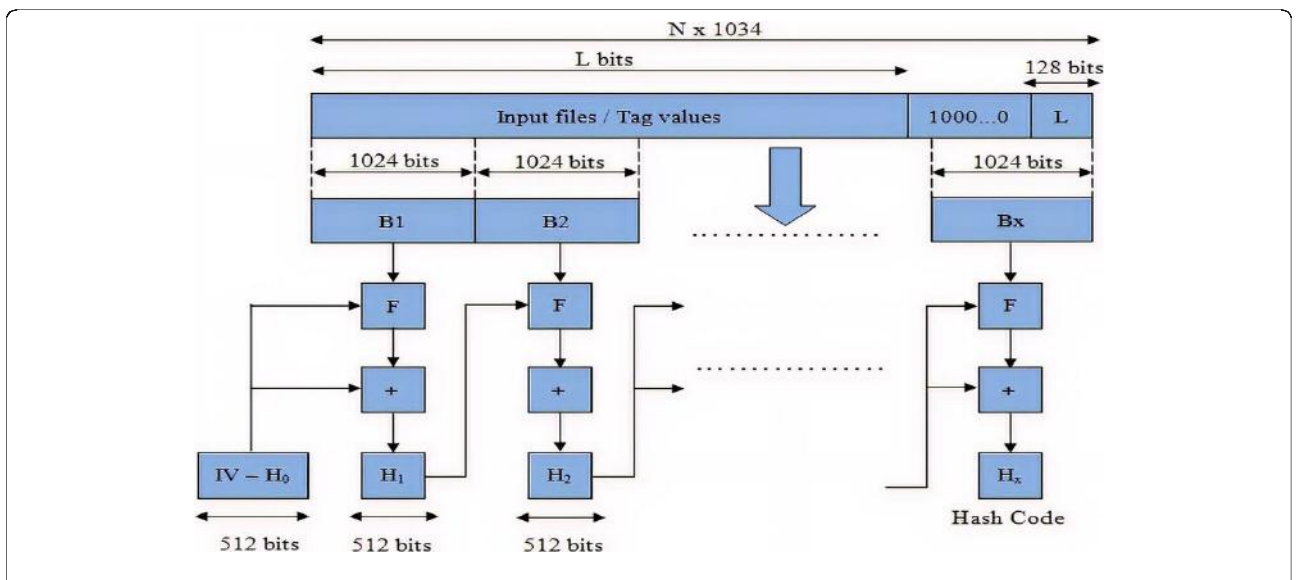


Fig 2 : Structure of SHA 512 Algorithm

The process of convergent encryption involves several phases. First, the original file is divided into blocks of fixed size. Each block is then encrypted using the CK to create an

encrypted data block. Next, a tag is derived for each encrypted data block using the same hash function used to generate the HC. This tag is used to identify duplicate data blocks in the CSP. If two data blocks have the same content, they will have the same tag, making it easy to detect duplicate data. The encrypted data block, along with its tag, is saved in the CSP. This ensures that the data is encrypted and secured before it is stored in the cloud, and the tags allow for efficient detection of duplicate data blocks.

It seems like you have shared a technical document describing a proposed methodology for secure and efficient data deduplication in cloud storage using convergent encryption and elliptic curve cryptography. The document describes the encryption and decryption process using the CE algorithm and highlights its limitation of being vulnerable to dictionary attacks. To overcome this, the proposed methodology utilizes the Modified ECC algorithm for additional encryption.

Convergent encryption

During encryption, the original file (f) and κ_h are given as input for the encryption algorithm and E_y is the encryption function. Finally, this encryption algorithm gives ciphertext(C_t) as output.

$$C_t = E_y(f; \kappa_h) \quad (1)$$

Convergent decryption

During decryption, the encrypted file f or C_t is inputted to the decryption algorithm. Finally, this decryption algorithm outputs f and C_t .

$$f = D_y(C_t; \kappa_h) \quad (2)$$

This function is utilized as a maximum limit. ECC is a kind of algorithm that is used in the implementation of public-key cryptography. The mathematical model of the ECC with g and e as integers is given below.

$$w^2 = v^3 + (e; 4g^3)27e^2 \neq 0 \quad (3)$$

The document also discusses the key generation process and the generation of public, private, and secret keys. It further explains the process of checking for duplicate data copies using hash values and hash trees. If the same user tries to upload the same file, the CS calculates the hash value and checks for the hashtag value in the HT. If it's available, the CS queries the path of the hash tree to verify the user id and avoids storing the file again. If different users try to upload the same file, the CS splits the file into blocks, generates a tag for each block, and creates a hashtag value to check for duplicates. The CS then sends a reference link to the user if the id is different.

The main step is to generate the public key (α_k) from the server and encrypting it. In the next step, a private key (β_k) is produced on the server-side, and the message is decrypted. The last step is to generate a secret key (o_k) from α, β and point on the curve (ρ). using succeeding equation, the α_k is generated,

$$\alpha_k = \frac{1}{4} \beta_k \omega \rho_c \quad (4)$$

The equation (5) elucidates O_k generation,

$$O_k = \frac{1}{4} \alpha_k \omega \beta \omega \rho \quad (5)$$

After o_k generation, the file is encrypted. This encrypted file contains two CTs, and mathematically, they are depicted as,

$$C_1 = \frac{1}{4} (f; K = 1; 2; \dots \dots; (n - 1)) \omega \rho_c O_k \quad (6)$$

$$C_2 = \frac{1}{4} (f; K = 1; 2; \dots \dots; (n - 1)) \omega \alpha_k O_k \quad (7)$$

Here, C_1 and C_2 represents the two CTs, K is the random number generated in $(1, \dots, (n - 1))$ interval. During encryption, o_k is added to the CTs. During decryption, o_k is subtracted with the two CTs, and the original file f is given by,

$$f = \frac{1}{4} C_2 - \beta_k \omega C_1 - O_k \quad (8)$$

Overall, the document proposes an interesting methodology for secure and efficient data deduplication in cloud storage using convergent encryption and elliptic curve cryptography. The proposed secure deduplication system is using the SHA512 algorithm to generate hashtag values for each block of data in a file. The hashtag value is checked against a hash table (HT) to determine if the block is a duplicate or not. If the block is a duplicate, the user is allowed to download the file, otherwise, the user is considered an invalid user.

After the OK generation, the file is encrypted using two ciphertexts (CT1 and CT2), as shown in Equation 6. When multiple users try to upload the same file, the file is split into blocks, and each block is assigned a hashtag value. The hashtag value is converted into a hash code (HC), and the HC is checked against the HT to determine if the block is a duplicate or not.

When a user tries to download a file, the user sends the tag value of the specified file, and the CS generates a hash value for the tag. The generated hash value is then checked against the HT to determine if the file is available for download. If the hash value matches an entry in the HT, the user is allowed to download the file, otherwise, the user is considered an invalid user.

The proposed approach considered the file block-sizes of 5 MB, 10 MB, 15 MB, 20 MB, and 25 MB. Then, the tag key is created for each of the divided blocks. Next, the hash

value is computed for all the tag keys utilizing the same SHA-512 algorithm.

In the uploading phase, the CS checks the hashtag (HT) for a particular input file. If the hashtag value of the input file is in that HT, then the CS queries the path of the hash tree to the users. If a user sends the correct path, then the CS verifies the user id. If the id is the same, then the CS does not store the file again. Generally, the hash tree path has the succeeding format,

$$P(H_t)1/4fRLL; RLR; etc: g \quad (9)$$

Where, $P(H_t)$ denotes the path of the hash tree, RLL represents the “Root, Left, Left”, RLR , denotes the “Root, Left, Right.” The leaf node is not added to the hashed tree path. The same user trying to upload the same file is mathematically denoted as,

$$\rightarrow D_t^{Uploads} \rightarrow S_t CS^{asks} P(H) \quad (10)$$

$$\rightarrow P(H_t) \rightarrow CS^{Send} R(f) \quad (11)$$

$$\rightarrow P(H_t)_{Not\ matched} \rightarrow CS^{Informed} \rightarrow I_u \quad (12)$$

whether it is in the HT. If the value is available, then the CS lets the user download the file, else the CS considers them as an invalid user. It is mathematically denoted as

$$H(T)_{matched} \rightarrow D_u \downarrow \quad (13)$$

$$H(T)_{Notmatched} \rightarrow I_u \quad (14)$$

To evaluate the encryption time of the proposed system, the starting and ending time of the encryption process is recorded, and the difference between the two times is computed. The encryption time is an important metric as it indicates the efficiency of the encryption algorithm in converting the plaintext data into ciphertext.

Algorithm 1: Uploading file into the Cloud Server

Input: Original file

Output : upload the file into the CS

Begin

Initialize key k, tagT, hashtagH(T),

Blocks(B1,B2,.....Bn) and hash tree Ht

For n number of files do

{

Generate k using SHA -512

```

Divide f into blocks (B1,B2,.....Bn)
Generate tag T
Generate H(T) using SHA -512
If (H(T) == Hash_table) then
Check PoW
Else
    generateHt and insert H(T) into the hash
table and encrypt the file f
store the file in Cloud Server(CS)
end if
}
End for
End.

```

Algorithm 2 : Downloading file from the Cloud Server

```

Input : Tag value
Output : Download the original file
Begin
    Initialize tag T, hashtaf H(T) original file f
For all tags do
{
Generate H(T) using SHA – 512
If(H(T) == Hashtable) then
Cloud Server allows the user to download – the
File f(↓)
Else
Cloud server informs as invalid user
End if
}
End for
End

```

4.0 Results & Discussion:

Performance analysis of proposed decryption technique Decryption time D_t is considered as the time that a decryption algorithm consumes to transform the encrypted data back to the original data. Decryption time is computed as the difference between the decryption ending time and decryption starting time. To evaluate the decryption time of the proposed system, the starting and ending time of the decryption process is recorded, and the difference between the two times is computed. The decryption time is an important metric as it indicates the efficiency of the decryption algorithm in converting the ciphertext data into plaintext.

Performance analysis of key generation Key generation time is the time taken to generate the public and private keys used in the encryption and decryption process. It is evaluated as,

To evaluate the key generation time of the proposed system, the starting and ending time of the key generation process is recorded, and the difference between the two times is computed. The key generation time is an important metric as it indicates the efficiency of the key generation algorithm in generating the required keys. Security analysis Security analysis is the process of evaluating the security of a system by testing it against various attacks. The proposed system is evaluated against various security attacks to ensure its robustness and reliability.

4.1 Performance Analysis:

The performance analysis of the proposed system is compared with existing security algorithms such as Diffie-Hellman (DH), ECC and Rivest Shamir Adelman (RSA) to ensure that the proposed system is efficient and performs better than the existing algorithms. The paper presents two proposed techniques, namely the Modified Elliptic Curve Cryptography (MECC) and the deduplication scheme. The performance of these techniques is compared with existing methods, such as DH, RSA, ECC, CRT, and SDM. The results indicate that the proposed MECC approach provides better performance in terms of encryption and decryption time, key generation time, and security level. Similarly, the proposed deduplication scheme outperforms existing techniques concerning the deduplication rate and tree generation time. Overall, the paper suggests that the proposed MECC and deduplication techniques can be effectively used in secure cloud storage systems to enhance data security and reduce storage overheads.

4.2 Performance analysis of proposed encryption technique:

Encryption time

E_t is considered as the time that an encryption algorithm consumes to generate encrypted data as of the inputted data. Encryption time is computed as the difference between the encryption ending time and encryption starting time. It is evaluated as, Table 1 Performance comparison of proposed MECC and Existing Techniques in terms of Encryption time

File size in MB	Encryption Time(sec)			
	Proposed MECC	DH	Existing EC	Existing RSA
5	6.21	10.22	14.47	12.44
10	10.04	19.64	22.56	21.76
15	15.54	28.32	28.00	30.35
20	21.0	36.33	36.65	35.61
25	27.55	46.29	44.32	47.28

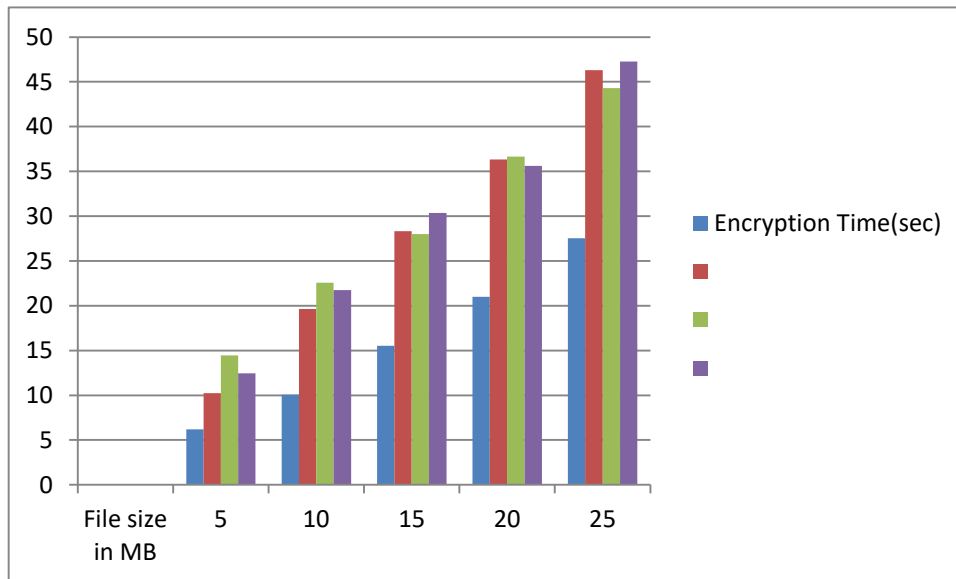


Table 2 Performance comparison of proposed MECC and existing techniques in terms of Decryption Time

File size in MB	Decryption Time (sec)			
	Proposed MECC	DH	Existing ECC	Existing RSA
5	5.28	12.21	13.50	11.51
10	10.22	18.11	22.66	20.53

15	16.74	26.43	29.43	28.54
20	20.36	34.56	38.64	36.87
25	25.44	45.44	45.81	48.43

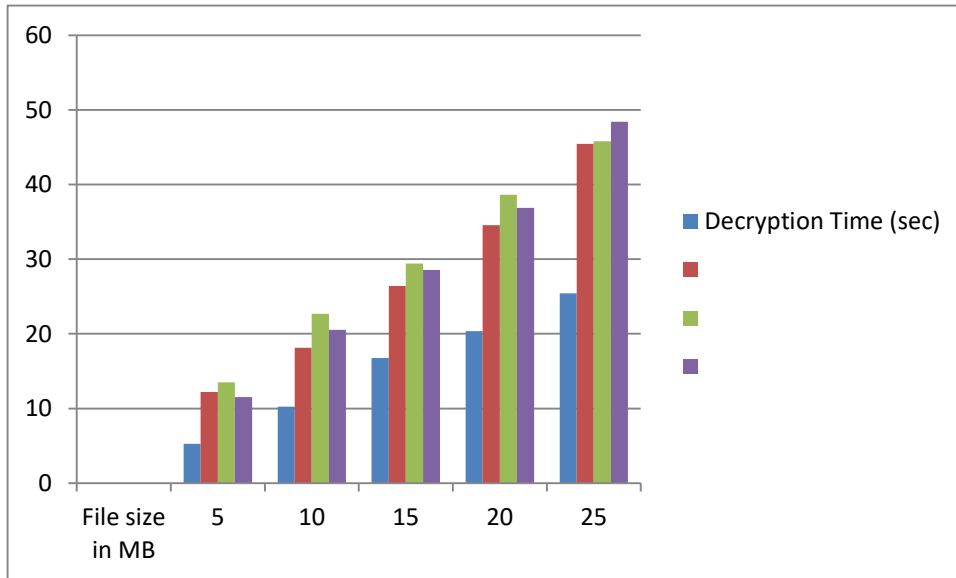
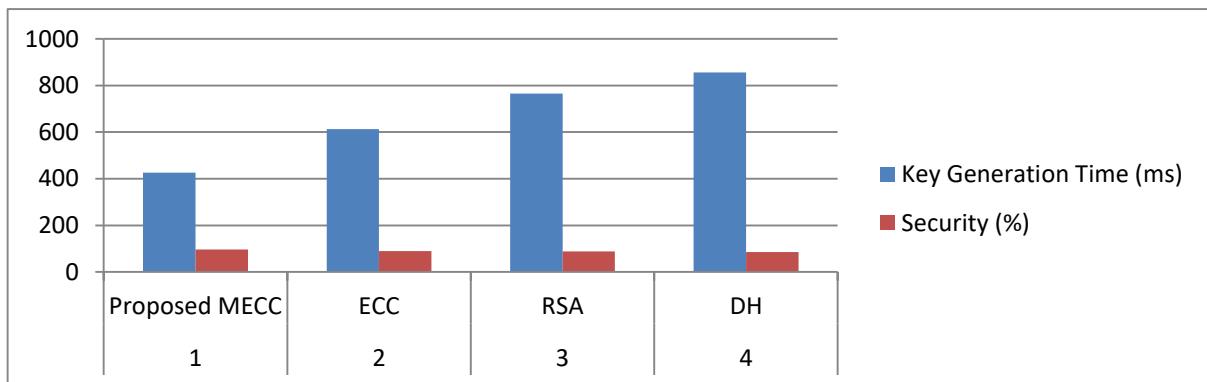


Table 3 Performance Comparison of Proposed MECC and Existing Techniques in terms of Key Generation Time and Security Level

Sl. No	Encryption Algorithms	Key Generation Time (ms)	Security (%)
1	Proposed MECC	425.21	96
2	ECC	612.32	90
3	RSA	765.54	87.5
4	DH	856.33	85



where S denotes the security level, H_d is a hacked data, and O_d denotes the number of original data.

The paper proposes a secure deduplication system that uses convergent and MECC algorithms in a cloud-fog environment. The system was analyzed in four scenarios: when a new user uploads a new file, when the same user uploads the same file, when different users upload the same file, and when different users download the file. The proposed system was tested with file sizes ranging from 5 MB to 25 MB, and the performance analysis showed that the system has 96% security, which is higher than other existing encryption methods.

5.0 Conclusion:

The paper suggests that the proposed model can be extended for IoT applications that use dynamic resource management at the edge environment and in building cyber-physical systems with different use cases and data formats. The proposed technique could increase security and optimize computation time and storage in an integrated environment like IoT or cyber-physical systems. The paper presents a promising approach for secure data deduplication in a cloud-fog environment, and it has potential applications in various domains. However, it would be helpful to provide more information on how the proposed system addresses potential security vulnerabilities and the potential limitations and challenges in implementing the proposed model in real-world scenarios.

6.0 References:

- [1] Lohstroh M, Kim H, Eidson JC et al (2019) On enabling Technologies for the Internet of important things. *IEEE Access* 7:27244–27256. <https://doi.org/10.1109/ACCESS.2019.2901509>
- [2] Abbas N, Zhang Y, Taherkordi A, Skeie T (2018) Mobile edge computing: a survey. *IEEE Internet Things J* 5:450–465. <https://doi.org/10.1109/JIOT.2017.2750180>
- [3] Ren J, Zhang D, He S et al (2019) A survey on end-edge-cloud orchestrated network computing paradigms: transparent computing, mobile edge computing, fog computing, and cloudlet. *ACM Comput Surv*:52. <https://doi.org/10.1145/3362031>
- [4] Zhang P, Liu JK, Richard Yu F et al (2018) A survey on access control in fog computing. *IEEE Commun Mag* 56:144–149. <https://doi.org/10.1109/MCOM.2018.1700333>
- [5] Menon VG, Jacob S, Joseph S, Almagrabi AO (2019) SDN powered humanoid with edge computing for assisting paralyzed patients. *IEEE Internet Things J*:1. <https://doi.org/10.1109/jiot.2019.2963288>
- [6] Menon VG, Prathap J (2017) Vehicular fog computing. *Int J Veh Telemat Infotain Syst* 1:15–23. <https://doi.org/10.4018/ijvtis.2017070102>
- [7] Liu J, Zhang Q (2018) Offloading schemes in Mobile edge computing for ultra-reliable low latency

- communications. *IEEE Access* 6:12825–12837. <https://doi.org/10.1109/ACCESS.2018.2800032>
- [8] Li S, Zhang N, Lin S et al (2018) Joint admission control and resource allocation in edge computing for internet of things. *IEEE Netw* 32:72–79. <https://doi.org/10.1109/MNET.2018.1700163>
- [9] Nadesh RK, Aramudhan M (2018) TRAM-based VM handover with dynamic scheduling for improved QoS of cloud environment. *Int J Internet Technol Secur Trans*:8. <https://doi.org/10.1504/IJITST.2018.093340>
- [10] Rajesh S, Paul V, Menon VG, Khosravi MR (2019) A secure and efficient lightweight symmetric encryption scheme for transfer of text files between embedded IoT devices, pp 1–21
- [11] Zhang J, Chen B, Zhao Y et al (2018) Data security and privacy-preserving in edge computing paradigm: survey and open issues. *IEEE Access* 6:18209–18237. <https://doi.org/10.1109/ACCESS.2018.2820162>
- [12] Nadesh RK, Srinivasa Perumal R, Shynu PG, Sharma G (2018) Enhancing security for end users in cloud computing environment using hybrid encryption technique. *Int J Eng Technol* 7
- [13] Abbasi M, Rafiee M, Khosravi MR et al (2020) An efficient parallel genetic algorithm solution for vehicle routing problem in cloud implementation of the intelligent transportation systems. *J Cloud Comput* 9. <https://doi.org/10.1186/s13677-020-0157-4>
- [14] Subramanian N, Jeyaraj A (2018) Recent security challenges in cloud computing. *Comput Electr Eng* 71:28–42. <https://doi.org/10.1016/j.compeleceng.2018.06.006>
- [15] Jiang S, Jiang T, Wang L (2017) Secure and efficient cloud data Deduplication with ownership management. *IEEE Trans Serv Comput* 12: 532–543. <https://doi.org/10.1109/TSC.2017.2771280>
- [16] Yoon MK (2019) A constant-time chunking algorithm for packet-level deduplication. *ICT Express* 5:131–135. <https://doi.org/10.1016/j.ict.2018.05.005>
- [17] Wang L, Wang B, Song W et al (2019) Offline privacy preserving proxy re-encryption in mobile cloud computing. *Inf Sci (Ny)* 71:38–43. <https://doi.org/10.1016/j.jksuci.2019.05.007>
- [18] Wang L, Wang B, Song W, Zhang Z (2019) A key-sharing based secure deduplication scheme in cloud storage. *Inf Sci (Ny)* 504:48–60. <https://doi.org/10.1016/j.ins.2019.07.058>
- [19] Kwon H, Hahn C, Kim D, Hur J (2017) Secure deduplication for multimedia data with user revocation in cloud storage. *Multimed Tools Appl* 76:5889–5903. <https://doi.org/10.1007/s11042-015-2595-4>
- [20] Akhila K, Ganesh A, Sunitha C (2016) A study on Deduplication techniques over encrypted data. *Procedia Comput Sci* 87:38–43. <https://doi.org/10.1016/j.procs.2016.05.123>

Chapter-4

Privacy preservation in dynamic data through synonymous linkage on microaggregation

Abstract

The rise of the big data age and the growth of the mobile Internet and intelligent gadgets are undoubtedly responsible for the digitalization of personal information. However, the dissemination of such information raises the possibility of privacy violations. To address this issue, privacy preserving data publishing methods have been proposed. However, current methods based on anonymous models may not be effective in protecting non-numerical sensitive information that may contain synonymous linkages leading to privacy breaches. This research proposes a microaggregation-based dynamic data publishing strategy to get around this restriction. The suggested approach adds a number of indicators to assess the relationships between values that are not numerically sensitive, which enhances the clustering impact of the microaggregation anonymous approach. In order to support the dynamic release and update of data, a dynamic update programme is also included. In comparison to current state-of-the-art methodologies, experimental research reveals that the suggested method offers greater privacy protection and publishable data availability. Therefore, the suggested microaggregation-based privacy-preserving dynamic data publishing strategy may provide a practical way to safeguard sensitive data while facilitating the sharing and publication of huge data.

Keywords : 1. Privacy preserving 2. Microaggregation 3. Big Data. 4. Clustering.

1.1 Introduction:

The value of big data has attracted significant attention from governments, industries, and research departments worldwide, leading to the development of numerous technological innovations and applications. Traditional privacy preserving data publishing methods, such as deleting identifying attributes, may not provide sufficient protection from linking attacks. The K-anonymity and l-diversity models were put out as solutions to this problem, however they might not be successful in securing naturally semantically relevant non-numerical sensitive information. This problem is addressed by the research, which suggests a microaggregation-based dynamic data release technique that protects privacy by avoiding the synonymous linking of sensitive variables. The proposed method offers a number of indicators to evaluate synonymous relationships between non-numerically sensitive variables, which strengthens the clustering effect of the microaggregation algorithm. The approach also includes a dynamic

update plan that enables the dynamic release and refresh of data. The work is divided into various sections, including a summary of relevant research, an introduction to the fundamental indicators utilised in the conventional microaggregation approach, and a description of the design indicators for the suggested algorithm. The study then finishes with a summary of its contributions and the presentation of experimental findings. Overall, the suggested privacy-preserving dynamic data release technique based on microaggregation offers a practical method for safeguarding naturally semantically relevant non-numerical sensitive information while facilitating the sharing and publication of huge data. The paper's contributions include the introduction of new indicators, an improved microaggregation algorithm, and a dynamic update program. These contributions provide a significant step forward in the development of privacy-preserving data publishing methods in dynamic and real-time big data scenarios.

1.2 Related Work

The K-anonymity model generalises the quasi-identifier properties of records in a dataset to a certain value range in order to divide records into comparable classes with at least K records having the same quasi-identifier values. Other more effective anonymous models, such as l-diversity and t-closeness, have been proposed to enhance the K-anonymity model's capacity to safeguard user privacy. The generalisation operation on quasi-identifiers is used by several existing anonymization approaches, which can be computationally costly and lead to considerable information loss. However, microaggregation and clustering techniques can also be used to produce the split of analogous classes based on quasi-identifiers. Various researchers have proposed different microaggregation and clustering methods to achieve K-anonymity and improve the privacy protection of static and dynamic data sets. These methods include knowledge-based numerical mapping, hierarchical clustering, genetic algorithms, distance metrics, information entropy, fuzzy possibilistic clustering, linear discriminant analysis, optimized prepartitioning strategy, efficient clustering method, weighted K-member clustering, K-center clustering approach, particle swarm optimization algorithm, and equi-cardinal clustering. Overall, the literature shows a growing interest in privacy-preserving data publishing, and researchers continue to propose new and improved methods for achieving K-anonymity and other anonymous models.

1.3 Prior knowledge:

Data records are divided into equivalent classes using the microaggregation approach based on the highest intra-class and lowest inter-class similarity. To evaluate similarity, a

distance metric is used. Attributes in a relational database can be continuous or discrete, with discrete attributes further classified as nominal or ordinal. While ordinal characteristics have a meaningful order or ranking among values, nominal qualities might have semantic connections or no associations between values. .

ID	Age	Pincode	Sex	Religion	Capitalgain
1	34	515001	F	Hinduism	fair
2	36	515401	M	Christian	excellent
3	48	515230	F	Sikhism	fair
4	52	515430	M	Christian	moderate

Table 1. Table of mixed attributes.

Capitalgain is a discrete ordinal property having three values—moderate, fair, and excellent—and non-semantic relationships. In order to evaluate the relationships between records with numerous attributes, a number of distance metrics have been established.

Definition 1 (Distance for a characteristic that is continuous). The distance between two values $v_i, v_j \in C$ for any continuous attribute C in data table T may be defined as:

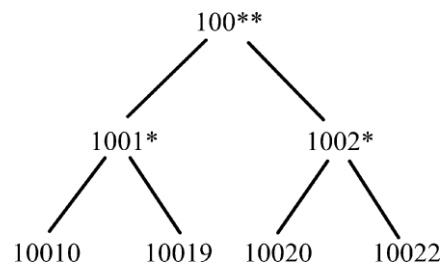


Figure 1. Tree of taxonomy for attribute Pincode.

$$d(v, v) = \frac{|v_i - v_j|}{c_{ij} [\max(C) - \min(C)]} \quad (1)$$

where $\max(C)$ and $\min(C)$ denote the highest and lowest values of a continuous attribute C , respectively

Definition 2 (Distance for nominal property with semantic association). The distance between two values v_i, v_j, N_s for any semantic correlation nominal property N_s in data table T may be written as:

$$d_{NS}(v_i - v_j) = \begin{cases} 0 ; & v_i = v_j \\ \frac{|Parent(v_i, v_j)|}{|TreeNs|} & v_i \neq v_j \end{cases} \quad (2)$$

Where $|TreeNs|$ is the total number of leaf nodes for TreeNs, where TreeNs is the taxonomy tree for the semantic correlation nominal property Ns.

Definition 3 (Distance for nominal characteristic with non-semantic association). The distance between two values v_i, v_j N for any non-semantic correlation nominal property N in data table T may be written as follows:

$$d(v, v) = \frac{p-match(v_i, v_j)}{N_i - P_j} \quad (3)$$

Definition 4 (Distance for the ordinal characteristic). Any ordinal property O in data table T may have a distance between two values v_i, v_j that may be defined as:

$$d_o(v_i, v_j) = |\varphi(v_i) - \varphi(v_j)| \quad (4)$$

$$\text{With : } \varphi(v) = \frac{rank(v)-1}{|O|} \quad (5)$$

where $|O|$ is the number of different values in the ordinal attribute O, and $rank(v)$ is the rank of value v in ascendant order.

Definition 5 (The separation of two recordings). The distance between two records r_1, r_2 T is defined as follows for a data table T with continuous characteristics $C_i (i = 1, \dots, m)$, semantic correlation nominal attribute $N_{sj} (j = 1, \dots, n)$, non-semantic correlation nominal attribute $N_g (g = 1, \dots, x)$, and ordinal attributes $O_h (h = 1, \dots, y)$:

$$d(r_1, r_2) = \frac{1}{|QIA|} \left(\sum_{i=1}^m d_C(r_1(C_i), r_2(C_i)) + \sum_{j=1}^n d_{N_s}(r_1(N_{sj}), r_2(N_{sj})) + \sum_{g=1}^x d_N(r_1(N_g), r_2(N_g)) + \sum_{h=1}^y d_O(r_1(O_h), r_2(O_h)) \right)$$

2.0 Micro aggregation for privacy protection in opposition to synonymous linking:

This approach is achieved by introducing two indicators: the semantic correlation index and the semantic similarity index. The semantic correlation index evaluates the degree of semantic correlation between two non-numerical sensitive values, while the semantic similarity index measures the degree of semantic similarity between two sensitive values. These indicators are used to supplement the traditional micro aggregation metrics, such as Euclidean distance, which only consider the numerical attributes of data records. By incorporating these indicators into the micro aggregation method, the proposed approach can effectively prevent the linkage of non-numerical sensitive information while ensuring the privacy protection of numerical attributes. Additionally, the proposed method can dynamically update data records, ensuring the long-term effectiveness of privacy protection measures. According to experimental findings, the suggested technique performs better in terms of data accessibility and the efficiency of privacy protection than current state-of-the-art solutions. Therefore, it is a potential strategy for disseminating data while protecting privacy, especially in big data scenarios that are dynamic and real-time. The privacy of people in the dataset is better secured by countering synonymous assaults in this manner.

2.1 Micro aggregation-based dynamic data release with privacy protection:

The proposed privacy-preserving micro aggregation approach addresses the challenge of privacy protection in the context of data dissemination. It aims to optimise the trade-off between disclosure risk and information loss by contrasting a unique micro aggregation metric versus synonymous linkage. The first two components of this metric aim to minimise information loss by reducing the distance between quasi-identifier qualities within each equivalent group, while the third component aims to broaden the semantic range of sensitive attributes to prevent synonymous linkage. The approach also includes a dynamic updating programme that allows for the insertion, deletion, and modification of data while using dynamic adjustment to prevent synonymous linkage.

3.0 Micro aggregation publishing algorithm:

Algorithm Based on what you have provided, it seems that the DRASL algorithm utilizes the K-anonymity model to obscure specific values into a range to safeguard personal data. To balance the trade-off between disclosure risk and knowledge loss, a unique micro aggregation metric versus synonymous linkage is used. The third and final component of this measure maximises the semantic variance of sensitive features in each comparable group in

an effort to prevent synonymous coupling. The input data table T and the privacy protection parameter K are used in the publishing process of the DRASL algorithm. Based on a predetermined set of sensitive variables, the approach delivers the anonymous data table T and the clustered equivalent groups GID .

Algorithm 1 DRASL for the first release

Input: Data table T ; parameter K ; predefined catalogue of sensitive values;

Output: Clustered equivalent groups GID ; anonymous data table T^*

```

1: if ( $|T| \leq K$ ) then
2:   Return
3: end if
4: Let  $GID = \emptyset$  //Create an empty list of equivalent groups
5: Let  $T^* = \emptyset$  //Create an empty anonymous table of  $T$ 
6: while ( $|T| > K$ ) do
7:   Select a record  $r$  from  $T$  randomly
8:    $T = T - \{r\}$ 
9:    $gid = \{r\}$ 
10:  while ( $|gid| < K$ ) do
11:    Find a record  $r' \in T$  s.t.  $\max\{f_{link_{SA}}(gid, (gid \cup \{r'\}))\}$ 
12:    Find a group  $gid_j \in GID$  s.t.  $\max\{f_{link_{SA}}(gid, (gid \cup \{gid_j\}))\}$ 
13:    if ( $f_{link_{SA}}(gid, (gid \cup \{r'\})) > f_{link_{SA}}(gid, (gid \cup gid_j))$ ) then
14:       $T = T - \{r'\}$ 
15:       $gid = gid \cup \{r'\}$ 
16:    else
17:       $GID = GID - gid_j$ 
18:       $gid = gid \cup gid_j$ 
19:    end if
20:  end while
21:   $GID = GID \cup gid$ 
22: end while
23: Create an empty list  $Q$ 
24: while ( $|T| \neq 0$ ) do
25:   Select a record  $r$  from  $T$  randomly
26:    $T = T - \{r\}$ 
27:   for each group  $gid_j \in GID$  do
28:      $Q \leftarrow f_{link_{SA}}(gid_j, (gid_j \cup \{r\}))$ 
29:   end for
30:    $j =$  the sequence number of the element with the maximal value in  $Q$ 
31:    $gid_j = gid_j \cup \{r\}$ 
32: end while
33:  $T^* \leftarrow generalization(GID)$ 
34: return  $GID, T^*$ 

```

3.1 Dynamic insertion of record:

To address these issues, the proposed DRASL algorithm 2 includes a dynamic adjustment process for insertion of record.

Algorithm 2 : The procedure works as follows:

Input:

D: the current dataset

r: the record to be inserted

K: the desired anonymity level

L: the number of nearest neighbors to consider

Output:

D': the updated dataset with the new record r inserted

Steps:

Compute the microaggregation metric for each equivalent group in D, based on Definition 11.

Find the L nearest neighbors of r in D, based on the quasi-identifier attributes.

For each of the L nearest neighbors, compute the microaggregation metric for the equivalent group that would result if r were added to that neighbor's group.

Choose the equivalent group with the best microaggregation metric among the L+1 options (i.e., the L nearest neighbors and the group without any nearest neighbors).

If the chosen group has fewer than K-1 records, add forged records to it until it has at least K-1 records.

Add r to the chosen group.

Perform dynamic adjustment to prevent synonymous linkage as follows: a. If the chosen group already has a forged record, update the sensitive value of the forged record randomly. b.

Otherwise, generate a new forged record with a different sensitive value from that of r and add it to the chosen group.

Output the updated dataset D'.

The proposed dynamic adjustment algorithm in situation 1 uses a two-step process to prevent synonymous linkage. Firstly, it generates a set of candidate forged records with different sensitive values, and then selects the best candidate based on the semantic diversity metric (Definition 10) to minimize the probability of synonymous linkage. In situation 2, the algorithm also generates multiple candidate forged records to prevent synonymous linkage. By doing so, the algorithm can effectively protect privacy and maintain data utility even in the face of dynamic data updates.

GID	Age	Pincode	Disease
1	[21-22]	[12***-14***]	Gastro
1	[21-22]	[12***-14***]	Cancer
2	[27-29]	[19***-26***]	Thyriod
2	[27-29]	[19***-26***]	Cardio

Table 1. Similar groupings from a micro patient table clustered together.

GID	Age	Pincode	Disease
1	[21-23]	[13***-16***]	Gastro
1	[21-23]	[13***-16***]	Cancer
1	[21-23]	[13***-16***]	Thyriod
1	[21-23]	[13***-16***]	Cholera
2	[26-28]	[19***-26***]	Thyriod
2	[26-28]	[19***-26***]	Cardio

Table 2. Equivalent groupings were updated once a new record was added.

The suggested approach (approach 3) looks to deal with the problem of sensitive values being exposed during dynamic record insertion by either establishing a new value for the forged record or a new fabricated record entirely. In particular, lines handle the situation where there is already a forgery in the group and the algorithm changes its value to a new value that is not identical to the sensitive value of the new record being inserted. On the other hand, lines handle the situation where there isn't a forgery in the group and the algorithm creates a new forgery with a random sensitive value that is different from the sensitive value of the tuple and has no connection.

GID	Age	Pincode	Disease
1	[22-24]	[13***-16***]	Gastro
1	[22-24]	[13***-16***]	Cancer
1	[22-24]	[13***-16***]	Thyriod
1	[22-24]	[13***-16***]	Asma
2	[27-29]	[19***-26***]	Thyriod
2	[27-29]	[19***-26***]	Cardio

Algorithm 3 DRASL dynamic adjustment for the protection of sensitive values

Input: Clustered equivalent groups GID ; predefined catalogue of sensitive values;

Output: Updated clustered equivalent groups GID with forged record

```
1: while (Group change  $\Rightarrow r \rightarrow SA_r \in D_{SA}$ ) do
2:   Let  $gid_j \in GID$  be the Group where the change occurs
3:   if (there is already  $fg_r \subseteq gid_j$ ) then
4:      $fg_r \rightarrow fg_r(D_{SA\_random} \in D_{SA})$  s.t.  $D_{SA\_random} \neq SA_r$  &  $Link_{SA}(D_{SA\_random}, SA_r) = 0$ 
5:   else
6:      $gid_j \cup gid_j \{fg_r \rightarrow D_{SA\_random} \in D_{SA}\}$  s.t.  $D_{SA\_random} \neq SA_r$  &  $Link_{SA}(D_{SA\_random}, SA_r) = 0$ 
7:   end if
8: end while
9: return  $GID$ 
```

Table 3. Implementing algorithm 3

The dynamic process for deletion in the suggested DRASL algorithm is designed to ensure that sensitive information is protected even after record deletion. Algorithm 4 provides a step-by-step guide on how this process works. When a record r is deleted from an equivalent group Gid_j and the size of the group is less than K , the algorithm first removes Gid_j from the clustered equivalent group GID (Line 7). Next, a random record r is selected from gid_j and removed (Lines 9-10). The algorithm then identifies the most suitable equivalent group for r to join using the improved microaggregation metric (Lines 11-13). Once the best group G_i has been identified.

If the size of Gid_j is still greater than or equal to K after the record deletion, then Algorithm 3 is called to update the group Gid_j (Lines 3-5 and 16-18). This ensures that the equivalent group continues to be properly protected and that sensitive information is not exposed. Overall, the dynamic adjustment process for record deletion in DRASL helps to stabilize secrecy of the data even after records have been removed.

Input: Clustered equivalent groups GID ; deleted record r ; parameter K ; predefined catalogue of sensitive values;
Output: Updated clustered equivalent groups GID

- 1: Let $gid_j \in GID$ be the equivalent group currently containing the deleted record r
- 2: $gid_j = gid_j - \{r\}$
- 3: **if** ($|gid_j| \geq K$) **then**
- 4: $gid_j \leftarrow recall_Algorithm\ 3(gid_j, r)$
- 5: $GID = GID \cup gid_j$
- 6: **else**
- 7: $GID = GID - gid_j$ //retrieve the group gid_j from the set GID
- 8: **while** ($|gid_j| \neq 0$) **do**
- 9: Select a record r from gid_j randomly
- 10: $gid_j = gid_j - \{r\}$
- 11: **for** each group $G_i \in GID$ **do**
- 12: $Q \leftarrow f_{link_{SA}}(G_i, (G_i \cup \{r\}))$
- 13: **end for**
- 14: $i =$ the sequence number of the element with the maximal value in Q
- 15: $G_i = G_i \cup \{r\}$
- 16: $G_i \leftarrow recall_Algorithm\ 3(G_i, r)$
- 17: $GID = GID \cup G_i$
- 18: **end while**
- 19: **end if**
- 20: **return** GID

Algorithm 4: Micro aggregation

Input: modified record r_{mod} and the original record r_{ori} a. Call Algorithm 4 to remove the original record r_{ori} from its equivalent group if the alteration is on quasi-identifiers. b. Invoke Algorithm 2 to add the altered record r_{mod} to the most appropriate equivalent group.

If the modification is on sensitive value: a. Call Algorithm 3 to update the sensitive value of the equivalent group containing the original record r_{ori} b. If the updated sensitive value is different from the original one, call Algorithm 4 to delete the original record r_{ori} from its equivalent group and add the new record r_{mod} into the equivalent group containing the updated sensitive value c. If the updated sensitive value is synonymous linked with the modified sensitive value, generate a new value and repeat Step 3a.

The algorithm handles the modification of records either by deleting the old record and inserting the modified record into a suitable equivalent group, or by updating the sensitive value and checking whether it is still protected from synonymous linkage. If the updated sensitive value is not protected, the algorithm generates a new value and repeats the process until the new value is suitable for insertion.

Algorithm 5:

Input: Clustered equivalent groups GID ; modify record r ; parameter K ; predefined catalogue of sensitive values;

Output: Updated clustered equivalent groups GID

```
1: Let  $gid_j \in GID$  be the equivalent group currently containing the modify record  $r$ 
2: while (record change  $r \rightarrow r'$ ) do
3:   while (modification includes only quasi-identifiers  $r[QID]$ ) do
4:      $gid_j \leftarrow recall\_Algorithm\ 4(gid_j, r)$ 
5:      $GID = GID \cup gid_j$ 
6:      $GID \leftarrow recall\_Algorithm\ 2(GID, r')$ 
7:   end while
8:   while (modification includes only sensitive value  $SA_r$ ) do
9:     if (Cash table  $H$  contains the modify record  $r$ ) then
10:       $H = H \cup H\{r \rightarrow SA_r = SA_{r'}\}$ 
11:     end if
12:      $gid_j = gid_j \cup gid_j\{r \rightarrow SA_r = SA_{r'}\}$ 
13:      $gid_j \leftarrow recall\_Algorithm\ 3(gid_j, r')$ 
14:      $GID = GID \cup gid_j$ 
15:   end while
16: end while
17: return  $GID$ 
```

The proposed DRASL algorithm can be evaluated by measuring the reduction in probability mass synonymous linkage between sensitive values in the published data. The DRASL algorithm introduces dynamic adjustments to the micro aggregation process, which can improve the protection of privacy in published data by reducing the possibility of synonymous linkage between sensitive values. Additionally, the paper highlights the trade-off between privacy protection and data utility. The suggested approach optimises the micro aggregation measure against synonymous linkage in an effort to strike a compromise between these two criteria. The DRASL method efficiently safeguards security while managing the usefulness of the released data by minimising information loss during micro aggregation and maximising the semantic variety of sensitive qualities within each analogous group. This compromise is crucial because too stringent privacy protection methods may result in substantial information loss and lower the usefulness of public data for analysis and decision-making.

4.0 Conclusion:

Overall, the suggested solution solves the shortcomings of previous methods in preventing privacy leakage caused by semantic links between non-numerically sensitive variables, which is a potential contribution to the field of privacy preserving data publication. The use of micro aggregation and updates enables more efficient clustering of comparable groups and improved privacy protection. It is crucial to remember that the suggested approach might not be appropriate for all large data publishing scenarios, including those involving

unstructured data and graph data. To create strategies for privacy preservation in these situations, more study is required. All things considered, the suggested approach has the potential to be used in a variety of publication scenarios and can support ongoing attempts to achieve privacy preservation in voluminous data.

5.0 References :

- [1] Ge, M., Bangui, H. & Buhnova, B. Big data for internet of things: a survey. *Fut. Gen. Comput. Syst.* **87**, 601–614 (2018).
- [2] Zhu, L., Yu, F. R., Wang, Y., Ning, B. & Tang, T. Big data analytics in intelligent transportation systems: a survey. *IEEE Trans. Intell. Transp. Syst.* **20**(1), 383–398 (2019).
- [4] Qi, C. Big data management in the mining industry. *Int. J. Miner. Metall. Mater.* **27**, 131–139 (2020).
- [5] Shamsi, J. A. & Ali, K. M. Understanding privacy violations in big data systems. *IT Professional.* **20**(3), 73–81 (2018).
- [6] Lv, Z. & Qiao, L. Analysis of healthcare big data. *Fut. Gen. Comput. Syst.* **109**, 103–110 (2020).
- [7] Anupam, D., Sarma, K. & Deka, S. Data security with DNA cryptography. *Proceedings of the World Congress on Engineering* **2019**, 246–252 (2019).
- [8] Samarati, P. & Sweeney, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. *SRI Computer Science Laboratory.* 1–19 (1998).
- [9] Sweeney, L. K-anonymity: a model for protecting privacy. *Internat. J. Uncertain. Fuzziness Knowl.-Based Syst.* **10**(5), 557–570(2002).
- [10] Samarati, P. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.* **13**(6), 1010–1027 (2001).
- [11] Machanavajjhala, A., Gehrke, J., Kifer, D. & Venkatasubramanian, M. *l*-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1**(1), 3 (2007).
- [12] Li, N., Li, T., Venkatasubramanian S. & CSMDL. *t*-closeness: Privacy beyond k-anonymity and *l*-diversity. *IEEE 23rd International Conference on Data Engineering.* 106–115 (2007).
- [13] Palanisamy, B., Liu, L., Zhou, Y. & Wang, Q. Privacy-preserving publishing of multilevel utility-controlled graph datasets. *ACM Trans. Internet Tech.* **18**, 1–21 (2018).

- [14] Temuujin, O., Ahn, J. & Im, D. H. Efficient l -diversity algorithm for preserving privacy of dynamically published datasets. *IEEE Access*. **7**, 122878–122888 (2019).
- [15] Xiao, Y. & Li, H. Privacy Preserving data publishing for multiple sensitive attributes based on security level. *Inf. (Switzerland)*.**11**, <https://doi.org/10.3480/info11030166> (2020).
- [16] Domingo-Ferrer, J. & Torra, V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Min. Knowl. Disc.* **11**(2), 195–212 (2005).
- [17] Domingo-Ferrer, J. & Mateo-Sanz, J. M. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* **14**(1), 189–201 (2002).
- [18] Domingo-Ferrer, J., Sanchez, D. & Rufian-Torrell, G. Anonymization of nominal data based on semantic marginality. *Inf. Sci.* **242**, 36–48 (2013).
- [19] Domingo-Ferrer, J., Soria-Comas, J. & Mulero-Vellido, R. Steered microaggregation as a unified primitive to anonymize data sets and data streams. *IEEE Trans. Inf. Forensics Secur.* **14**(12), 3298–3411 (2019).