Privacy Preservation of Cluster Integrity on Web-Scraped Hospital Data

Hrishik Sai Bojnal, Dharmisht SVK, J Krishna Kaarthik, Dhyaan Kotian, Shashidhar Virupaksha
Department of Computer Science and Engineering
RV Institute of Technology and Management
Bengaluru, Karnataka, India,
Visvesvaraya Technological University
Belagavi, Karnataka, India
Emails: hrishiksai@outlook.com, dharmisht_11@yahoo.co.in, kaarthikj511@gmail.com, dhyaan.kotian@outlook.com, shashidhar.virupaksha@gmail.com

***Abstract*— This paper proposes a method for anonymizing geospatial healthcare data that preserves the natural clustering of locations. A dataset is built with each healthcare facility's location, ratings, and review count by scraping it off an online maps platform. This enables the calculation of each facility's regional influence. The approach retains spatial clusters by restricting the displacements on each facility's geospatial coordinates to at most the minimum radius of the cluster while also factoring for the prominence of the facility in the region. This ensures that the shape and magnitude distribution of clusters is maintained, thereby preserving dataset quality. Impactful insights can be drawn from this anonymized dataset without having to identify any individual facility.**

Keywords— Web scraping, Healthcare data, Anonymization techniques, Privacy preservation, Geospatial data, Cluster-preserving anonymization, Spatial data integrity

I. Introduction

In the era of big data, the ability to collect and analyze vast amounts of information has transformed numerous fields, including healthcare. One method that has gained prominence is web scraping, which involves the automated extraction of data from websites. This technique enables researchers to gather large datasets from various online sources, facilitating comprehensive analyses that were previously challenging to undertake. For instance, in the healthcare sector, web scraping can be used to compile detailed information about hospitals, including their ratings, locations, and patient reviews, providing valuable insights into healthcare accessibility and quality across different regions.

However, as we harness the power of web scraping to gather this data, we encounter critical concerns regarding privacy and data security. As a result, it is necessary to effectively anonymize the sensitive data collected. Anonymization refers to the process of removing or altering identifiable information in a manner that preserves the overall dataset quality. This is essential for maintaining privacy, especially in the case of geospatial and demographic data.

Privacy-preservation is even more necessary in web scraping workflows, and the risk of potential misuse is higher. Thus, implementing effective anonymization strategies is crucial for preserving individual privacy and maintaining the usability of the data for analysis. By balancing data usability and privacy, researchers can protect sensitive information without compromising the integrity of their findings.

In this study, we explore cluster-preserving anonymization to preserve the integrity of hospital data scraped using online maps platforms in a way that maintains the overall geographic influence that various healthcare facilities exert on their surroundings. Thus, important insights, such as the distribution of quality healthcare across the country, can still be drawn without identifying a single facility.

II. Related Works

In this section, we go over related works in the Privacy Preserving Data Aggregation, Privacy-Preserving AI, Distributed Data Mining, Web Scraping, and Privacy-Preserving Clustering, outlining key methodologies that inform our approach to secure data handling, spatial analysis, and anonymization.

A. Privacy Preserving Data Aggregation

Recently, notable progress has been made in techniques aimed at safeguarding data privacy in multiple areas. J. Zhang et al. created LVPDA, a simple and verifiable data aggregation framework that maintains privacy for Internet of Things (IoT) environments. This approach utilizes a form of encryption and a specific signing method to ensure that the data remains secure and accurate while keeping the computational demands manageable [1]. Yan et al. introduced a data aggregation technique for fog-assisted mobile crowdsensing that emphasizes both privacy and reliability. Their solution allows users to verify the accuracy of collected data and safeguards the information against untrusted servers and fog nodes [2].

Zhou et al. presented EPDA, an energy-efficient method developed for wireless sensor networks. This method maintains data security while extending the operational lifespan of the network by organizing it in a tree structure

with connected leaf nodes [3]. Finally, Chang et al. tackled issues of privacy that are associated with smart meters in smart grids through their 3PFT scheme. This approach incorporates a mechanism for fault tolerance while minimizing computational requirements, utilizing a masking technique combined with secret sharing to maintain privacy during data aggregation [4]. Collectively, these studies highlight the importance of preserving privacy in data-collection across diverse technologies.

B. Privacy Preserving Artificial Intelligence

In the field of privacy-preserving artificial intelligence, various techniques have been created to tackle data privacy issues, especially in fields like healthcare. Khalid et al. conducted a detailed survey of privacy-preserving techniques used in AI for healthcare. They pointed out key methods like Federated Learning and Hybrid Techniques, which allow secure sharing of data while protecting patient privacy, an important requirement as a result of complex healthcare records and strict privacy laws [5]. Torkzadehmahani et al. also looked at privacy in AI for biomedical data, highlighting risks related to genome-wide studies. They organized different privacy-preserving approaches into categories and suggested that combining Federated Learning with other methods could improve privacy protection, although this approach requires more computational power [6].

Dodda et al. reviews various techniques on Federated Learning as a way to train models collaboratively while keeping data on local devices, which helps maintain privacy in AI development. They discussed important parts of this method, such as how model updates are combined as well as features for privacy necessary for data protection [7]. Qu et al. introduced a new training structure meant for edge computing. Their approach includes two main phases: a federated pre-training phase that involves both cloud and edge servers, which incentivizes edge contributions, and a model segmentation phase that uses homomorphic encryption to protect data [8]. Lastly, Aminifar et al. presented a method called k-PPD-ERT, which is a privacy-preserving version of extremely randomized trees for analyzing sensitive healthcare information. This approach keeps patient data safe while still enabling effective training of models on distributed datasets [9]. These studies reflect how important privacy-preserving techniques are for making progress in AI while protecting sensitive data.

C. Privacy Preserving Distributed Data Mining

In the area of privacy-preserving distributed data mining, several advancements have emerged to enhance data privacy while enabling effective data analysis. Merani et al. introduced a framework called Rings, which incorporates Multi-Party Computation (MPC) to improve privacy in large-scale data mining. Unlike traditional MPC methods that relied on a limited number of peers, this new approach uses a distributed setup in which every data provider is involved in the aggregation process. The framework addresses issues like data unavailability during network disruptions by employing multiple data sources termed as rings, ensuring system reliability and privacy throughout the process [10]. Q. Zhang et al. focused on the increasing amount of distributed medical data resulting from the Internet of Health (IoH) and proposed PDFM (Privacy-free Data Fusion and Mining) to enable secure retrieval of medical records. This method addresses the challenges of privacy when integrating cross-departmental data, showing that it can effectively support privacy-preserving searches and improve healthcare services [11].

Li et al. tackled the challenge of maintaining privacy in Distributed Privacy-Preserving Data Mining (DPPDM) by introducing a semi-supervised learning method that utilizes both labeled and unlabeled data from different sites. Their approach employs a parameter-masking Expectation-Maximization algorithm, allowing a site to learn without exposing individual data or enabling traceability [12]. In addition, X. Zhang et al. presented MRMondrian, a scalable multidimensional anonymization method designed for big data applications. By using the MapReduce paradigm, this approach recursively partitions data until each subset fits into the memory of individual computing nodes, using a tree indexing structure to support efficient recursive operations. It aligns with differential privacy principles, allowing for effective privacy management in large datasets while significantly improving the scalability of traditional anonymization methods [13]. Collectively, these studies underscore the significance of innovative privacy-preserving techniques in the realm of distributed data mining, highlighting their potential to balance data utility and privacy protection in various applications.

D. Web Scraping

In the realm of web scraping, Krotov et al. examined the often-overlooked legal and ethical challenges associated with web scraping, a practice increasingly employed in both industry and academic research. The work done underscores the importance of balancing the utility of automated data-collection tools with compliance to legal frameworks and ethical considerations. By reviewing legal and ethical literature, critical questions that researchers and practitioners must address to prevent controversies and potential lawsuits were highlighted. This signifies the necessity of responsible web scraping practices to ensure adherence to privacy laws and mitigate reputational risks for organizations [14]. Uzun introduced a new technique called UzunExt, which improves the efficiency of web scraping by using string-based methods instead of traditional DOM-based extraction, resulting in significant time savings [15].

Additionally, Dogucu and Çetinkaya-Rundel advocated for the incorporation of web scraping into statistics and data science education, highlighting its benefits in helping students collect real-world data more effectively. Activities were presented that link web scraping techniques to classical statistical concepts, making sure that students gain valuable experience with the entire data science workflow. While web scraping offers valuable opportunities, such as working with current and relevant datasets, challenges are prevalent for instructors, including the technical complexity and evolving web structures, offering strategies to mitigate these difficulties [16].

E. Privacy Preserving Clustering

In the area of privacy-preserving clustering, Majeed et al. introduced a novel anonymization technique that

effectively balances privacy and utility by considering both similarity and diversity during the clustering process. In contrast to traditional methods that focus solely on one aspect, their approach leverages machine learning algorithms that ensure effective anonymization, even in imbalanced clusters. Their results showed notable improvements, significantly reducing common privacy risks by 13.01% and AI-based risks by 24.3%, while also enhancing data utility by up to 20.21% [17].

In another study, Hu et al. developed a lightweight k-means clustering method specifically designed for industrial IoT environments. This method mitigates privacy risks by ensuring that no sensitive attributes are exposed during clustering, achieved through a secure and efficient initialization of cluster centers that adapts to changing data [18].

Additionally, Majeed et al. also provided a comprehensive review of various clustering-based anonymization mechanisms, categorizing them by data type and evaluating their strengths and weaknesses across different applications such as social networks and cloud computing [19].

Lastly, Bi et al. introduced PriKPM, a privacy-preserving k-prototype clustering scheme that employs splitting of data across two servers and additive secret sharing to secure mixed data processing in cloud environments. Their experimental results confirmed PriKPM's efficiency and accuracy, making it a viable solution for secure data outsourcing [20].

In our forthcoming sections, we propose a novel approach that lies at the intersection of web scraping, aggregation, and clustered anonymization.

III. Methodology

This section outlines the proposed approach. Our study aims to demonstrate how cluster-preserving anonymization techniques produce a result closer to original, unaltered data as opposed to random additive anonymization methods. We showcase this using the following methodology: dataset acquisition through web scraping of publicly available information, cleaning and modeling the data into a suitable representation, plotting histograms for original data, randomly anonymized data and cluster-preserving data anonymization to draw comparisons between the different techniques.

A. Dataset acquisition

The initial dataset used is extracted from a document published by the National Institute of Technology Jalandhar [21]. It contained the names, addresses, and the cities and states of a majority of hospitals across India. We further expand the dataset by utilizing additional data scraped for each hospital via an online maps platform. This includes the star ratings, number of ratings, and latitude and longitude coordinates.

Acquisition of said additional data is done by driving Microsoft Edge to automatically browse the maps platform pages of each hospital and perform web scraping of the aforementioned data using the Selenium framework. This is done by searching for the hospital name concatenated with its city and state, which often leads directly to its page on maps platform. In the occasional case where it leads to a search results page instead, we consider the first result. The first result in the search query will on occasion contain erroneous values, which can possibly introduce faulty values into the dataset. However, these deviations are insignificant to the point that their impact on the final results are negligible. After this stage of data extraction is complete, we will have the finalized dataset, to be cleaned and inferred from.

B. Data cleaning

On observing the acquired dataset closely, we notice occasional omissions exist in the tabulated dataset due to the nature of the web scraping process. For the purposes of our result, we resolve this issue by backward filling the coordinates of missing values where location data is unavailable. This is acceptable due to the inherent nature of the dataset, where hospitals are often situated close to each other in terms of location throughout the country. Additionally, we fill the missing values of ratings and number of ratings with the median value. This is generally safe as it will not significantly affect the overall influence at any given location.

C. Data modeling

A potential issue in the analysis of the dataset is that the raw rating alone is not an accurate metric of the relative prominence of a hospital in a region. This is illustrated in extreme cases where, for example, a hospital with a single 5-star rating could misleadingly appear more reputable than a hospital with a 4.5-star rating with thousands of reviews. We must therefore incorporate the rating R and number of ratings N into a single field that more accurately defines the influence of a hospital. We therefore define the *effective rating* η as:

$$\eta = \alpha * R * \log(N) \quad (1)$$

Where, α is an arbitrary coefficient used to normalize the magnitude of the effective rating such that it reasonably defines a radius of influence around the hospital.

With the effective rating computed for each hospital, we proceed to perform a preliminary analysis on the dataset. The cleaned and processed dataset is plotted as a bubble plot, where the geospatial coordinates serve as the x and y axes. The radius of the bubble corresponds to the effective rating, amplified to enhance visibility.

D. Data Representation

We now proceed to represent the influence of hospitals spatially, by using their scraped latitude and longitude values. This is done at a resolution of 0.01 degrees of the geospatial coordinates. The value of 0.01 degrees is chosen as it roughly corresponds to a 1x1 square kilometer grid which we found to be appropriate for the application of the model.

We fill each cell of the obtained grid map by applying calculated influence values of the hospitals. The influence

value for a given cell is calculated as a sum of the values of all hospitals that have influence over that particular cell. The influence value is calculated to be an exponentially decaying function of the distance to the hospital given by the piecewise function:

$$i = \begin{cases} ae^{-\lambda r}, & \text{if } r < \eta k \\ 0, & \text{otherwise} \end{cases}$$

where i is the influence value to be calculated, r is the Euclidean distance from the cell of interest to the hospital, while a, λ, and k are constants determining the influence scale, decay rate and maximum radius of influence, respectively[1].

Having represented all hospitals geospatially, we now identify cluster by hospital influence. First, we remove any insignificant values by considering only the 50th percentile and above of cells that previously had any influence at all. The remaining cells that are not empty can be grouped into clusters. In this context, a cluster is defined as a set of 8-adjacent cells with an influence value that are bounded by empty cells. We then calculate the magnitude of each cluster by summing up the influence values of each cell in the cluster. The logarithms of the influence values are plotted as a histogram. We treat this histogram as a distinctive signature of the dataset and that we can use to gauge the signatures of all datasets that have undergone anonymization. We can evaluate the effectiveness of the anonymization metric based on the similarity of its signature with that of the original dataset; both visually and quantitatively with measures such as the Earth Mover's Distance (EMD).

We now define our method of cluster-preserving anonymization, and contrast it with Gaussian noise addition. For performing cluster-preserving anonymization, we first need to identify the edge points of each cluster. We use the 'Quickhull' algorithm, as implemented in the SciPy Spatial library to compute the convex hull of each cluster. Upon identifying the clusters, we can compute the centroid of each one by averaging the coordinates of all points within the convex hull. From this centroid, we determine the nearest cell on the convex hull, which gives us a radius r that defines an inner circle within the cluster. This radius indicates the maximum distance a point can be moved within the cluster.

It is important for us to ensure that hospitals with high influence are not displaced too much from their original positions, as this could affect the overall influence in the region. To address this issue, we adjust the displacement based on a metric that reflects each hospital's influence. In this case, we have found that the square of the effective rating preserves the signature well.

We perform anonymization of the coordinates of each hospital by selecting a cell within the cluster and displacing the cell in a random direction by a random distance, from 0 to r scaled down by the square of the effective rating. We repeat this process an arbitrary number of times, or until it falls within the boundary of the cluster. We call this a valid displacement (Fig. 1).

Mathematically, anonymized geospatial coordinates x' and y' of a hospital with effective rating η and initial coordinates x and y is given by:

$$(x', y') = (x, y) + \hat{H} \cdot \frac{k}{\eta^2} \tag{3}$$

where $\hat{H}$ is a random unit vector and k is a random number between 0 and r.

Following which, we obtain the histogram of the cluster magnitude of this anonymized dataset as its data signature.

For the randomly anonymized data, we apply random additive noise to the geospatial coordinates of the hospitals. To somewhat preserve the original clustering after noise addition, we restrict the noise to follow a Gaussian distribution, with a mean and standard deviation based on the inner radii of all clusters of the original dataset. Specifically, we use

$$(x', y') = (x, y) + \mathcal{N}_{2D}(\mu_r, \sigma_r) \tag{4}$$

where $\mathcal{N}_{2D}$ refers to a random Gaussian 2-dimensional vector, μ_r is the mean of inner circle radii of all clusters in the dataset and σ_r refers to the standard deviation of these radii.

It is worth noting that, while there is definitely a noticeable deterioration of dataset quality after crude noise addition as shown in the forthcoming sections, restricting the Gaussian noise to the parameters of the cluster inner radii does not cause total degradation of the original dataset

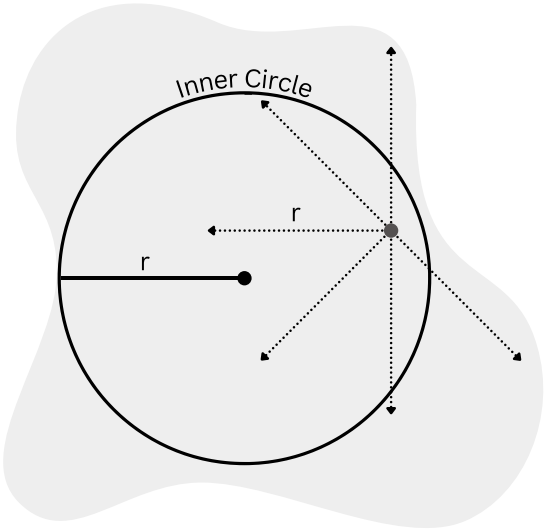

Fig. 1. A typical cluster is shown here, along with its inner circle of radius r. The dotted lines, each also measuring r, represent the valid displacements of a sample point within the cluster.

[1]A note on the arbitrary constants: After testing various values, we determined that setting α=0.1 in (1), and a=1, λ=0.01, k=0.1 in (2) yielded effective results for our model.

signature. This demonstrates the efficacy of the inner radius as a reliable metric in the anonymization of clustered data.

To achieve full anonymization, we obfuscate the names of each hospital, add Gaussian noise to the ratings and update the 'number of reviews' field such that η given by (1) is preserved.

IV. Results

In this section, we present the results of our analysis, demonstrating the effectiveness of the cluster-preserving anonymization technique.

We can first plot the scraped coordinates in a scatter plot (Fig. 2). Note that some points may eclipse others and the figure is in no way indicative of the number of hospitals in the country. Notably, there is a distinct clustering of points in the major metropolitan areas, specifically, Delhi, Kolkata, Mumbai, Hyderabad, Bengaluru and Chennai. Additionally, grouping can also be observed in regional population centers.

We plot the logarithms of the influence values for each cell in the grid in calculated as per (2), to emphasize areas with lower influence. The visualization (Fig. 3) highlights metropolitan regions as the most influential areas, where the density of influence is notably higher. In contrast, rural and less populated regions exhibit significantly lower influence levels. This pattern aligns with the conjecture that influence correlates with population centers.

The map indicating clusters with the top 50% of non-zero hospital influences is shown in Fig .4. These clusters are generated based on calculated effective ratings and their geospatial spread across different regions. Each cluster is shown as a distinct region or group within the plot, formed around points of influence concentration. In considering just the upper half of the values, we ensure that hospital influences are represented appropriately with respect to denser zones of influence. This map is used to compare the accuracy of the anonymized approaches.

Fig. 5. and Fig. 6. show clusters after various techniques of anonymization. Hospital Influence clusters after cluster-preserving anonymization are shown in Fig. 5. The clusters are anonymized based on the cluster-preserving anonymization method explained in the proposed approach, where cells within the cluster are displaced by the radius value r of the inner circle of the convex hull, preserving the clusters formed by the original data. By comparing the map to that pf the unaltered data in Fig. 4. we can observe that there is minimal, if any, change between the two datasets.

After performing random noise addition to the original dataset, hospital influences are shown in Fig. 6. The clusters are anonymized by applying random additive noise to the geospatial coordinates of the cells within each cluster, which results in the original dataset clusters not being preserved. This is evident when comparing the two anonymized map results, where there are noticeable differences in the structure, scale, and cluster distribution visually between the two maps, especially in the deccan region, in relation to the original dataset.

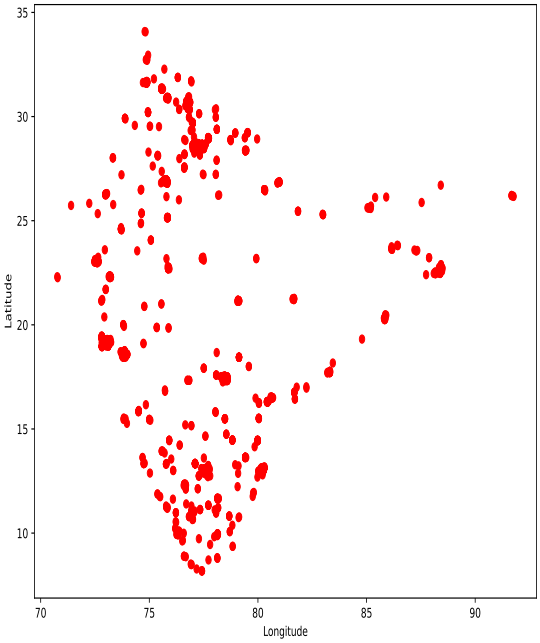

Fig. 3. Scatter plot highlighting the geospatial distribution of hospitals in India, with notable clusters in major metropolitan areas and regional population centers.

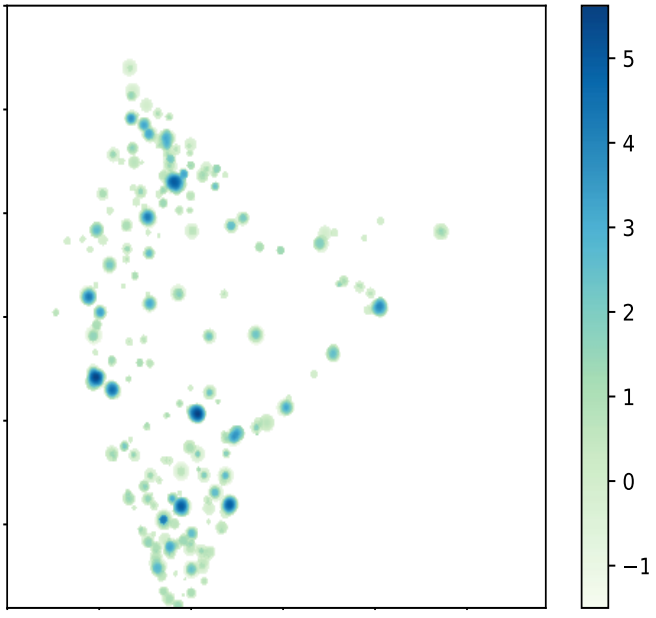

Fig. 2. Logarithmic visualization of influence values across the grid, highlighting metropolitan areas as regions of highest influence and illustrating the disparity in healthcare accessibility between urban and rural areas

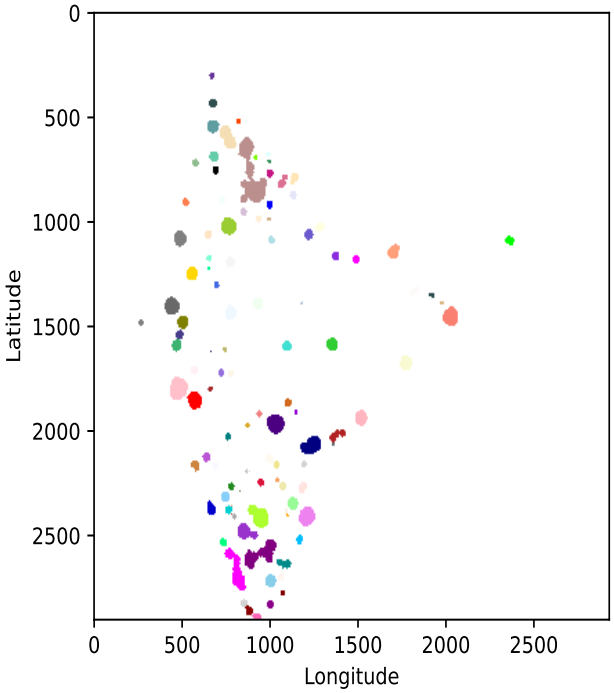

Fig. 4. Clusters representing the top 50% of hospital influences by effective rating, highlighting regions with concentrated healthcare presence for comparison with anonymized data.

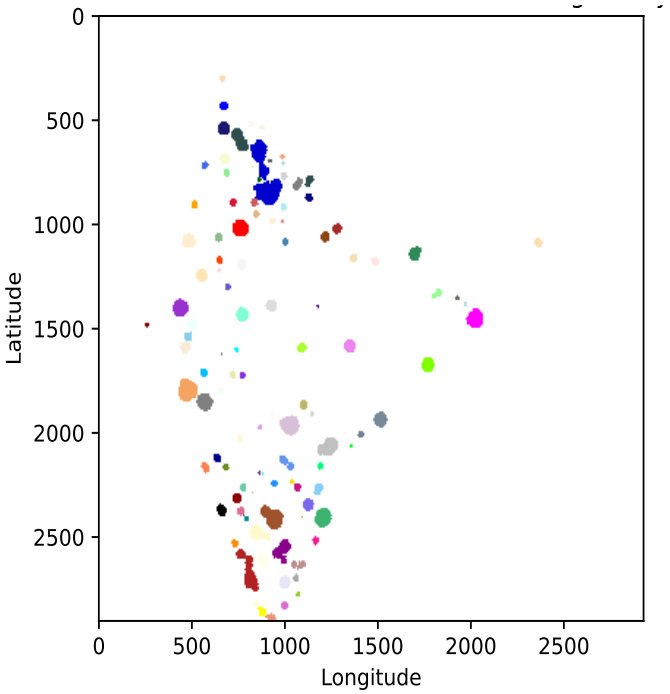

Fig. 5. Hospital-influence clusters after applying clustered anonymization, where cells are displaced within each cluster's inner circle radius. Minimal differences are observed when compared to the original data.

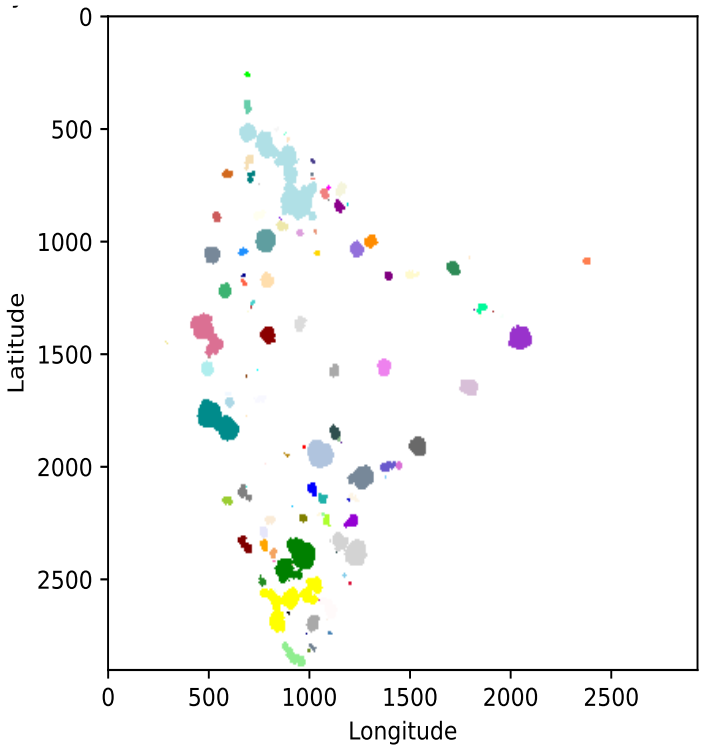

Fig. 5. Hospital-influence clusters after applying clustered anonymization, where cells are displaced within each cluster's inner circle radius. Minimal differences are observed when

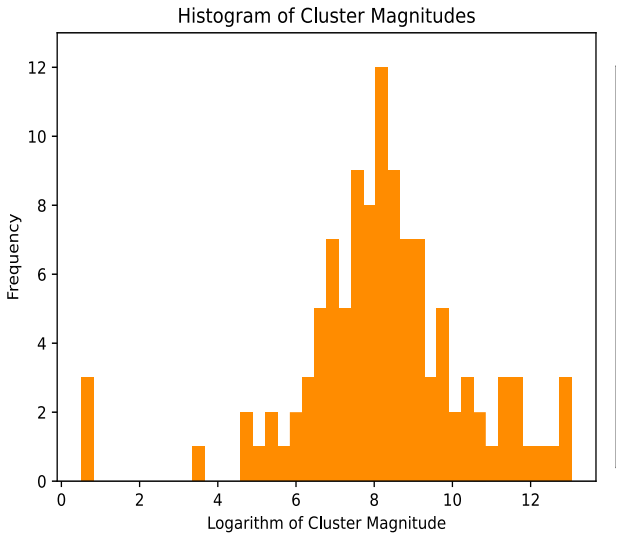

Fig. 7. Signature of the original dataset, depicting the distribution of hospital influence based on effective ratings.

The signature of the original dataset (Fig. 7.) provides a visual representation of the distribution of hospital influence. The central peak reflects the concentration of hospitals with similar influence values, while the tails indicate the presence of hospitals with either low or high effective ratings. This distribution serves as a baseline for comparison with anonymized datasets, highlighting the original data's inherent characteristics.

Fig. 8. shows the signature of the dataset anonymized using clustering-based methods. This histogram retains a similar spread and central peak height to that of the original dataset, indicating that clustering-based anonymization more effectively preserves the distributional characteristics compared to naive anonymization. To quantify this visual similarity, we found the Earth Mover's Distance (EMD) of this signature to that of the original dataset to be as low as 0.4002, further suggesting a close similarity in distribution.

Fig. 9. presents the signature of the naively anonymized dataset. Random noise is added to hospital locations as per (4). It is visually evident that the histogram shows a significantly larger spread compared to the original dataset, accompanied by a reduction in the central peak. The EMD between this signature and the original dataset was found to be much higher at 1.3015.

The EMD for both techniques are compared in Table 1.

TABLE 1. EARTH MOVER'S DISTANCE (EMD) BETWEEN ORIGINAL AND ANONYMIZED DATASETS

Dataset Type	EMD Value
Cluster preserving anonymization	0.4002
Naive Anonymization	1.3015

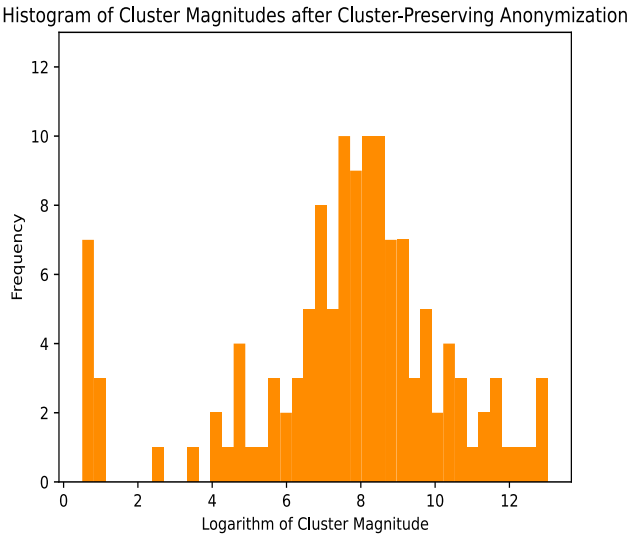

Fig. 8. Signature of the dataset anonymized using clustering-based methods, exhibiting similar spread, central peak height and overall shape to the original dataset, demonstrating effective preservation of cluster characteristics.

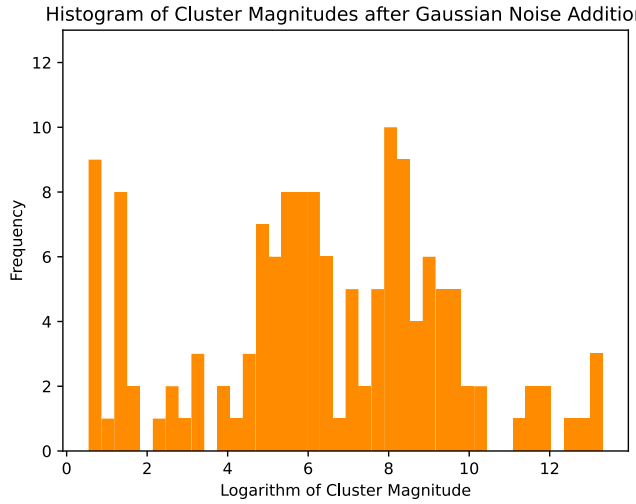

Fig. 9. Signature of the naively anonymized dataset, showing irregular patterns and reduced central peak compared to the original dataset.

V. CONCLUSION

In this study, we compiled a detailed dataset of hospitals in India by scraping information from various sources like online maps platforms. We collected important details such as hospital ratings, the number of reviews, and exact locations.

This dataset contains valuable information that lets us analyze how prominent and influential hospitals are in different areas. Given the sensitive nature of the geographical and demographic data, it is essential to anonymize the information to ensure privacy protection.

We explored different anonymization techniques to find a balance between keeping the data useful and ensuring

confidentiality. Our research highlights that clustering-based anonymization is a better method for preserving the natural patterns and density of the data. Unlike naive anonymization, which scatters data points and distorts their natural grouping, the clustering approach keeps the overall layout, density, and frequency of hospital locations intact. Our analysis shows that this technique helps maintain a distribution structure similar to the original dataset.

By successfully protecting privacy while keeping the integrity of the dataset, clustering-based anonymization proves to be a great option for scenarios where accurate spatial representation is key.

REFERENCES

[1] J. Zhang, Y. Zhao, J. Wu and B. Chen, "LVPDA: A Lightweight and Verifiable Privacy-Preserving Data Aggregation Scheme for Edge-Enabled IoT," *IEEE Internet of Things Journal,* vol. 7, no. 5, pp. 4016-4027, 2020.

[2] X. Yan, W. W. Y. Ng, B. Zeng, C. Lin, Y. Liu and L. Lu, "Verifiable, Reliable, and Privacy-Preserving Data Aggregation in Fog-Assisted Mobile Crowdsensing," *IEEE Internet of Things Journal,* vol. 8, no. 18, pp. 14127-14140, 2021.

[3] L. Zhou, C. Ge, S. Hu and C. Su, "Energy-Efficient and Privacy-Preserving Data Aggregation Algorithm for Wireless Sensor Networks," *IEEE Internet of Things Journal,* vol. 7, no. 5, pp. 3948-3957, 2020.

[4] Y. Chang, J. Li, N. Lu, W. Shi, Z. Su and W. Meng, "Practical Privacy-Preserving Scheme With Fault Tolerance for Smart Grids," *IEEE Internet of Things Journal,* vol. 11, no. 2, pp. 1990-2005, 2024.

[5] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha and J. Qadir, "Privacy-preserving artificial intelligence in healthcare: Techniques and applications," *Computers in Biology and Medicine,* vol. 158, p. 106848, 2023.

[6] R. Torkzadehmahani, R. Nasirigerdeh, D. B. Blumenthal, T. Kacprowski, M. List, J. Matschinske, J. Spaeth, N. K. Wenke and J. Baumbach, "Privacy-Preserving Artificial Intelligence Techniques in Biomedicine," *Methods of information in medicine,* vol. 61, no. S 01, pp. e12-e27, 2022.

[7] S. B. Dodda, S. Maruthi, R. R. Yellu, P. Thuniki and S. R. Byrapu Reddy, "Federated Learning for Privacy - Preserving Collaborative AI: Exploring federated learning techniques for training AI models collaboratively while preserving data privacy," *Australian Journal of Machine Learning Research & Applications,* vol. 2, no. 1, pp. 13-23, 2022.

[8] X. Qu, Q. Hu and S. Wang, "Privacy-preserving model training architecture for intelligent edge computing," *Computer Communications,* vol. 162, pp. 94-101, 2020.

[9] A. Aminifar, M. Shokri, F. Rabbi, V. K. I. Pun and Y. Lamo, "Extremely Randomized Trees With Privacy Preservation for Distributed Structured Health Data," *IEEE Access,* vol. 10, pp. 6010-6027, 2022.

[10] M. L. Merani, D. Croce and I. Tinnirello, " Rings for Privacy: An Architecture for Large Scale Privacy-Preserving Data Mining," *IEEE Transactions on Parallel and Distributed Systems,* vol. 32, no. 6, pp. 1340-1352, 2021.

[11] Q. Zhang, B. Lian, P. Cao, Y. Sang, W. Huang and L. Qi, "Multi-Source Medical Data Integration and Mining for Healthcare Services," *IEEE Access,* vol. 8, pp. 165010-165017, 2020.

[12] Z. Li, L. Yang and Z. Li, "Mixture-Model-Based Graph for Privacy-Preserving Semi-Supervised Learning," *IEEE Access,* vol. 8, pp. 789-801, 2020.

[13] X. Zhang, L. Qi, W. Dou, Q. He, C. Leckie and R. Kotagiri, "MRMondrian: Scalable Multidimensional Anonymisation for Big Data Privacy Preservation," *IEEE Transactions on Big Data,* vol. 8, no. 1, pp. 125-139, 2022.

[14] V. Krotov, L. Johnson and L. Silva, "Legality and Ethics of Web Scraping," *Communications of the Association for Information Systems,* vol. 47, 2020.

[15] E. Uzun, "A Novel Web Scraping Approach Using the Additional Information Obtained From Web Pages," *IEEE Access,* vol. 8, pp. 61726-61740, 2020.

[16] M. Dogucu and M. Çetinkaya-Rundel, "Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities," *Journal of Statistics and Data Science Education,* vol. 29, no. sup1, pp. S112-S122, 2021.

[17] A. Majeed, S. Khan and S. O. Hwang, "Towards Optimization of Privacy-Utility Trade-Off Using Similarity and Diversity Based Clustering," *IEEE Transactions on Emerging Topics in Computing,* vol. 12, no. 1, pp. 368-385, 2024.

[18] C. Hu, J. Liu, H. Xia, S. Deng and J. Yu, "A Lightweight Mutual Privacy Preserving k-Means Clustering in Industrial IoT," *IEEE Transactions on Network Science and Engineering,* vol. 11, no. 2, pp. 2138-2152, 2024.

[19] A. Majeed, S. Khan and S. O. Hwang, "Toward Privacy Preservation Using Clustering Based Anonymization: Recent Advances and Future Research Outlook," in *IEEE Access*, 2022.

[20] R. Bi, D. Guo, Y. Zhang, R. Huang, L. Lin and J. Xiong, "Outsourced and Privacy-Preserving Collaborative k-Prototype Clustering for Mixed Data via Additive Secret Sharing," *IEEE Internet of Things Journal,* vol. 10, no. 18, pp. 15810-15821, 2023.

[21] "List of Hospitals - Pan India - Dr. BR Ambedkar National Institute of Technology, Jalandhar," [Online]. Available: https://v1.nitj.ac.in/nitj_files/links/List_of_Hospital_-_Pan_India_28496.pdf. [Accessed 25 October 2024].