IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Clustering Based Anonymization for Privacy Preservation on Social Network

[1]Rutuja Prakash Jagtap, [2]Mukta Ganpat Warungase, [3]Chaitali Rajendra Narayane,[4] Kalpesh Bapurao Ahire and [5] Prof. Dhanajay M. Kanade

[1,2,3,4] Students, Department of Computer Engineering

[5]Assistant Prof, Department of Computer Engineering

[1]K. K. Wagh Institute of Engineering Education and Research, Nashik, India

Abstract: The current surge in social network popularity has generated massive amounts of data about social network interaction. Because these data contain many personal characteristics about individuals, anonymization is essential. Anonymization is a realistic technique for protecting consumers privacy when publishing data. This important step is required before final data can be used in research and data mining because it is no longer personal data. Because online social networks (OSNs) include sensitive information about unique members, it is necessary to anonymize network data before making it public. In this paper we proposed a novel method for privacy preserving. The proposed system is implemented in python on the dataset. Our system is evaluated with the evaluation parameter silhouette score and the results shows that our proposed approach outperforms.

Index Terms—Clustering, Privacy preserving, Social Networking.

I. INTRODUCTION

Recently, social networks [1,2] have received a lot of attention in research and development, partly because more and more social networks are being built online and the development of Web 2.0 applications. Social networks model social relationships using graph structures with vertices and edges. Vertices model individual social actors in a network, while edges model relationships between social actors. Social network analysis [3, 1, 4, 5] has become a crucial tool in contemporary sociology, geography, economics, and information science as a result of the explosive rise of social networks. Finding hidden social patterns is the aim of social network analysis. The effectiveness of social network analysis has been demonstrated to be substantially greater than that of conventional approaches that concentrate on evaluating the characteristics of individual social actors.

Privacy Preservation is a recent research area which consists of two major categories as given in Figure 1.1. One of the categories is Privacy Preservation in Data Mining (PPDM) and another is Privacy Preservation in Data Publishing (PPDP). In PPDM, after applying data mining functionalities mined patterns can be hidden from the third parties (intruders).

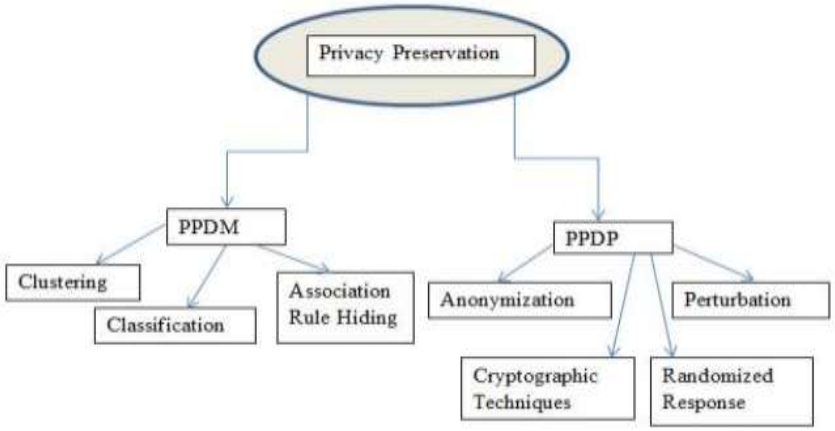

Fig. 1. Categories of Privacy Preservation Mechanism

There has been a lot of research done on relational data privacy preservation. Re-identifying people by combining a published table holding sensitive information with some external tables modelling attacker background knowledge is a significant category of privacy attacks on relational data. Numerous effective algorithms and significant models have been put out. However, the majority of current studies can only work with relational data. Straightforwardly using those procedures on social network data is not possible. Comparatively speaking, anonymizing social network data is substantially more difficult [6].

In this paper we present a clustering-based method for privacy preserving in social network. The proposed system is built in python with the help of various libraries. The rest of the paper is organized as, section II analyzed state of art systems, section III presents the proposed model explanation. Section IV shows the results and section V concludes the paper.

II. LITERATURE SURVE

In this section we present a deep literature survey on state of art systems for privacy preserving in social networks. The findings of the single pass k-means anonymization technique and the anonymized viewpoint of a data collection are the main topics of the [7] paper. Using generalization and suppression techniques, the dataset is made anonymous. Researchers in [8] analyzes the likely issues in these crucial areas of privacy, background knowledge, and data utility. It focuses on the current techniques for anonymization for maintaining the privacy of disclosing data on social networks, acknowledges the challenges associated with maintaining secrecy while publishing social network information. The foundation is provided by the clustering- and graph-based anonymization method.

The authors [9] presents a k-anonymity approach to minimize information loss during the generalization process for anonymized data, Since the clustering-based k-Anonymity technique employs separate anonymous sets of data and runs in O (n2/k) time, it is crucial to combine related data types into a single group. The author makes a useful comparison between their techniques and other clustering-based k-Anonymity techniques. The goal of this study is to gain an anonymized view of data without revealing any personal information about the users or their connections to other users. Researchers [10] offer examples of privacy protection challenges on social media. The sequential clustering-based anonymization process is presented in two different iterations by the author, starting with the centralized scenario.

A new method called slicing is designed by Li et al. (2012) [11] that divides the data horizontally and vertically. Slicing maintains data usage than generalization and is employed for membership disclosure protection. Slicing based privacy preserving micro data publishing is used to manage high- dimensional data. Slicing is employed for attribute disclosure protection and serves as an efficient algorithm for computing the sliced data with l-diversity needs. Slicing maintains better usage of data than generalization and bucketization in workloads with the sensitive attribute. Two new anonymization methods have been designed by Ghinita et al. (2011) [12] for sparse high-dimensional data. They depend on Nearest-Neighbor (NN) search in high-dimensional spaces which uses Locality-Sensitive Hashing (LSH). The slicing process is revisited by Vani and Jayanthi (2013) [13] and is utilized for the attribute disclosure protection.

A Privacy Preserving Framework for Supervisory Control and Data Acquisition (PPFSCADA) is Strategy-based permutation method introduced by Adil Fahad et al. (2014) [14] where the data privacy and data mining techniques are managed simultaneously. The designed technique includes the vertically partitioning original dataset for increasing the perturbation results. A framework is introduced with many network traffic data with arithmetical, definite and hierarchical attributes. It is also used for clustering the partitioned sets into many clusters depending on the designed framework. The perturbation process is realized through the variation of original attribute value with a new value.

[15] makes the assumption that the vertices are broken down into equivalence classes and that each class is appropriately anonymized using a relational data anonymization method that is already in use. Then, to more effectively anonymize the social network, examine whether edges should be included in the collapsed graph after condensing all of the vertices in an equivalence class into a single vertex. Publishing the number of edges of each edge type connecting two vertices in an equivalence class is one practical method. The method in question is known as cluster-edge anonymization.

III. PROPOSED APPROACH

In this section we present our proposed model architecture and its details. Figure 2 depicts the system architecture diagram. Our proposed model takes a dataset as an input. This data is preprocessed to remove redundant, null and unwanted data from the dataset. After preprocessing the dataset is splitted into training and testing dataset. Our proposed model performs several clustering techniques, including K-Member, Greedy K-Member, c-means, one-pass k-means and k-means clustering with generalization and suppression on a dataset. The dataset is first loaded using Pandas and unnecessary columns are dropped. Categorical variables are then encoded, and numerical variables are scaled using the MinMaxScaler method from sklearn.preprocessing.

K-Member

The first clustering technique applied is K-Member, which is performed using the KMedoids function from sklearn extra.cluster. The number of clusters is set to 5, and the random state is set to 0. The Silhouette Coefficient is then calculated using the silhouette score function from sklearn.metrics.

Greedy K-Member

The second clustering technique applied is Greedy KMember, which is performed using the SpectralClustering function from sklearn.cluster. The number of clusters is set to 5, the affinity is set to 'nearest neighbors', the number of neighbors is set to 5, and the labels are assigned using 'discretize'. The Silhouette Coefficient is then calculated using the silhouette score function.

C-means

The third clustering technique applied is c-means clustering using k-medoids, which is performed using the KMedoids function from sklearn.cluster. The number of clusters is set to 5, the metric is set to 'euclidean', the initialization is set to 'k-medoids++', and the maximum number of iterations is set to 300. The Silhouette Coefficient is then calculated using the silhouette score function.

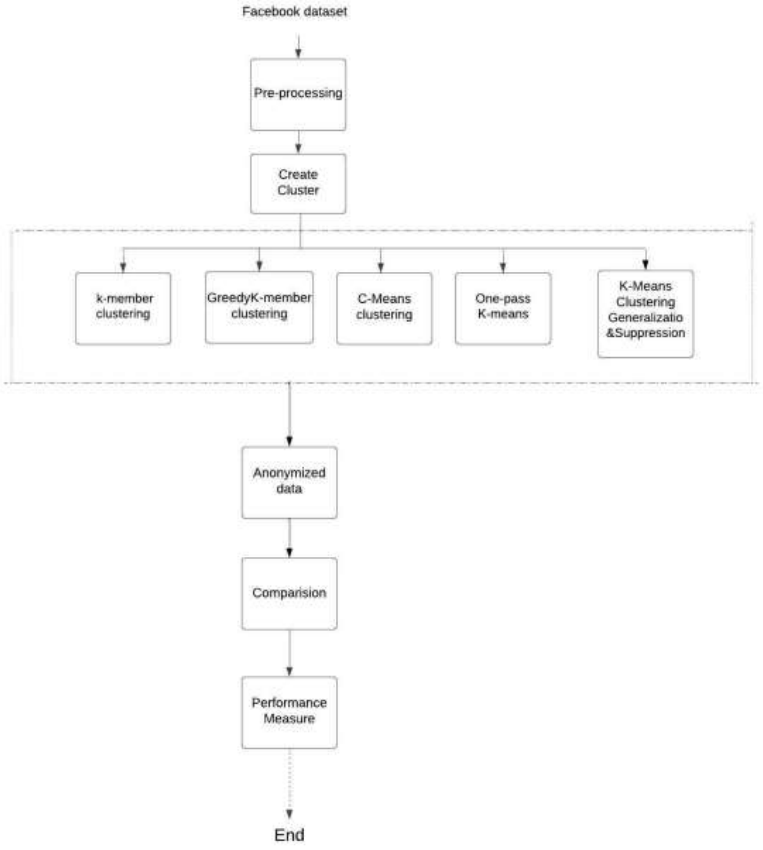

Fig. 2. System Architecture

One-pass k-means

The fourth clustering technique applied is one-pass k-means clustering, which is performed using the MiniBatchKMeans function from sklearn.cluster. The number of clusters is set to 5, the random state is set to 0, and the maximum number of iterations is set to 100. The Silhouette Coefficient is then calculated using the silhouette score function.

k-means clustering with generalization and suppression

The final clustering technique applied is k-means clustering with generalization and suppression. This is implemented using a for loop that performs the k-means algorithm multiple times, suppressing sensitive attributes and generalizing Quasi identifiers. The number of clusters is set to 5, the maximum number of iterations is set to 100, the convergence threshold is set to 1e-4, the suppression factor for sensitive attributes is set to 0.5, and the generalization factor for quasi-identifiers is set to 0.5. The sensitive attributes and quasi-identifiers are defined, and centroids are initialized by sampling from the dataset. The Silhouette Coefficient is not calculated in this final step. Finally, the anonymized dataset is output as a CSV file for each clustering technique. The output files include the cluster labels for each record.

IV. RESULTS

A. Dataset

To test the efficiency of our system we have downloaded our own dataset of Facebook users from Kaggle with features like gender, marital status, city, zip code, and country. The dataset consists of 1500 records.

B. Evaluation parameter Silhouette Coefficient

Silhouette Coefficient

The Silhouette Coefficient is a metric used to evaluate the quality of clustering results. It measures how well each data point in a cluster is separated from the other data points in the same cluster (cohesion) compared to how well it is separated from the data points in the neighboring clusters (separation). The Silhouette Coefficient ranges from -1 to 1, where a value of 1 indicates that the data point is well matched to its own cluster and poorly matched to neighboring clusters, a value of 0 indicates that the data point is on the boundary between two clusters, and a value of -1 indicates that the data point is poorly matched to its own cluster and well matched to a neighboring cluster. A higher Silhouette Coefficient value indicates a better clustering result. Figure 3 shows the comparative results analysis of our proposed system

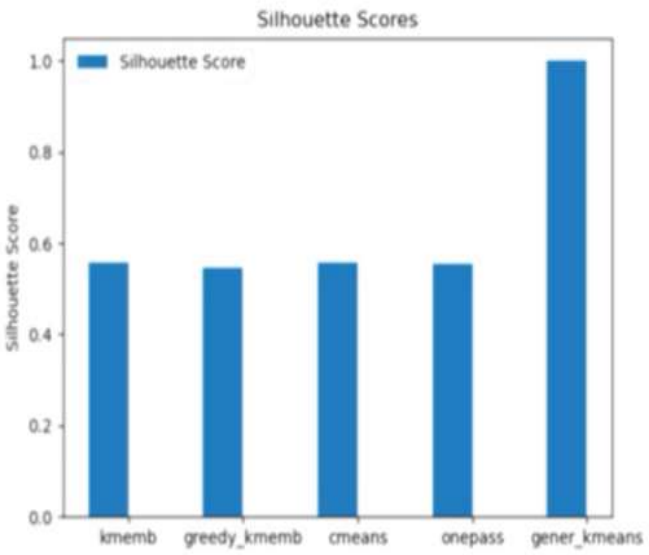

Fig. 3. Comparative analysis

V. CONCLUSION

In this paper we present a clustering-based approach for privacy preserving on social networks. By analysing Different Clustering Method, The Centroid-based Clustering i.e K Means algorithm is best because it it is straightforward and effective. By applying generalization and suppression process on dataset to get the anonymized view of data set. As social network data is much more complicated than relational data, privacy preserving in social networks is much more challenging and needs many serious efforts in the near future. Particularly, modeling adversarial attacks and developing corresponding privacy preservation strategies are critical

REFERENCES

[1] J. Scott. Social Network Analysis Handbook. Sage Publications Inc., 2000.

[2] B. Wellman. For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community. In Proceedings of the 1996 ACM SIGCPR/SIGMIS Conference on Computer Personnel Research (SIGCPR'96), pages 111, New York, NY, USA, 1996. A CM Press.

[3] L. C. Freeman, D. R. White, and A. K. Romney. Research Methods in Social Network Analysis. George Mason University Press, Fairfax, VA, 1989.

[4] S. Wasserman and K. Faust. Social network analysis: Methods and applications. Cambridge University Press, 1994.

[5] J. Srivastava, M. A. Ahmad, N. Pathak, and D. K.-W. Hsu. Data mining based social network analysis from online behavior. Tutorial at the 8th SIAM International Conference on Data Mining (SDM'08), 2008.

[6] B. Zhou and J. Pei. Preserving privacy in social networks against neighborhood attacks. In Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE'08), pages 506515, Cancun, Mexico, 2008. IEEE Computer Society.

[7] Rashmi B. Ghate, Rasika Ingle "Clustering Based Anonymization for Privacy Preservation" 2013

[8] Tamir Tassa and Dror J. Cohen, "Anonymization of centralized and distributed social networks by sequential clustering" IEEE Transactions on Knowledge and data Engineering, Vol. 25, pp. 311-324, Feb 2013.

[9] Jun-Lin Lin, Meng-ChengWei, " An Efficient Method for Kanonymization" Journal ACM 08 proceeding of International Workshop o Privacy and Anonymity in Information Society, pp. 46-50, 2008

[10] Bin Zhou, Jian Pei and Wo-Shun Luk," A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data", ACM Newsletter Journal, Vol. 10, pp. 12-22, December 2008.

[11] Li, T, Li, N, Zhang, J & Molloy, I 2012, 'Slicing: A new approach for privacy preserving data publishing', IEEE transactions on knowledge and data engineering, vol. 24, no. 3, pp. 561-574

[12] Ghinita, G, Kalnis, P & Tao, Y 2011, 'Anonymous publication of sensitive transactional data', IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 2, pp. 161-174.

[13] Vani, B., & Jayanthi, D. 2013, 'Efficient approach for privacy preserving microdata publishing using slicing'. International Journal of Research in Computer and Communication Technology, 4, 225

[14] Haghnegahdar, A, Khabbazian, M & Bhargava, VK 2014, 'Privacy risks in publishing mobile device trajectories', IEEE Wireless Communications Letters, vol. 3, no. 3, pp. 241-244.

[15] E. Zheleva and L. Getoor. Preserving the privacy of sensitive relationships in graph data. In Proceedings of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD (PinKDD'07), 2007