International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Privacy Preservation in Online Social Networks Using Graph-Attribute-Driven Optimal Clustering Algorithm

R. Suresh[1*], Dr. A. Devendran[2], Dr.V.N. Rajavarman[3]

Submitted: 26/01/2024 **Revised**: 04/03/2024 **Accepted**: 12/03/2024

Abstract: The surge in global online social network (OSN) users, especially post-COVID-19, has made it an integral part of daily life. As more rely on social networks, there's an increasing demand for privacy protection because these platforms store sensitive user information. This creates risks of potential intruders trying to access and misuse that information. The research aims to create a method for safeguarding privacy in OSN, minimizing the risk of sensitive information leaks, reducing processing time, and achieving high-level privacy preservation with minimal information loss (IL). The outcomes show that, in terms of privacy preservation, the suggested strategy performs better than cutting-edge techniques. The metrics employed, including anonymization degree, IL, and execution time (ET), demonstrate the high-level effectiveness of the proposed approach in achieving the specified privacy goals. The proposed approach effectively anonymizes OSNs through Graph-Attribute-Driven Optimal Clustering Algorithm (GADOCA); the research offers a practical solution to the ongoing challenges in ensuring k-anonymity, l-diversity, and t-closeness. The evaluation against real-world data adds credibility to the proposed method, emphasizing its potential to enhance privacy preservation in OSNs and contribute to the broader field of information security.

Keywords: *Privacy preservation, online social networks, Sensitive information, Risk mitigation, anonymization degree, information loss and execution time.*

1. Introduction

As of late, the global user base of OSN has witnessed a significant surge, particularly in the aftermath of the COVID-19 pandemic [1]. OSN have now ingrained themselves as an integral part of daily life for many individuals [2]. This increasing reliance on OSNs has also heightened the need for privacy safeguards to shield users from potential malicious activities [3]. Given that OSNs contain sensitive user information, there is a risk of intruders attempting to exploit and leak this data for various purposes, be it commercial or non-commercial. Consequently, the establishment of different privacy levels within OSNs has become imperative. Even though a number of user and network-level privacy preservation techniques have been presented, obtaining k-anonymity and fulfilling higher privacy model criteria [4], such as t-closeness and l-diversity within OSNs, remain a significant research issue [5].

The main issue this study found is that there hasn't been enough effort made to guarantee k-anonymity, l-diversity, and t-closeness in OSNs [6]. Conventional techniques often fall short [7] in addressing the intricate interplay of factors that threaten user privacy.

The study recognizes the need for a comprehensive and effective method to anonymize OSNs, considering the sensitivity of user data and the potential threats from malicious entities [8]. The aim of this study is to present a new technique for OSN anonymization [9], which is based on a clustering strategy that takes into account certain graph features. This method aims to achieve privacy preservation at different levels, specifically targeting edge, node [10], and user attributes within the OSN graph. Ensuring k-anonymity, l-diversity, and t-closeness in each cluster of the suggested model is the main objective [11]. Using a clustering approach for k-anonymization, a one-pass anonymization strategy for l-diversity and t-closeness privacy, and a data normalisation algorithm to enhance the quality of OSN data are the objectives of the research [12]. The research methodology is systematically structured to comprehensively address privacy preservation challenges in OSN [13]. First, a data normalization algorithm is meticulously designed to preprocess and refine raw OSN data, ensuring a standardized and optimized input. Subsequently, a clustering method, harnessing the power of multiple graph properties, is employed to categorize OSN data into distinct clusters while simultaneously guaranteeing k-anonymization.

We present a novel one-pass anonymization approach to significantly improve privacy protection within each cluster. By focusing on the l-diversity and t-closeness requirements in particular, this novel technique [14]

[1]*Research Scholar, Computer Science, Dr.M.G.R Educational and Research Institute, Chennai, India. Email: sureshrmsscmphil@gmail.com*
[2]*Professor, Dr.M.G.R Educational and Research Institute, Chennai, India. Email: devendran.mba@drmgrdu.ac.in*
[3]*Professor, Dr.M.G.R Educational and Research Institute, Chennai, India. Email: nrajavarrman2003@gmail.com*

guarantees a reliable and effective post-processing mechanism for the clusters. The effectiveness of the proposed methodology is then evaluated through rigorous performance assessments against state-of-the-art approaches, employing a real-world dataset sourced from Yelp. Quantitative metrics, including anonymization degree, IL, and ET, serve as crucial indicators for assessing the proposed method's efficacy [15]. These metrics offer a quantitative perspective on the balance achieved between privacy enhancement and computational efficiency. The empirical evaluation utilizing a "Yelp real-world dataset" provides a practical context, enabling a meaningful comparison and validation of the proposed methodology's practical applicability. The significance of this research lies in its contribution towards addressing the critical issue of privacy preservation in OSNs. By proposing a novel method that effectively anonymizes OSNs through GADOCA, the research offers a practical solution to the ongoing challenges in ensuring k-anonymity, l-diversity, and t-closeness [16, 17]. The evaluation against real-world data adds credibility to the proposed method, emphasizing its potential to enhance privacy preservation in OSNs and contribute to the broader field of information security.

2. Literature review

The study by Gangarde et al. (2021) highlights the growing global demand for online social networks (OSNs) and the need for privacy preservation methods. They propose a novel method using multiple-graph-properties-based clustering to achieve k-anonymity, l-diversity, and t-closeness in OSNs. The method is evaluated utilizing a "Yelp real-world dataset" and shows high-level privacy preservation compared to state-of-the-art methods [18].

Singh et al. (2022) discuss the risk of leakage of sensitive information on social networks. They propose a technique to prevent disclosure and reduce noisy nodes by perturbing sensitive attributes. The technique is evaluated using metrics like APL, ACSPL, RRTI, number of noisy nodes, and IL, and shows that it preserves sensitive attributes with minimal loss of information, thereby preserving the utility of published data [19].

Gangarde et al. (2022) highlighted the importance of securing users' privacy on social media networks. They highlighted the potential for ethical and unethical users to misuse their information, highlighting the need for anonymization before public publication. On the Adult dataset, the suggested Mondrian algorithm which improves privacy and utility balance is contrasted with metrics like PIRL, NCP, and DP [20].

In 2022, Mehta et al. discussed the importance of data privacy in big data analytics. Traditional anonymization methods like k-anonymity and l-diversity are insufficient for large data sets. They proposed an improved scalable l-diversity (ImSLD) approach, which improves running time and lower IL compared to existing methods. Because of the careful record organisation in the original equivalency class, this method preserves the same level of secrecy [21].

2.1. Research Motivation

The existing studies reveal persistent challenges in privacy preservation for OSN emphasizing the need for solutions that address minimum sensitive structural IL, provide high-level privacy protection, and minimize complexity. Current state-of-the-art methods, including cryptography-based, grouping/clustering-based, swarm-intelligence-based clustering, and other OSN graph-based approaches, exhibit limitations. Cryptography-based methods offer secure communication but lack comprehensive privacy across all OSN elements and depend on trusted authorities. Grouping/clustering-based methods show promise but struggle with performance and privacy tradeoffs. Clustering based on swarm intelligence only reaches k-anonymity at a greater computational complexity. Some OSN graph-based approaches fail to achieve privacy in all components (nodes, edges, attributes), leading to IL. Differential privacy and customizable privacy schemes have limitations, with the former assuming uniform privacy demands and the latter triggering unexpected relationships, compromising privacy preservation and leaking sensitive information. This paper's unique model is motivated by these inadequacies in existing approaches.

3. Proposed Methodology

In order to improve privacy protection in OSNs while lowering IL and computational costs, this study presents the GADOCA technique. The suggested anonymization scheme for OSN privacy is based on a three-phase model: as shown in Figure 1, the first phase involves normalizing input OSN data and initial clustering; the second phase, known as cluster optimization, refines clusters utilizing graph properties like eccentricity and distance to ensure k-anonymity with little information leakage; and the third phase, known as privacy preservation, further refines clusters to ensure l-diversity and t-closeness. Applying these concepts helps mitigate the risk of privacy breaches and unauthorized disclosure of personal information in OSN.

The raw data for OSNs includes user information, attributes, and connections. Data normalization is a crucial step, transforming data through statistical operations to standardize node sets and identify missing or messy data. This enhances data quality and ensures anonymization accuracy.

Once the data normalization is complete, the next stage involves forming the initial clusters. This is accomplished by utilising the fuzzy c-means algorithm. The fuzzy c-means algorithm is utilized to group nodes into clusters, taking into account the attributes of each node. This initial clustering sets the foundation for subsequent optimization phases, contributing to the overall effectiveness of the proposed method in achieving privacy goals within OSNs. The following are the three main justifications for choosing fuzzy c-means:

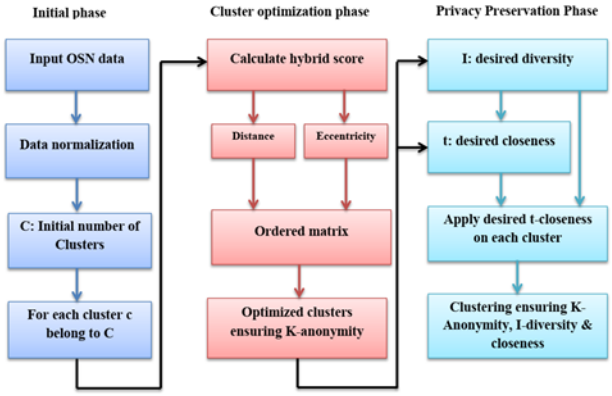
1. Accommodates Ambiguity: Fuzzy c-means handles uncertain or ambiguous data, allowing nodes to belong to multiple clusters with varying degrees of membership.

2. Flexible Cluster Formation: Allows nodes to have partial membership in multiple clusters, providing flexibility in grouping based on user attributes and connections.

3. Robust to Initialization: Less sensitive to different starting conditions, ensuring stable and reliable initial clusters for subsequent optimization phases in the proposed privacy preservation model for OSNs.

The research optimizes clusters using distance measures and graph properties to enhance reliability and privacy protection. A post-processing step ensures k-anonymity and fulfills l-diversity and t-closeness privacy notions. To describe the issue and lay out the suggested fix for OSN privacy preservation, a system model is provided. This structured approach provides a comprehensive understanding of the research process.

Fig 1: Suggested system model strategy for OSN privacy protection



3.1. Data Normalization

The initial phase of the proposed model is pivotal for refining raw input data obtained from OSN. This critical stage involves two key procedures: data normalization and the formation of initial clusters through the utilization of the fuzzy c-means algorithm. In the context of OSNs, where datasets often consist of a substantial number of users with diverse attributes reflecting connections and behaviors, the raw data is susceptible to outliers and messy entries. To address these issues and establish a clean, standardized dataset for subsequent analyses, data normalization becomes essential. This process employs statistical modeling techniques to enhance OSN data quality, involving the extraction of user profiles, calculation of various attributes, and aggregation into sets of vertices and attributes. The significance of data normalization lies in its capacity to elevate dataset quality, ensuring a standardized representation of attributes. This structured dataset, characterized by normalized and meaningful attributes, forms a foundational resource for subsequent steps in the proposed model. Particularly crucial in the context of privacy preservation within OSNs, a clean dataset facilitates the effective implementation of privacy protection measures. Eventually, the initial phase serves as a vital preprocessing step, laying the foundation for subsequent stages, including the creation of initial clusters and optimization phases. This progression aims to attain the overarching objective of preserving privacy in OSN.

3.2. Fuzzy C-means

The clustering algorithm relies on normalized inputs of vertices (V) and attributes (A) to accurately compute initial cluster centroids and their cluster members (CMs). This process aids in reducing sensitive information by utilizing normalized attribute values. Subsequently, the initial clusters are formed using the Fuzzy C-Means algorithm follow Equation (1):

$$CM_{ij} = \mu_{ij} \times \left(A_j - V_i\right) \quad (1)$$

Here, μ_{ij} represents the degree of membership of vertex V_i in cluster C_j. Unlike K-means, which generates initial clusters based on the mean of user attributes, Fuzzy C-Means allows for more flexible cluster formation by considering the degree of membership of each vertex in multiple clusters. Each cluster C_i, where i limits from 1 to the total number of clusters c, is required to have at least k users to satisfy k-anonymity in the network. Importantly, at the initial level, the value of k may vary for each

cluster, meaning different clusters can have different numbers of users. Let us consider a dataset where there are 100 vertices/users and 4 clusters altogether. Applying the Fuzzy C-Means algorithm ensures flexible cluster sizes based on the degree of membership. This approach addresses the limitation observed with K-means clustering, where k-anonymity was not consistently achieved across all clusters. To guarantee consistent cluster sizes, each cluster is further optimized by utilizing various graph properties, including eccentricity and distance.

3.2.1. Cluster Optimization Phase

The principal goal of the suggested approach is to achieve k-anonymity by means of clustering on preprocessed OSN data. However, as was previously said, k-anonymity is not guaranteed when utilizing Fuzzy C-means. The parameter K signifies that each cluster should have a minimum of K anonymous users to mitigate information leakage. Variations among cluster members are problematic as they may outcome in sensitive IL. For instance, consider two clusters C1 and C2 with K users and p = K + 10 users, correspondingly. In this case, there is a disparity in the anonymity levels between the clusters because the vertices in C1 are K-anonymous and the vertices in C2 are p-anonymous.

In order to minimize IL and ensure uniform cluster sizes, the cluster optimization phase is introduced to handle this difficulty. Two graph features are utilized to achieve this reorganization: the eccentricity of each vertex and the distance between the attributes of two vertices, which are used to evaluate each user's score inside each cluster. The constraint (n/k), where n is the total number of users and k is the number of clusters, must be followed by the number of users in a cluster. Interestingly, this creative approach represents the first attempt to use several graph features to protect every OSN graph element's privacy.

The distance and eccentricity graph attributes for each vertex or user in the current cluster are utilized to calculate the hybrid score. The Cluster Optimization Algorithm describes what happens during the cluster optimization stage. It includes two main tasks: calculating the hybrid score matrix and then optimizing the clusters.

Cluster Optimization Algorithm:

Inputs:

C: Set of clusters with its centroid, c: Number of clusters, A: Set of attributes, n: Total number of vertices in the network

Outputs:

SD: Sorted users list according to hybrid score

C: Optimized clusters ensuring K-anonymity

Steps:

```
D ← ones(n, 2)
m = 1
for i = 1 : c
        for j = 1 : size(Ci)
                uid ← CMi(j)
Huidi ← getScore(Auid, Ai_cent)
D(m, 1) ← uid
D(m, 2) ← Huidi
                m ← m + 1
end for
end for
SD ← Sort(D(:, 2), "ascending")
Cluster Optimization:
for i = 1 : n
        for j = 1 : c
if (status(SD(i :, 1)) != assigned) && (length(Cj) ≤ n/c)
Cj ← join(SD(i :, 1))
status(SD(i :, 1)) ← assigned
end if
end for
end for
```

We take advantage of graph features like eccentricity and distance to calculate the hybrid score for each vertex or user. We measure the distance between each user and the cluster centroid for the distance property. In the meantime, each vertex's maximum connections are measured in order to establish the eccentricity attribute. We evaluate the number of edges between two vertices in the context of measuring the distance property, with the lowest possible number of edges as the outcome. Essentially, in the given circumstance, the distance property records the greatest degree of similarity or closeness between the characteristics of two vertices.

Conversely, each vertex in the network has a maximum number of connections indicated by the eccentricity feature. In our example, we employ an attribute more precisely, the number of friends to determine each user's or vertex's eccentricity (NF^{uid}). Because it makes it easier to group vertices according on their connections, the eccentricity feature guarantees a balanced and dependable approach to clustering with little possibility for IL. Taking into consideration both eccentricity and distance features, we use a weight-based approach to produce a comprehensive hybrid normalized score. In the end, we

arrange every vertex into an ascending hybrid matrix, where a sorted list is represented by the first column. This hybrid scoring system effectively captures both proximity to the cluster centroid and the network connections of each vertex, offering a robust measure for vertex/user evaluation.

Size (C^i) indicates the number of CMs in the i^{th} cluster, while H_i^{uid} reflects the hybrid score value of the user/vertex uid of the i^{th} cluster. A^{uid} denotes the attributes of the j^{th} user/vertex, and A_{cent}^{i} signifies the attributes of the centroid of the i^{th} cluster. According to the algorithm, the getScore (.) function computes the hybrid value for each user/vertex in every cluster, bypassing the attribute of the j^{th} CM of the ith cluster and A_{cent}^{i}. Before that, we first obtain the vertex ID i.e., uid to obtain its corresponding set of attributes. Utilizing the getScore (.) function, we measure the two graph properties distance and eccentricity, as discussed above. The distance between A^{uid} and A_{cent}^{i} is computed by Equation (2):

$$d^{uid,cent} = \frac{\sum_{r=1}^{R} \left| a_r^{uid} - a_r^{cent} \right|}{R} \quad (2)$$

Where a_r^{uid} represents the r^{th} attribute of vertex uid, and a_r^{cent} represents the r^{th} attribute of the centroid vertex of the current cluster. R is the total amount of properties that every vertex in the network has. Divide the total distance of all the attributes by the total number of attributes, and utilize the absolute dissimilarity between the two attributes as the shortest distance between them.

The second graph property the eccentricity of each vertex A^{uid} is computed from its value of NF^{uid} by Equation (3):

$$e^{uid} = \left(\frac{1}{NF^{uid}} \right) \times \lambda \quad (3)$$

The minimum value of e^{uid} represents the maximum eccentricity of the vertex. The eccentricity outcome's scaling factor is represented by the symbol λ. Compute this scaling factor by taking the mean of NF^{uid} of all of the vertices, using Equation (4):

$$\lambda = \frac{\sum_{i=1}^{n} NF^{i}}{n} \quad (4)$$

The normalized eccentricity score of every vertex is guaranteed by the scaling factor. Each vertex's hybrid score is calculated utilizing a weighted approach in Equation (5):

$$H_i^{uid} = \left(w1 \times d^{uid,cent} \right) + \left(w2 \times e^{uid} \right) \quad (5)$$

where each graph property's weights are denoted by w1 and w2. Both weight parameters should have a value of w1 + w2 = 1. We give the eccentricity and distance parameters in this study similar weights, w1 = 0.5 and w2 = 0.5. The two vertices with a minimum value of H_i^{uid} represent more closeness or similarity between them. This increases the likelihood of more comparable vertices forming clusters. Each vertex and its hybrid score is stored individually in matrix D. The vertices in matrix D are then sorted into matrix SD in ascending order based on their hybrid score values.

3.3. Privacy Preservation Phase

Optimized clusters ensure k-anonymity in OSN, but l-diversity addresses threats like attribute disclosure, background knowledge, and homogeneity, ensuring comprehensive privacy preservation [22]. However, l-diversity itself presents difficulties, as preventing attribute disclosure is insufficient and often unnecessary to achieve. Recognizing these issues, the authors propose the use of t-closeness to address the drawbacks of both k-anonymity and l-diversity. This paper introduces a unified approach called the "one-pass algorithm" to address privacy preservation notions. It extends optimized clusters to ensure k-anonymity, l-diversity, and t-closeness. Standard definitions for l-diversity and t-closeness are provided before delving into the algorithm.

l-diversity Definition: If an equivalence class has at least l well-represented values for the sensitive property, it is said to have l-diversity. If each equivalency class in a table has l-diversity, the table is said to have l-diversity.

t-closeness Definition: If there is a threshold t that separates the distribution of a sensitive attribute within an equivalency class from the attribute's distribution throughout the entire table, the class is considered t-close. If every equivalency class in the table satisfies the t-closeness requirement, the table is said to have t-closeness.

While l-diversity addresses background knowledge and homogeneity attacks, it may not sufficiently mitigate attribute disclosure. On the other hand, t-closeness resolves all the issues associated with l-diversity but falls short in dealing with identity disclosure. Meanwhile, k-anonymity effectively tackles identity disclosure.

To address these considerations comprehensively, the proposed approach involves developing a common technique for extending k-anonymized clusters with both l-diversity and t-closeness. This is achieved through the one-pass algorithm, designed to post-process the optimized clusters. One-pass Privacy Preservation Algorithm outlines the functionality of the one-pass algorithm, ensuring t-closeness with a predefined threshold value (t) and subsequently applying the l-

diversity using the entropy-based method. Thus, the goal of the one-pass approach is to achieve total privacy protection within OSNs.

One-pass Privacy Preservation Algorithm

Inputs:

C: Optimized clusters ensuring K-anonymity

c: Number of clusters, A: Set of attributes

l: Desired diversity

Output:

C: Clusters ensuring l-diversity and t-closeness

Steps:

```
for i = 1 : c
for j = 1 : length(Ci)
uid ← Ci(j)
temp ← getDist(Auid, A ← Ci)
T(j, 1) ← uid
T(j, 2) ← temp
end for
t ← mean(T)
% Anonymize all users in clusters using t-closeness:
        for i = 1 : length(T)
if (T(i, 2) < t)
 L1 ← join(T(i, 1))
            else
             L2 ← join(T(i, 1))
   end if
  end for
Ci ← append(L1, L2) % Return the t-closeness anonymized cluster
LDi ← getDiversity(Ci)
  end for
 while (diversity(LD) < 1) do
Max ← cluster with maximum diversity value from LD
Min ← cluster with minimum diversity value from LD
Temp ← Max + Min
 C ← C - {Max, Min} + Temp
 end while
```

An easy approach for achieving l-diversity and t-closeness for input k-anonymized clusters is the One-pass Privacy Preservation Algorithm. The algorithm groups users within each cluster according to their dynamically computed t-value, which comes after the specification of t-closeness. The Earth Mover's Distance [23] is a useful tool for calculating the t-value. As described in One-pass Privacy Preservation approach, the approach uses the 'getDist(.)' function to guarantee that there is an equal distance between every user and every other member of the current cluster. If we suppose that the attribute set of the current user uid is A^{uid}, and the attribute set of all members of the same cluster is represented as $A \leftarrow C^i$, then the equal distance can be computed in the temp variable in Equation (6), as follows:

$$temp = \frac{1}{2}\sum_{j=1}^{q}\left|A^{uid} - A \leftarrow C^{i}(j)\right| \tag{6}$$

In this methodology, matrix T serves as a repository for Earth Mover's Distance values calculated for all cluster members. The t-value is determined by taking the mean of all the distances in matrix T, which helps to anonymize users inside a cluster. Users express varying levels of satisfaction or dissatisfaction with the t-value. Ultimately, users from both satisfaction and dissatisfaction lists are combined to reconstruct the cluster, ensuring resilience against similarity attacks and attribute disclosure threats.

Following t-closeness anonymization, clusters are expanded to adhere to the l-diversity privacy notion, leveraging the entropy l-diversity concept [24]. For each t-closeness-anonymized cluster, diversity is computed using entropy, and the outcomes are stored in the matrix LD. The application of a greedy algorithm guarantees that each cluster satisfies the l-diversity criterion. This iterative process continues until all clusters successfully achieve l-diversity.

Notably, throughout the entire one-pass algorithm, no users are eliminated from the cluster, preserving the privacy notion of k-anonymity for the optimized clusters. This comprehensive approach ensures a multi-faceted privacy preservation strategy, encompassing k-anonymity, t-closeness, and l-diversity within the optimized clusters. The suggested clustering approach minimises computational complexity while balancing privacy for users and vertices. It validates clusters against privacy preservation notions k-anonymity, l-diversity, and t-closeness, followed by edge anonymization within OSN. The algorithm is used for the designed model, as proposed in [23].

Each cluster is represented by the cluster head node, and edge anonymization is achieved through the computation of super-edges. By achieving k-anonymity, l-diversity, and t-closeness, this all-encompassing approach guarantees privacy protection and reduces dangers.

4. Results and Discussion

The experimental phase involved a comprehensive performance analysis of the proposed model compared to state-of-the-art approaches utilizing MATLAB on a Windows 10 operating system and an Intel I3 processor with 8 GB of RAM, utilizing a real-life dataset from Yelp [24]. An experimental analysis was conducted utilizing the Yelp dataset, which is a customer review set. Two files, focusing on friends and users, were used to analyze user attributes and the network of connections among users. These files, including user ID and 18 associated attributes, provided crucial data for a thorough examination of the methods' performance. Three important performance metrics were used to compare the suggested GADOCA methodology to cutting-edge techniques: ET, IL, and degree of anonymization (DoA). The ET is the mean amount of time required for each of the 25 scenarios to complete the anonymization process for OSN data. We calculated the total number of users assigned to each user's cluster in order to get the DoA for that user. Stated otherwise, the DoA of the user is comparable to that of its corresponding cluster.

Therefore, the DoA in Equation (7) is denoted as,

$$DOA = \deg ree\left(C_{u_i}\right) \times i \quad (7)$$

Where, Cu_i represents the DOA of user u_i that belongs to cluster C.

We utilised the methodology presented in Equation (8) [25] to compute the IL metrics:

$$IL = \frac{SSE}{SST} \quad (8)$$

In this case, SST is the sum of squares between clusters, and SSE is the sum of squares within the cluster.

4.1. Differences in Cluster Sizes

The study explores performance analysis with an emphasis on cluster size fluctuations, namely the number of clusters that are produced. During the experiment, the maximum user count was fixed at 10,000 while the cluster size parameter was varied between 20 and 100. Given the scale of the dataset, a minimum of 20 clusters was considered necessary. This experimental study set out to determine how three important performance metrics: DoA, IL and ET were affected by differences in cluster size. Upon analyzing the outcomes, a notable observation emerged. With an increase in the cluster size, the level of anonymization exhibited a decrease. This pattern can be explained by the observation that fewer clusters typically retain a higher proportion of K-anonymous users. The Figure 2 presents a comprehensive analysis of four distinct clustering techniques PSO-GA, LECC, Multiple Graph Properties-Based Clustering (MGPC), and GADOCA across varying numbers of clusters. Notably, as the number of clusters increases, PSO-GA consistently demonstrates improved performance, achieving lower values in the DoA. LECC and MGPC show enhanced anonymization, while GADOCA consistently outperforms them across all cluster configurations, demonstrating its effectiveness in achieving higher levels of anonymization in OSN, despite increasing cluster complexity.

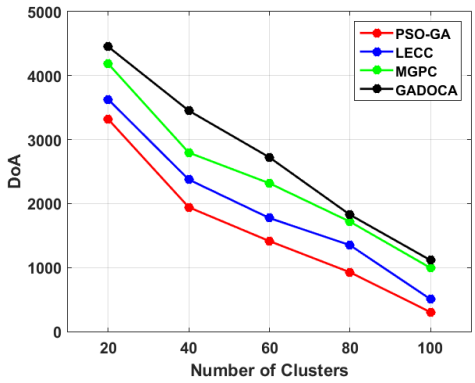


Fig 2: Impact of Varying Cluster Numbers on DoA Performance

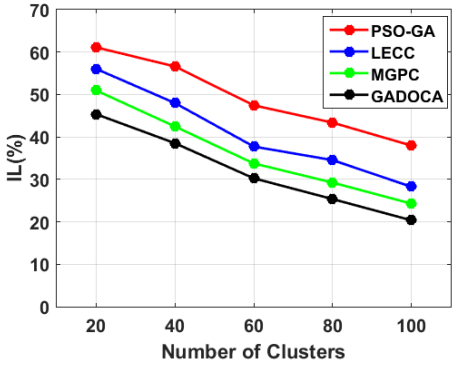


Fig 3: Impact of Varying Cluster Numbers on IL Performance

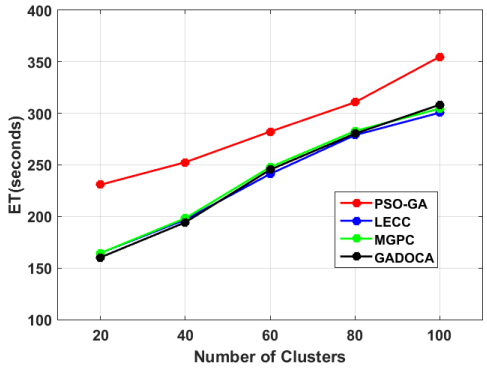


Fig 4: Impact of Varying Cluster Numbers on ET Performance

The Figure 3 demonstrates the impact of varying cluster numbers on IL performance for four clustering techniques PSO-GA, LECC, MGPC, and GADOCA. GADOCA consistently outperforms other methods in preserving information with lower IL scores, while PSO-GA and LECC show improved preservation with increasing cluster numbers. In essence, the results emphasize GADOCA robust performance in minimizing IL across varying cluster configurations. In Figure 4, four different clustering techniques: PSO-GA, LECC, MGPC, and GADOCA—across various cluster topologies are compared for execution time and ET. Notably, as the number of clusters increases, PSO-GA consistently exhibits higher ET, suggesting a moderate increase in computational complexity. GADOCA consistently outperforms LECC and MGPC in achieving faster ET, making it a promising technique for OSN data anonymization, despite increasing cluster numbers.

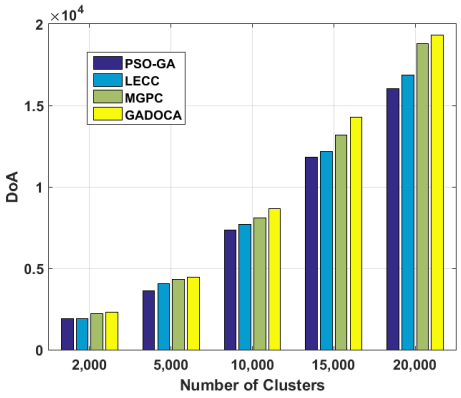


Fig 5: Impact of Varying Numbers of user on DoA Performance

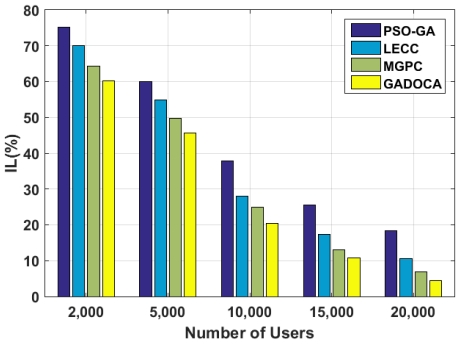


Fig 6: Impact of Varying Numbers of user on IL Performance

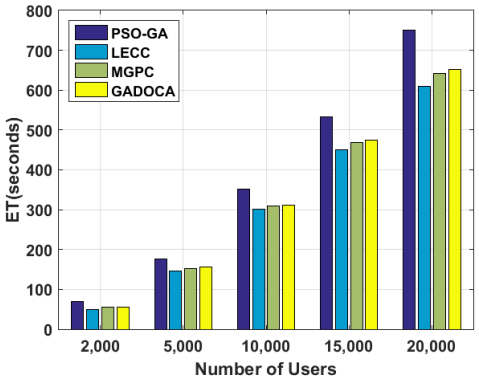


Fig 7: Impact of Varying Numbers of user on ET

4.2. Variations in Density

This analysis compares four clustering techniques: PSO-GA, LECC, MGPC, and GADOCA, varying user density from 2000 to 20,000. The study aims to assess their scalability and robustness, as higher user density leads to more significant clusters, affecting the overall performance of the algorithms. By exploring user densities from 2000 to 20,000, the research conducted a thorough investigation into the methods' effectiveness. Figures 5 to 7 present the outcomes of this experiment, illustrating the impact of user density on the performance parameters: DoA, IL, and ET, respectively. The Figure 5 illustrates a comprehensive comparison of the DoA performance for four clustering techniques: PSO-GA, LECC, MGPC, and GADOCA, under varying user densities. As the number of users increases from 2000 to 20,000, the DoA values for all

techniques demonstrate an upward trend. PSO-GA consistently maintains relatively lower DoA values, indicating its effectiveness in achieving higher levels of anonymity across different user densities. LECC and MGPC exhibit comparable DoA trends, with slightly higher values. Notably, GADOCA consistently outperforms the other methods, showcasing its ability to achieve enhanced anonymity even as user density increases. This analysis provides valuable insights into the scalability and performance of these clustering techniques concerning the DOA under varying numbers of users. The Figure 6 reveals the IL values for four distinct clustering techniques PSO-GA, LECC, MGPC, and GADOCA across different numbers of clusters and user densities. As the number of users increases from 2000 to 20,000, all techniques exhibit a consistent decrease in IL values. Lower IL values signify reduced IL during the anonymization process. PSO-GA consistently maintains the lowest IL values, indicating its ability to achieve higher privacy preservation. LECC and MGPC demonstrate similar trends with slightly higher IL values, while GADOCA consistently outperforms other methods. This figure 7 illustrates the ET values for four clustering techniques PSO-GA, LECC, MGPC, and GADOCA across different numbers of clusters and user densities. Lower ET values indicate quicker ET.

5. Conclusion

The presented research introduces a novel anonymization model tailored for OSN, addressing the crucial need for minimal loss of sensitive structural information and a heightened level of anonymity. The model operates through three distinctive phases initial, cluster optimization and privacy preservation effectively overcoming challenges encountered by current state-of-the-art techniques. The utilization of a statistical approach for normalizing input OSN data enhances the model's functionality. The hybrid score computation, incorporating multiple graph properties, contributes to more robust cluster formation compared to singular graph properties. Utilizing a one-pass approach reduces IL while simultaneously providing protection against known threats such as similarity and attribute disclosure. Experimental results, based on evaluations using the Yelp OSN dataset and various metrics IL, DoA, and ET, affirm the efficiency of the proposed model. Moreover, comparative studies with different cluster sizes and user densities show a significant 20% improvement in DoA and a 10% decrease in IL over current state-of-the-art techniques. Future research suggestions include dynamic clustering enhancements, assessments with other OSN datasets, and evaluations considering additional attack vectors beyond knowledge graph attacks.

Disclosure of Potential Conflicts of Interest

The authors confirm that there are no conflicts of interest related to the research, authorship, and/or publication of this article.

Funding

The authors declare that no financial support was received for the research, authorship, and/or publication of this article.

Reference

[1] Wani, M.A., Agarwal, N., and Bours, P., "Impact of unreliable content on social media users during COVID-19 and stance detection system". Electronics, 10(1), pp.5, (2020). https://doi.org/10.3390/electronics10010005

[2] Fire, M., Goldschmidt, R., and Elovici, Y., "Online social networks: threats and solutions". IEEE Communications Surveys & Tutorials, 16(4), 2019-2036 (2024). DOI: 10.1109/COMST.2014.2321628.

[3] Jain, A. K., Sahoo, S. R., and Kaubiyal, J., "Online social networks security and privacy". comprehensive review and analysis. Complex & Intelligent Systems, 7(5), pp.2157-2177, (2021). https://doi.org/10.1007/s40747-021-00409-7

[4] Wei, R., Tian, H., and Shen, H., "Improving k-anonymity based privacy preservation for collaborative filtering". Computers & Electrical Engineering, 67, pp.509-519, (2018). https://doi.org/10.1016/j.compeleceng.2018.02.017

[5] Temuujin, O., Ahn, J., and Im, D. H., "Efficient L-diversity algorithm for preserving privacy of dynamically published datasets". IEEE Access, 7, pp.122878-122888, (2019). DOI: 10.1109/ACCESS.2019.2936301

[6] Sathiya Devi, S., and Indhumathi, R., "A study on privacy-preserving approaches in online social network for data publishing". In Data Management, Analytics and Innovation: Proceedings of ICDMAI 2018, 1, pp. 99-115, Springer Singapore, (2019). https://doi.org/10.1007/978-981-13-1402-5_8

[7] Pan, X., and Hamilton, A. F. D. C., "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape". British Journal of Psychology, 109(3), pp. 395-417, (2018). https://doi.org/10.1111/bjop.12290

[8] Papaioannou, M., Karageorgou, M., Mantas, G., Sucasas, V., Essop, I., Rodriguez, J., and Lymberopoulos, D., "A survey on security threats and countermeasures in internet of medical things (IoMT)". Transactions on Emerging Telecommunications Technologies, 33(6), pp.e4049, (2022). https://doi.org/10.1002/ett.4049

[9] Tripathy, B. K., Sishodia, M. S., Jain, S., and Mitra, A., "Privacy and anonymization in social networks". Social Networking: Mining, Visualization, and Security, pp.243-270, (2014). https://doi.org/10.1007/978-3-319-05164-2_10

[10] Zhang, J., Chen, B., Zhao, Y., Cheng, X., and Hu, F., "Data security and privacy-preserving in edge computing paradigm: Survey and open issues". IEEE access, 6, pp.18209-18237, (2018). DOI: 10.1109/ACCESS.2018.2820162

[11] Tu, Z., Zhao, K., Xu, F., Li, Y., Su, L., and Jin, D., "Protecting trajectory from semantic attack considering ${k} $-anonymity, ${l} $-diversity, and ${t} $-closeness". IEEE Transactions on Network and Service Management, 16(1), pp.264-278, (2018). DOI: 10.1109/TNSM.2018.2877790

[12] Oishi, K., Tahara, Y., Sei, Y., and Ohsuga, A., "Proposal of l-diversity algorithm considering distance between sensitive attribute values". In 2017 IEEE Symposium Series on Computational Intelligence (SSCI) pp. 1-8, (2017). IEEE. DOI: 10.1109/SSCI.2017.8280973

[13] Mauger, C., Mahec, G. L., and Dequen, G., "Multi-criteria Optimization Using l-diversity and t-closeness for k-anonymization". In Data Privacy Management, Cryptocurrencies and Blockchain Technology: ESORICS 2020 International Workshops, DPM 2020 and CBT 2020, Guildford, UK, September pp. 17–18, 2020, Revised Selected Papers 15, pp. 73-88. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-66172-4_5

[14] Sei, Y., Okumura, H., Takenouchi, T., and Ohsuga, A., "Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness". IEEE transactions on dependable and secure computing, 16(4), pp. 580-593 (2017). DOI: 10.1109/TDSC.2017.2698472

[15] Mortazavi, R., and Erfani, S. H., "GRAM: An efficient (k, l) graph anonymization method". Expert Systems with Applications, 153, pp.113454, (2020). https://doi.org/10.1016/j.eswa.2020.113454

[16] Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J., "Beyond k-Anonymity: l-Diversity and t-Closeness". In Database Anonymization: Privacy Models, Data Utility, and Microaggregation-based Inter-model Connections, pp. 47-51, Cham: Springer International Publishing 2016. https://doi.org/10.1007/978-3-031-02347-7_6

[17] Kaushik, P., and Tayde, V. D., "Data Privacy in Hadoop Using Anonymization and T-Closeness". In Smart and Innovative Trends in Next Generation Computing Technologies: Third International Conference, NGCT 2017, Dehradun, India, October 30-31, 2017, Revised Selected Papers, Part II 3, pp. 459-468.,Springer Singapore (2018). https://doi.org/10.1007/978-981-10-8660-1_35

[18] Gangarde, R., Sharma, A., Pawar, A., Joshi, R., and Gonge, S., "Privacy preservation in online social networks using multiple-graph-properties-based clustering to ensure k-anonymity, l-diversity, and t-closeness". Electronics, 10(22), pp.2877, (2021). https://doi.org/10.3390/electronics10222877

[19] Singh, A., & Singh, M., "Social Networks Privacy Preservation: A Novel Framework". Cybernetics and Systems, pp.1-32, (2022). https://doi.org/10.1080/01969722.2022.2151966

[20] Gangarde, R., Shrivastava, D., Sharma, A., Tandon, T., Pawar, A., and Garg, R., "Data anonymization to balance privacy and utility of online social media network data". Journal of Discrete Mathematical Sciences and Cryptography, 25(3), pp.829-838, (2022). https://doi.org/10.1080/09720529.2021.2016225

[21] Mehta, B. B., and Rao, U. P., "Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing". Journal of King Saud University-Computer and Information Sciences, 34(4), pp.1423-1430, (2022). https://doi.org/10.1016/j.jksuci.2019.08.006

[22] Siddula, M., Li, Y., Cheng, X., Tian, Z., and Cai, Z., "Anonymization in online social networks based on enhanced equi-cardinal clustering". IEEE Transactions on Computational Social

Systems, 6(4), pp.809-820, (2019). DOI: 10.1109/TCSS.2019.2928324

[23] Jamshidi, M. B., Alibeigi, N., Rabbani, N., Oryani, B., and Lalbakhsh, A., "Artificial neural networks: A powerful tool for cognitive science". In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 674-679, IEEE (2018). DOI: 10.1109/IEMCON.2018.8615039

[24] Chen, J., Lin, X., Wu, Y., Chen, Y., Zheng, H., Su, M., and Ruan, Z., "Double layered recommendation algorithm based on fast density clustering: Case study on Yelp social networks dataset". In 2017 International Workshop on Complex Systems and Networks (IWCSN) pp. 242-252, IEEE (2017). DOI: 10.1109/IWCSN.2017.8276534

[25] Domingo-Ferrer, J., Farras, O., Ribes-González, J., and Sánchez, D. "Privacy-preserving cloud computing on sensitive data: A survey of methods, products and challenges". Computer Communications, 140, pp.38-60, (2019). https://doi.org/10.1016/j.comcom.2019.04.011