FALL 2023–24 · TARP

Review 2

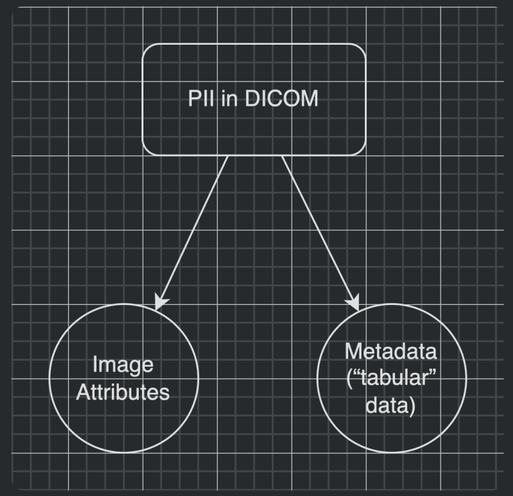
medical record anonymisation

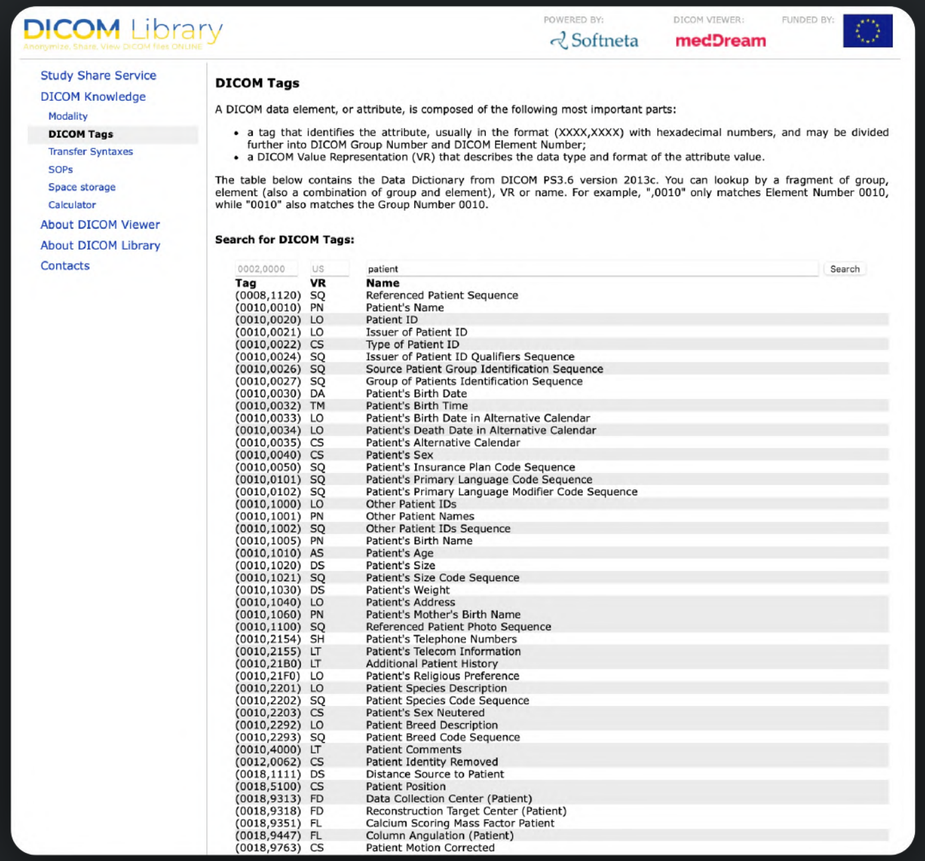
augmenting the $k\text{-}anonymity$ algorithm using $l\text{-}diversity$

Bedant (20BCE0568) · Shreedhar (20BCE0576)
Sai Ruthvik (20BCE0929) · Asmita (20BCE2454),
Shreyoshi (20BCI0244) · Anuj Aghi (20BCEI0266)

The Need for Anonymity

- DICOM image format — a lot of personal identifiable information (PII)
- Complete removal/redaction of info not great for research
- Manipulation = better alternative?





DICOM Library
Anonymize, Share, View DICOM files ONLINE

POWERED BY: Softneta | DICOM VIEWER: medDream | FUNDED BY:

- Study Share Service
- DICOM Knowledge
  - Modality
  - **DICOM Tags**
  - Transfer Syntaxes
  - SOPs
  - Space storage
  - Calculator
- About DICOM Viewer
- About DICOM Library
- Contacts

DICOM Tags

A DICOM data element, or attribute, is composed of the following most important parts:

- a tag that identifies the attribute, usually in the format (XXXX,XXXX) with hexadecimal numbers, and may be divided further into DICOM Group Number and DICOM Element Number;
- a DICOM Value Representation (VR) that describes the data type and format of the attribute value.

The table below contains the Data Dictionary from DICOM PS3.6 version 2013c. You can lookup by a fragment of group, element (also a combination of group and element), VR or name. For example, ",0010" only matches Element Number 0010, while "0010" also matches the Group Number 0010.

Search for DICOM Tags:

0002,0000 | US | patient | Search

Tag	VR	Name
(0008,1120)	SQ	Referenced Patient Sequence
(0010,0010)	PN	Patient's Name
(0010,0020)	LO	Patient ID
(0010,0021)	LO	Issuer of Patient ID
(0010,0022)	CS	Type of Patient ID
(0010,0024)	SQ	Issuer of Patient ID Qualifiers Sequence
(0010,0026)	SQ	Source Patient Group Identification Sequence
(0010,0027)	SQ	Group of Patients Identification Sequence
(0010,0030)	DA	Patient's Birth Date
(0010,0032)	TM	Patient's Birth Time
(0010,0033)	LO	Patient's Birth Date in Alternative Calendar
(0010,0034)	LO	Patient's Death Date in Alternative Calendar
(0010,0035)	CS	Patient's Alternative Calendar
(0010,0040)	CS	Patient's Sex
(0010,0050)	SQ	Patient's Insurance Plan Code Sequence
(0010,0101)	SQ	Patient's Primary Language Code Sequence
(0010,0102)	SQ	Patient's Primary Language Modifier Code Sequence
(0010,1000)	LO	Other Patient IDs
(0010,1001)	PN	Other Patient Names
(0010,1002)	SQ	Other Patient IDs Sequence
(0010,1005)	PN	Patient's Birth Name
(0010,1010)	AS	Patient's Age
(0010,1020)	DS	Patient's Size
(0010,1021)	SQ	Patient's Size Code Sequence
(0010,1030)	DS	Patient's Weight
(0010,1040)	LO	Patient's Address
(0010,1060)	PN	Patient's Mother's Birth Name
(0010,1100)	SQ	Referenced Patient Photo Sequence
(0010,2154)	SH	Patient's Telephone Numbers
(0010,2155)	LT	Patient's Telecom Information
(0010,21B0)	LT	Additional Patient History
(0010,21F0)	LO	Patient's Religious Preference
(0010,2201)	LO	Patient Species Description
(0010,2202)	SQ	Patient Species Code Sequence
(0010,2203)	CS	Patient's Sex Neutered
(0010,2292)	LO	Patient Breed Description
(0010,2293)	SQ	Patient Breed Code Sequence
(0010,4000)	LT	Patient Comments
(0012,0062)	CS	Patient Identity Removed
(0018,1111)	DS	Distance Source to Patient
(0018,5100)	CS	Patient Position
(0018,9313)	FD	Data Collection Center (Patient)
(0018,9318)	FD	Reconstruction Target Center (Patient)
(0018,9351)	FL	Calcium Scoring Mass Factor Patient
(0018,9447)	FL	Column Angulation (Patient)
(0018,9763)	CS	Patient Motion Corrected

Introduction

- Critical need for anonymisation of sensitive medical datasets such as X-rays, MRIs, and CT-scans.
- These images (**DICOM** format: $digital\ imaging\ and\ communications\ in\ medicine$) contain highly confidential patient information (PHI) and must be redacted.
- **Redaction can limit research.**
- What if the records were manipulated to remove unique, identity information, yet **maintained statistic fidelity**?
- Scale: (2021) ~130 PACS ($picture\ archiving\ and\ communication\ system$) actively hosting **8.5 million case studies** and records. The data represents more than 2 million patients, with approximately 275 million images related to their exams. DICOM — archaic standard, patient-tagged images. Greenbone Networks research; 97 vuln. systems in India in 2020 (121 mil leaked images).

~130 PACS ~275 mil records ~2 mil patients ~8.5 mil case studies

~1 mil records publicly available on *DICOM Library*

Motivation (Technical)

Dated, Insecure Solutions: (2021) ~130 PACS systems actively exposing 8.5 million case studies. The data represents more than 2 million patients, with approximately 275 million images related to their exams. DICOM — archaic standard, patient-tagged images. Greenbone Networks research; 97 vuln. systems in India in 2020 (121 mil leaked images).

Digitalisation: Electronic Health Records (EHR) — increase in accessibility and portability of patient data, as well as risk (DICOM plays a role too). In 2014, the implementation of electronic records in US hospitals rose from 10% to 70%, indicating a significant shift towards digital record-keeping.

Motivation (Social)

Health Professionals' Ethical Responsibility and Regulatory Compliance: Healthcare professionals are required to provide quality care to patients, which includes safeguarding their medical information. This encourages the adoption of secure practices such as image slicing for safeguarding critical information in the form of images.

Patient-Centered Approach: It involves respecting patient values, preferences, and autonomy while providing high-quality medical services. This approach has a significant impact on healthcare practices, policies, and decision-making, and it's closely linked to the social feasibility of addressing the problem of secure medical image transmission.

Public Awareness and Concern: With the increasing awareness of data breaches and privacy issues, the general public is increasingly concerned about the security of their personal and medical information. This heightened awareness creates social pressure on healthcare institutions to implement secure transmission solutions.

Motivation (Economic)

Healthcare continues to have the highest data breach costs of all industries, according to a new report from IBM. It revealed that the average cost of a healthcare data breach is now $10.93 million — up from $10.10 million in 2022.

Over the past three years, the average cost of a healthcare data breach has risen by 53.3%, the report said. In 2020, it was $7.13 million.

Healthcare organizations worldwide averaged 1,463 cyberattacks per week in 2022. As of July 2023, more than 330 breaches have been reported so far, affecting 43 million people, which is rapidly approaching 2022's total of 52 million impacted patients.

Motivation (Political)

DP (Personal Data Protection Bill): IAMAI (Internet & Mobile Assoc. of India) put healthcare explicitly under PDP. NDHB proposal: district-level patient-record databases, NHIA (National Health Information Architecture).

Frameworks: HIPAA, HITECH acts promote the secure and ethical use of electronic health data. (HIPAA — health insurance portability & accountability: ensuring privacy of patient records, HITECH — health IT for economic & clinical health: promotion of privacy-preserving EHR technology)

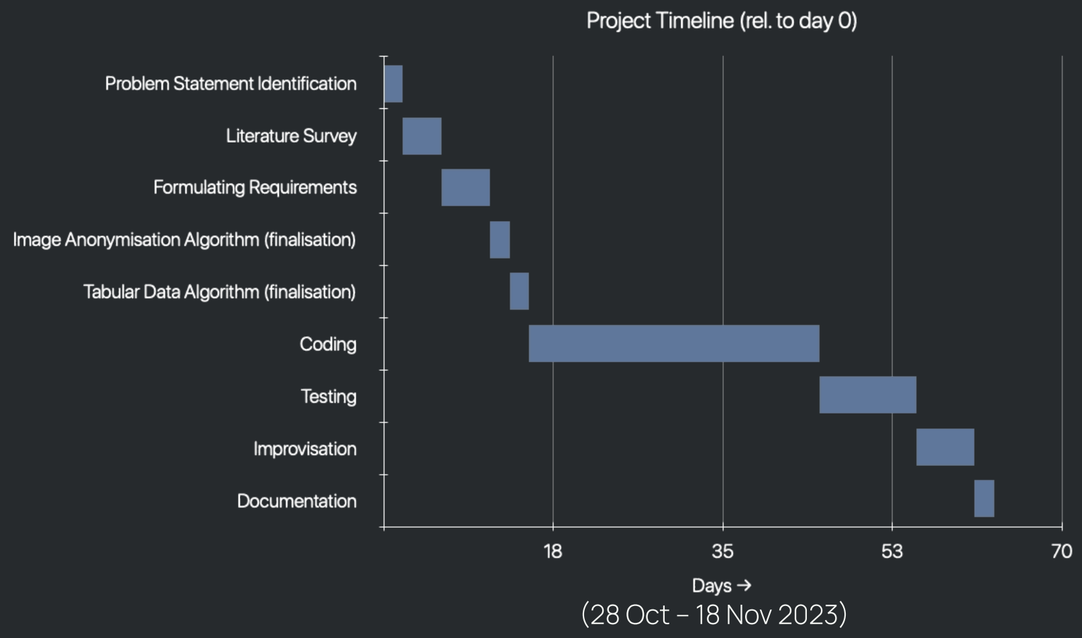
Project Outcome

The project's outcome for achieving k-anonymity of medical records is the successful implementation of a comprehensive framework that ensures the privacy and security of patient data. By employing encryption, data masking, and de-identification techniques, along with robust access controls and an audit trail system, the project aims to protect sensitive information and restrict access only to authorized personnel.

As a short term goal, we intend to make a research paper on the topic for everyone to be able to gain access to methodologies for data anonymization.

Timeline

Activity	Days	Start	End
Problem Statement Identification	2	28/8/23	30/8/23
Literature Survey	4	30/8/23	3/9/23
Formulating Requirements	5	3/9/23	8/9/23
Image Anonymisation Algorithm (finalisation)	2	8/9/23	10/9/23
Tabular Data Algorithm (finalisation)	2	10/9/23	12/9/23
Coding	38	12/9/23	20/10/23
Testing	10	12/10/23	22/10/23
Improvisation	6	22/10/23	28/10/23
Documentation	2	28/10/23	30/10/23



(28 Oct – 18 Nov 2023)

Review of Existing Literature

fuzzy clustering-based k-anonymization

- The paper outlines a novel approach based on fuzzy clustering for achieving k-anonymization, highlighting the importance of preserving privacy in the context of crowd movement analysis with face recognition technology.
- A comprehensive comparative analysis is conducted, pitting the proposed fuzzy clustering-based method against various existing k-anonymization strategies, emphasizing the trade-off between privacy preservation and data utility in the specific context of crowd movement analysis.
- Results indicate that the suggested fuzzy clustering-based approach outperforms other methods, demonstrating superior data quality (utility) and enhanced privacy protection, thereby showcasing its efficacy in balancing the need for privacy with the requirements for accurate analysis in scenarios involving crowd movement and face recognition.
- The study underscores the significance of customizing the anonymization approach based on the unique requirements of the application, highlighting the need for a tailored evaluation process to align privacy objectives with the specific demands and constraints of crowd movement analysis and face recognition technologies. This underscores the importance of not adopting a one-size-fits-all solution in privacy-preserving data analytics.

Review of Existing Literature

Face Picture Anonymization using N-D k-anonymity

- The research highlights the complex challenge of balancing privacy preservation with the imperative to enable meaningful analysis of face photographs, particularly in contexts where both privacy concerns and the utility of data are of paramount importance.
- Through a detailed analysis, the paper demonstrates that the proposed approach based on N-D k-anonymity for anonymizing face pictures outperforms alternative methods, reaffirming its capability to strike a balance between ensuring high data quality and safeguarding privacy in the realm of face picture analysis.
- The study underscores the critical role of adopting sophisticated multidimensional data k-anonymization techniques in addressing the intricacies associated with protecting the privacy of individuals in face picture datasets, providing further evidence of the effectiveness of the proposed approach in meeting the dual requirements of data utility and privacy preservation.

Review of Existing Literature

Weighted k-$member$ Clustering for k-$anonymity$

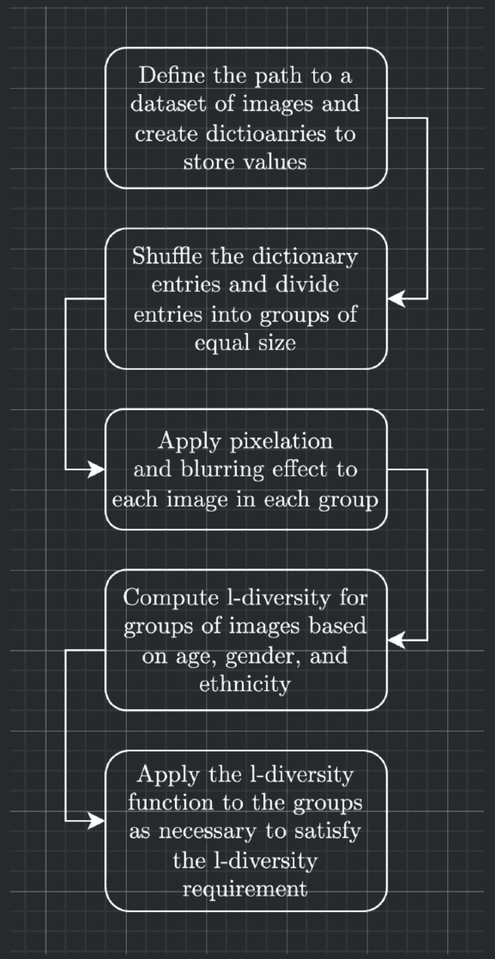
- The paper introduces a distinctive k-anonymization approach, employing a weighted clustering mechanism designed to fortify the protection of sensitive information while simultaneously ensuring the maintenance of high data quality, catering to the intricate demands of privacy preservation in diverse datasets.
- The suggested method involves the segregation of data into clusters based on their similarities, and subsequently assigns weights to each record based on the sensitivity of the information, allowing for a nuanced approach to privacy preservation that accounts for varying degrees of sensitivity within the dataset.
- The proposed algorithm is demonstrated to be versatile and applicable in various domains, including but not limited to healthcare, banking, and social media, underscoring its potential for widespread adoption across different industries and sectors grappling with the challenges of data privacy and security. By leveraging weighted k-member clustering, the method showcases its adaptability in catering to the nuanced requirements of diverse applications, extending its utility beyond a singular domain and highlighting its potential for broader implementation in privacy-sensitive contexts.

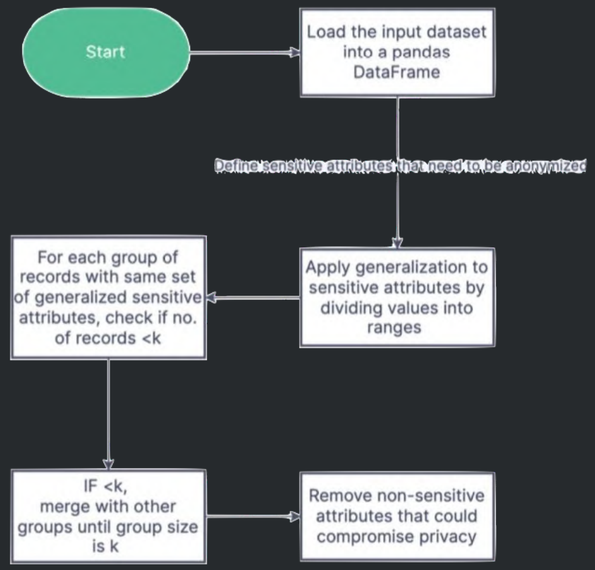
Review of Existing Literature

Scalable Anonymisation for Utility in Big Data Publishing

- The paper introduces an enhanced approach to l-diversity for scalable anonymization of large-scale datasets, emphasizing the critical balance between achieving heightened levels of privacy protection and minimizing data distortion, particularly in the context of big data publishing.
- The strategy operates through a two-step process, commencing with the partitioning of the data into subgroups based on similarity, followed by the manipulation of sensitive attributes within each subset using a probabilistic method. This dual approach enables the preservation of data utility while fortifying the privacy of individuals within the dataset, addressing the challenges posed by the anonymization of large-scale datasets.
- The research underscores the significance of scalable anonymization techniques in the context of big data, highlighting the imperative to develop sophisticated strategies that can efficiently handle the anonymization process for massive datasets without compromising the utility of the data. By leveraging the partitioning and probabilistic manipulation approach, the method demonstrates its potential for widespread adoption in various big data applications, offering a scalable solution that balances the imperatives of privacy protection and data utility in large-scale data publishing scenarios.

Proposed Methodology- Block diagrams





Proposed Methodology

The paper introduces a pioneering algorithm designed to enhance the anonymization of medical records, with a particular focus on preserving data utility in the context of DICOM (Digital Imaging and Communications in Medicine) format records. Our approach builds upon the well-established foundation of k-anonymity by incorporating the innovative concept of l-diversity. This amalgamation addresses a critical limitation of traditional k- anonymity, where individual records can still be vulnerable to attribute disclosure. By leveraging l-diversity, our algorithm ensures that within each equivalence class of k-anonymous records, there is a diversified representation of sensitive attributes, thereby fortifying the protection of patient identities. Notably, the novelty of our improved algorithm lies in its adaptability to the complex and multifaceted nature of DICOM data, which encompasses both graphical image information and tabular data containing identifying details. This unique feature facilitates comprehensive anonymization, rendering the algorithm highly effective in safeguarding patient privacy while simultaneously optimizing the utility of the anonymized data for research purposes. Researchers can confidently utilize the anonymized DICOM records, knowing that the risk of attribute disclosure has been significantly mitigated, thus unlocking new opportunities for valuable medical research without compromising patient confidentiality.

Implementation Screenshots

```
def get_partition_rects(df, partitions, column_x, column_y, indexes, offsets=[0.1, 0.1]):
    rects = []
    for partition in partitions:
        xl, xr = get_coords(df, column_x, partition, indexes, offset=offsets[0])
        yl, yr = get_coords(df, column_y, partition, indexes, offset=offsets[1])
        rects.append(((xl, yl),(xr, yr)))
    return rects
```

The function get_partition_rects essentially uses the coordinates from get_coords function) to construct rectangles representing the partitions.

```
def plot_rects(df, ax, rects, column_x, column_y, edgecolor='black', facecolor='none'):
    for (xl, yl),(xr, yr) in rects:
        ax.add_patch(patches.Rectangle((xl,yl),xr-xl,yr-yl,linewidth=1,edgecolor=edgecolor,facecolor=facecolor, alpha=0.5))
    ax.set_xlim(*get_bounds(df, column_x, indexes))
    ax.set_ylim(*get_bounds(df, column_y, indexes))
    ax.set_xlabel(column_x)
    ax.set_ylabel(column_y)
```

The function iterates through each rectangle in the rects list, adds each rectangle as a patch to the provided axis ax, and sets the appropriate limits and labels for the plot based on the bounds obtained from the get_bounds function.

Implementation Screenshots

```python
def build_anonymized_dataset(df, partitions, feature_columns, sensitive_column, max_partitions=None):
    aggregations = {}
    for column in feature_columns:
        if column in categorical:
            aggregations[column] = agg_categorical_column
        else:
            aggregations[column] = agg_numerical_column
    rows = []
    for i, partition in enumerate(partitions):
        if i % 100 == 1:
            print("Finished {} partitions...".format(i))
        if max_partitions is not None and i > max_partitions:
            break
        grouped_columns = df.loc[partition].agg(aggregations, squeeze=False)
        sensitive_counts = df.loc[partition].groupby(sensitive_column).agg({sensitive_column : 'count'})
        values = grouped_columns.iloc[0].to_dict()
        for sensitive_value, count in sensitive_counts[sensitive_column].items():
            if count == 0:
                continue
            values.update({
                sensitive_column : sensitive_value,
                'count' : count,

            })
            rows.append(values.copy())
    return pd.DataFrame(rows)
```

The build_anonymized_dataset function takes in a dataset df, partitions it based on provided partitions, and performs aggregations on specified feature_columns while considering a sensitive_column. It creates a dictionary aggregations to store aggregation functions for each feature column, distinguishing between categorical and numerical columns based on the categorical list. It iterates through the partitions, aggregates the data for each partition, and counts the occurrences of values in the sensitive column. The function then constructs a DataFrame with the aggregated values and the sensitive column counts, providing an anonymized representation of the original dataset. If provided, the parameter max_partitions limits the number of partitions processed, and the function prints the progress every 100 partitions.