electronics

MDPI

Article

Privacy Preservation in Online Social Networks Using Multiple-Graph-Properties-Based Clustering to Ensure k-Anonymity, l-Diversity, and t-Closeness

Rupali Gangarde [1,*], Amit Sharma [2], Ambika Pawar [1], Rahul Joshi [1] and Sudhanshu Gonge [1]

[1] Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, Maharashtra, India; ambikap@sitpune.edu.in (A.P.); rahulj@sitpune.edu.in (R.J.); sudhanshu.gonge@sitpune.edu.in (S.G.)

[2] School of Computer Applications, Lovely Professional University, Punjab (Deemed University), Ludhiana 144411, Punjab, India; amit.25076@lpu.co.in

* Correspondence: rupali.gangarde@sitpune.edu.in

Abstract: As per recent progress, online social network (OSN) users have grown tremendously worldwide, especially in the wake of the COVID-19 pandemic. Today, OSNs have become a core part of many people's daily lifestyles. Therefore, increasing dependency on OSNs encourages privacy requirements to protect users from malicious sources. OSNs contain sensitive information about each end user that intruders may try to leak for commercial or non-commercial purposes. Therefore, ensuring different levels of privacy is a vital requirement for OSNs. Various privacy preservation methods have been introduced recently at the user and network levels, but ensuring k-anonymity and higher privacy model requirements such as l-diversity and t-closeness in OSNs is still a research challenge. This study proposes a novel method that effectively anonymizes OSNs using multiple-graph-properties-based clustering. The clustering method introduces the goal of achieving privacy of edge, node, and user attributes in the OSN graph. This clustering approach proposes to ensure k-anonymity, l-diversity, and t-closeness in each cluster of the proposed model. We first design the data normalization algorithm to preprocess and enhance the quality of raw OSN data. Then, we divide the OSN data into different clusters using multiple graph properties to satisfy the k-anonymization. Furthermore, the clusters ensure improved k-anonymization by a novel one-pass anonymization algorithm to address l-diversity and t-closeness privacy requirements. We evaluate the performance of the proposed method with state-of-the-art methods using a "Yelp real-world dataset". The proposed method ensures high-level privacy preservation compared to state-of-the-art methods using privacy metrics such as anonymization degree, information loss, and execution time.

Keywords: anonymization; clustering; k-anonymity; l-diversity; online social network; privacy preservation; t-closeness

check for updates

Citation: Gangarde, R.; Sharma, A.; Pawar, A.; Joshi, R.; Gonge, S. Privacy Preservation in Online Social Networks Using Multiple-Graph-Properties-Based Clustering to Ensure k-Anonymity, l-Diversity, and t-Closeness. *Electronics* **2021**, *10*, 2877. https://doi.org/10.3390/electronics10222877

Academic Editors: Hemant Ghayvat and Sharnil Pandya

Received: 19 October 2021
Accepted: 15 November 2021
Published: 22 November 2021

1. Introduction

An online social network (OSN) provides a powerful platform for users to interact and share information between one another [1]. According to the latest global digital report, at present, there are 800 million users of online social networks [2]. Privacy has become a significant concern with respect to many emerging technologies, such as the Internet of Things (IoT) and cloud computing, which generate tremendous data [3]. Moreover, severe concerns have arisen with respect to OSN privacy [4–7]. Due to the involvement of sensitive data, privacy in OSNs is a topic of interest to many researchers [8,9]. Data aggregated from different sources give rise to the problem of data privacy [10]. Many data privacy challenges—such as private information leakage, misuse of personal data, etc.—are commonly observed in OSNs [11]. Some of the most well-known spamming attacks include context-aware spamming and broadcast spamming attacks. The network structure is also prone to structural attacks such as Sybil attacks and shilling attacks. These attacks can spread worms and enable botnets to propagate in the OSN via profile interaction and third-party applications [12]. Many other intelligent systems are available to simplify our life like human–machine interactions so as to increase industrial production [13]. Models to analyze human relationships in cognitive science [14] and AI-based platforms can help to accelerate the early detection of diseases [15]. Data generated by these systems are enormous, and are prone to privacy attacks.

Security mechanisms alone do not guarantee the privacy of data. Thus, it is necessary to devise privacy solutions separately in order to preserve privacy. Sensitive data can include individuals' names, addresses, location information, phone numbers, e-mail ID, health and insurance details, social security numbers (SSNs), financial records, personal photos, videos, notes, credit card details, etc. Leaks in sensitive data could lead to lawsuits, loss of customer confidence, brand damage, erosion of privacy, bad press, loss of money or profit, etc. In India, the Indian Information Act 2000 has the provision to take action and enforce punishment if anybody is found to be responsible for the unethical exposure of personal information. Real-world data are temporary, but information on the web persists for a limitless time, eventually threatening online users' privacy. Users often end up sharing sensitive information while being unaware that they are at risk [16]. Privacy is always a concern when owners share data with third parties, and personal identifiable information (PII) is at stake. Hence, it is not easy to preserve privacy in a domain that is inherently intended for sharing. No unauthorized person should be able to acquire any sensitive information related to users. An unauthorized person can significantly breach data privacy if they gain access to users' sensitive information. Figure 1 gives an overview of some examples of sensitive information, and how it can be used to damage and affect users' privacy.

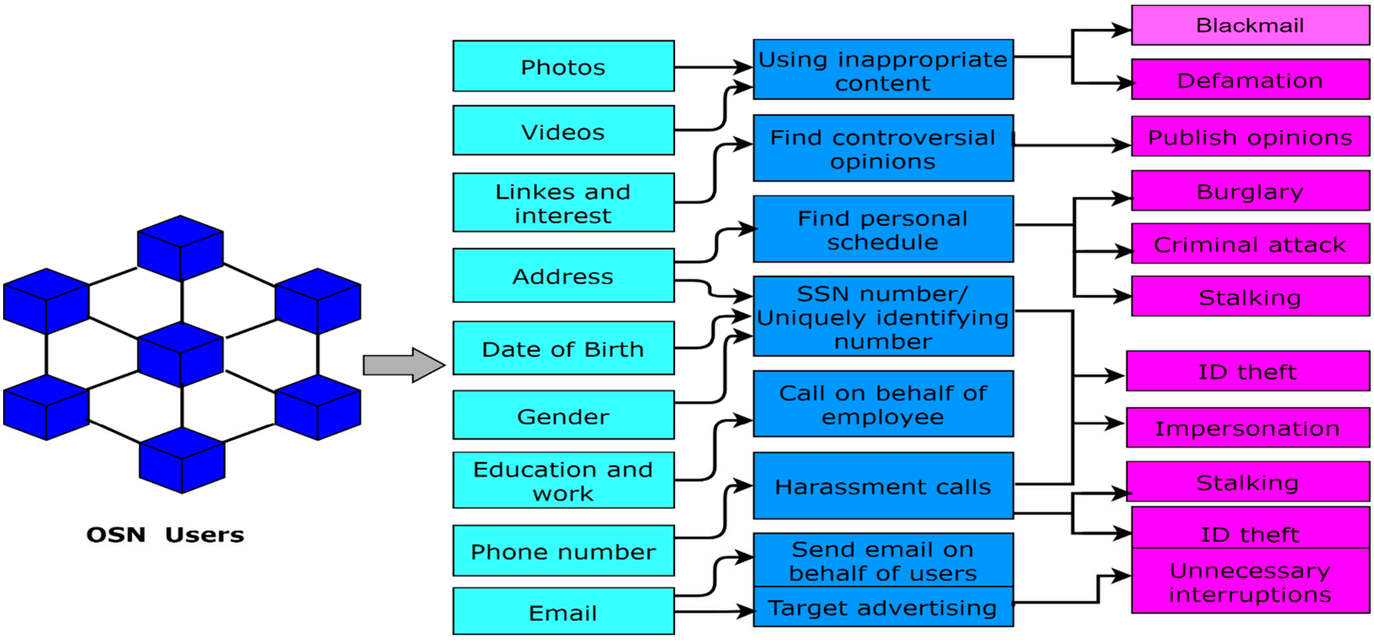

Figure 1. Sensitive information of OSN (online social network) users.

Users' photos and videos from their profiles could be doctored and used for blackmailing and defaming individuals [17]. Likes and interests reveal a lot about a person, and can lead to the revelation of controversial opinions. The address of a person can reveal their location, resulting in a criminal attack or burglary [18]. An individual's social security number (SSN) can be determined using a combination of address, date of birth, and gender, resulting in ID theft or impersonation [19]. Companies may use e-mails and phone numbers for targeted advertising, leading to unnecessary interruptions and spam. Therefore, it is an open challenge to protect confidential and sensitive data from unauthorized persons, and to ensure that the actual data are available only to legitimate users [20] of OSNs.

Privacy preservation using anonymization in OSNs has received significant interest. We can present OSNs in a graph form; the end user represents a node in the OSN network, and an edge is a connection between two nodes [21,22]. A node can have many edges in a social network graph, such as user-to-user, user-to-attribute, and attribute-to-attribute nodes. Therefore, it is necessary to protect all components of the OSN graph by anonymizing them, i.e., nodes, edges, and attributes. However, the purpose of any anonymization technique is that it should not eliminate ample information that induces utility of the original graph, causing structural information loss (IL). Identity preservation is one of the vital aspects in preserving the privacy of OSNs [22]. A pure, naive anonymization method substitutes an identifiable attribute user's name in the information with arbitrary identifiers, but the invaders can utilize the background knowledge of the structure of this anonymized graph to identify a user. Thus, the OSN should be anonymized so that ethical and trusting promoters can obtain information from it, but it remains ineffective to unethical parties who require it to obtain the sensitive personal data of individual users. In OSNs, two vital pieces of data need to be preserved: knowledge regarding a users' sensitive attributes, and relationships between users, i.e., edge/link details in graphs of the end users [23–26]. However, protecting these details in OSNs is a challenging research problem. Various techniques have been introduced for anonymization in social networks in order to achieve privacy preservation notions such as k-anonymity and l-diversity, but these privacy preservation notions have limitations, and fail to anonymize all social network elements [27] (i.e., nodes, edges, and attributes) effectively.

This research paper presents a novel approach for privacy preservation in OSNs using improved multiple-attribute-based clustering to satisfy all privacy preservation requirements with minimal information loss and reduced computational cost. We systematically plan the proposed anonymization approach to cover the privacy of all of the OSN graph elements. In this regard, first, we apply data normalization to address noisy data, messy data, and empty fields using the lightweight approach; this step helps to enhance the quality of raw OSN data for efficient anonymization. Then, the appropriate graph properties are selected to form the clusters of input social networks using K-means. This is a simple approach of clustering that ensures the minimum IL with the privacy of nodes, edges, and attributes. The formation of clusters takes place according to similar characteristics of nodes so as to satisfy k-anonymity.

Furthermore, clusters ensuring k-anonymity are improved with l-diversity and t-closeness to address every privacy notion in an OSN. The formation of clusters provides k-anonymity by default, but it is also necessary to address the concepts of l-diversity and t-closeness. For this aim, we propose a novel one-pass algorithm that ensures the l-diversity and t-closeness for each k-anonymized cluster. Section 2 presents related works on the privacy-preserving OSNs, along with the motivations and contributions of our research. The problem statement and the explanation of the proposed model are given in Section 3. The experimental results and the privacy preservation analysis using real-world datasets are shown in Section 4. Finally, Section 5 discusses conclusions and prospects for future work.

2. Related Works

Privacy preservation in OSNs is a novel investigative field that still in progress. Many of the studies in this prominent field are dependent on a computational viewpoint. Here, we discuss relevant recent works in the area of privacy preservation in OSNs. As discussed earlier, our privacy study considers the OSN in the form of a graph with nodes, edges, and node attributes. Therefore, we can achieve privacy for all of these OSN graph elements [28–30]. The related works we review in Sections 2.1 and 2.2 present research motivations. Finally, Section 2.3 shows the contributions of our research.

2.1. Privacy Methods in OSNs

The earlier pervasive social network (PSN) method was proposed in [31] for privacy preservation. The anonymous authentication algorithm helps to authenticate trust levels and pseudonyms in order to provide trustworthy PSNs and privacy preservation [32]. This achieves secure, anonymous authentication using trusted authorities. Blockchain-based frameworks provide authorization and identity management to increase confidentiality and preserve data privacy [33]. In [34], the author addresses identity disclosure threats in weighted social network graphs; the weighted 1*-neighborhood threats are identified for OSN users under the assumption of knowledge about target node connections, corresponding edge weights, and node degrees; the author designed heuristic indistinguishable group anonymization (HIGA) to address 1*-neighborhood attack.

The survey study in [35] presents recent privacy preservation techniques for OSNs. These privacy preservation methods include perturbation, building the entire alternative network, and naïve anonymization, along with their limitations. The authors of [36] proposed a hybrid privacy preservation approach for social networks; they considered both identity and location privacy in order to address privacy leakage and robustness; the game-based Markov decision process system achieved improved data utility with higher privacy preservation. A local differential privacy scheme was proposed in [37] for OSN publishing in order to preserve information about community structure; the synthetic social network information was generated in this model as the published versions according to structural constraints on edge probability reconstruction.

An efficient and fast social network de-anonymization technique that relied on structural information was proposed in [38]. The authors designed a novel pairwise node similarity metric and effective node-matching technique. The clustering algorithm in [39] achieves k-anonymity in OSNs using swarm intelligence; initially, the author designed the clustering algorithm using particle swarm optimization (PSO) to reduce the IL; however, PSO-based clustering leads to a high computational burden; therefore, for OSN clustering, the author proposed a hybrid genetic algorithm (GA) and PSO-based algorithm (PSO-GA). Another recent study [40] proposed a de-anonymization scheme for OSNs to reveal the impact of user attributes on de-anonymization accuracy; the authors quantified user attributes and chose vital attributes to produce a multipartite graph, which was divided into different communities. In [41], the authors present another clustering-based privacy preservation scheme for OSNs, aiming to achieve the privacy of all of the social network elements—nodes, links, and attributes—via proposed clustering, with the OSN nodes clustered using the similarity metrics to achieve k-anonymity; the k-anonymity is further enhanced to achieve the l-diversity privacy notion.

Recently, a feature learning model was developed in [42] to achieve privacy preservation; the authors used a feature learning approach to define the social connections between the social users and then build inferred social graphs, which they used for the purpose of privacy preservation. The research presented in [43] identifies privacy bounds on the data of individuals' unique mobility traces; individuals' privacy is preserved by coarsening data spatially and temporally for anonymity. A mechanism for the preservation of privacy during message transmission was introduced in [44] via message obfuscation, using the message replication and sensitive attributes replacement strategy; the authors analyzed the social behaviors of each user in order to compute their credibility for privacy preservation in OSNs. The differential privacy scheme proposed in [45] combines different series to achieve privacy of all of the graph elements; the dK-1 series holds the degree frequency, the dK-2 series holds the joint degree frequency, and the linking knowledge between the edges is stored in the dK-3 series. A customizable reliable differential privacy (CRDP) scheme was proposed in [46] to achieve customizable privacy for every individual; the authors measured the social distance in order to ascertain the shortest path between two vertices, and then utilized those vertices as the metrics to customize the levels of privacy protection.

2.2. Motivation

The above studies show that privacy preservation is still a challenging problem for OSNs when considering minimum sensitive structural IL, high-level privacy protection, and minimum complexity. The summary of the research gaps in current state-of-the-art methods that motivated us to propose the novel model in this paper are as follows:

Cryptography-based methods [31–33] achieve secure communications with a certain level of privacy, but cannot achieve privacy in all of the elements of OSNs. Moreover, they rely on trusted authorities for secure communications and privacy.

Grouping/clustering-based methods [39–41] have shown promising outcomes, but have yet to simplify and improve performance and privacy protection tradeoff requirements. Swarm-intelligence-based clustering achieves only k-anonymity privacy notions in OSNs with higher computational complexity. Such clustering techniques fail to achieve high-level privacy preservation in OSNs because of the poor quality of anonymization. The clustering in [41] focuses on only a single graph attribute for clustering, which limits the privacy protection in OSNs.

Other different types of OSN graph-based methods [34–38,42–44] fail to achieve privacy in all of the components, i.e., nodes, edges, and attributes of nodes. In such techniques, attackers can efficiently utilize the structural information of anonymized network graphs, leading to information loss.

The differential privacy [45] and customized privacy [46] schemes for OSNs have received attention. However, differential privacy schemes assume that each data owner shares similar privacy demands and, therefore, fail to address different notions of privacy. On the other hand, a customizable privacy scheme triggers the formation method of differential privacy, resulting in unexpected relationships between added noises that minimize privacy preservation and leak more sensitive information and privacy demands.

2.3. Contributions

Considering the above research gaps in recent OSN graph-based privacy preservation methods, we propose a novel clustering-based privacy preservation scheme using multiple graph properties for cluster formation to achieve high-level privacy protection in OSNs with minimal IL and reduced computational cost.

The contributions stated below summarize the novelties of the proposed model:

1. After data normalization and K-means clustering, we propose the novel cluster optimization algorithm to achieve k-anonymity using two graph properties: distance, and eccentricity. Multiple graph properties ensure reliability in cluster optimization, with minimal sensitive information leakage and a higher degree of anonymization. The cluster optimization phase produces clusters with at least k-anonymized users;
2. To enhance the privacy preservation of k-anonymized clusters, we propose the novel one-pass algorithm to ensure that the clusters have l-diversity and t-closeness, and to protect data from similarity threats and attribute disclosure threats;
3. The notion of equal-distance-based t-closeness privacy ensures the prevention of disclosure of users' attributes, and of similarity threats. The notion of l-diversity privacy ensures the prevention of the disclosure of sensitive attributes at the cluster head (CH);
4. The analysis of our results presents performance comparison of the proposed method with similar methods, using a real-world dataset, and varying the number of users and the number of clusters.

3. Proposed Method

This section presents the methodology of the proposed anonymization scheme for OSN privacy using clustering. The proposed model has three phases: the initial phase, the cluster optimization phase, and the privacy preservation phase, as shown in Figure 2. The initial phase performs the input OSN data normalization and initial clustering.

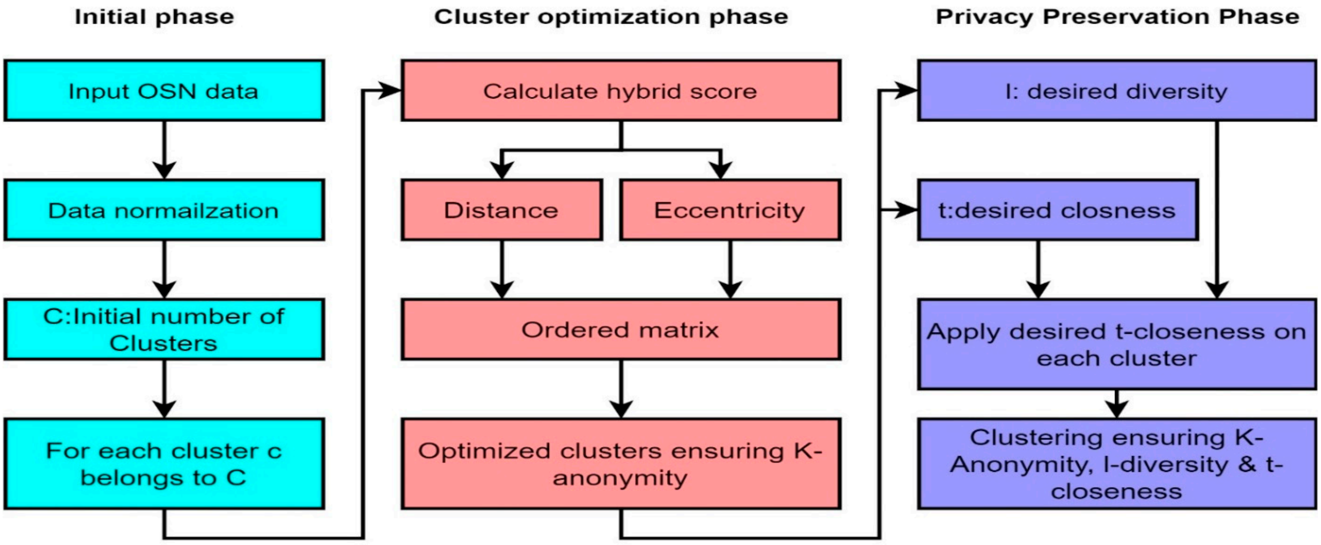
Initial phase
Input OSN data
Data normailzation
C:Initial number of Clusters
For each cluster c belongs to C
Cluster optimization phase
Calculate hybrid score
Distance
Eccentricity
Ordered matrix
Optimized clusters ensuring K-anonymity
Privacy Preservation Phase
l: desired diversity
t:desired closness
Apply desired t-closeness on each cluster
Clustering ensuring K-Anonymity, l-diversity & t-closeness

Figure 2. Proposed system model approach for privacy preservation in OSNs.

The cluster optimization phase optimizes the clusters using two graph properties—distance and eccentricity—to ensure k-anonymity with minimal information leakage. The privacy preservation phase further enhances the clusters to ensure l-diversity and t-closeness.

Table 1 shows the functionality of the proposed model. The three phases, and the processing within each, provide a clear understanding of the proposed system.

Table 1. The functionality of the proposed model.

Phases	Processing
Initial Phase	1. Acquire OSN data 2. Data pre-processing using Algorithm 1 3. Apply K-means clustering
Cluster Optimization Phase	1. For each cluster, initiate cluster optimization using Algorithm 2 2. Compute distance of each vertex 3. Compute eccentricity for each vertex 4. Compute hybrid score using distance and eccentricity properties for each vertex 5. Sort each cluster according to hybrid score 6. Optimize clusters to satisfy k-anonymity
Privacy Preservation Phase	1. Acquire k-anonymized clusters 2. Apply one-pass algorithm (Algorithm 3) 3. Within Algorithm 3, ensure l-diversity for each cluster 4. Within Algorithm 3, ensure t-closeness for each cluster 5. Produce clusters ensuring k-anonymity, l-diversity, and t-closeness

The input raw OSN data consists of users, their attributes, and edges between them. The data normalization step is required in order to prepare the sets of nodes and their corresponding attributes by performing the statistical operations. Additionally, we apply the function to detect the missing or messy data and replace them with relevant values. This step ensures not only data quality enhancement, but also the accuracy of anonymization in OSNs. After the data normalization step, we form the initial clusters using the conventional K-means clustering algorithm. The reasons behind selecting K-means clustering are as follows: (1) it is simple to cluster users according to their similarities; (2) it is fast and produces efficient clusters; and (3) K-means cannot prevent the outliers, and supports privacy for all outliers. After the initial clustering, we optimize the clusters according to distance measures using the multiple graph properties of each user. The multiple graph properties ensure reliable cluster formation with high-level privacy protection. After forming the clusters that provide k-anonymity, we apply the post-processing on each cluster using a one-pass algorithm. The one-pass algorithm ensures both l-diversity and t-closeness privacy notions in the proposed model. First, we present the system model to formulate the proposed problem, and then we illustrate the design of the proposed solution.

3.1. System Model

Let us consider the input OSN data, which are represented in the form of graph G that consists of vertices V and edges E. V represents the social network users, while E represents a link or connection between two users/vertices. We have n vertices, where $V = \{v^1, v^2, \ldots v^n\}$ in the network, where each vertex v^i has m associated attributes represented in set A as $A = \{a^i_1, a^i_2, \ldots a^i_m\}$. The edges between the vertices have directions; therefore, the total number of edges in the network is:

$$E = n \times (n-1) \quad (1)$$

where E is a set of edges. $E = \{e^{12}, e^{21}, e^{23}, \ldots e^{n,n-1}\}$. e^{12} denotes an edge from v^1 to v^2, which is not the same as e^{21}—an edge from v^2 to v^1. Each edge is assigned a weight value. The primary goal of this paper is to achieve privacy preservation for social network graph G such that each component (E, V, and A) of graph G should be anonymized to satisfy the following objectives:

1. To achieve complete privacy notions—k-anonymity, l-diversity, and t-closeness—for giving the constant value k;
2. To minimize the sensitive information leakage and reduce processing time, with minimal IL, and achieve high-level privacy preservation;
3. To improve the reliability of anonymization considering the multiple graph properties and data normalization.

The proposed method utilizes the optimized clustering approach to group the vertices with similar properties and build meaningful clusters of social users. Each cluster has its $\mathrm{n_u}$ to represent the number of users of that cluster. The $\mathrm{n_u}$ selection anonymizes the K users. We process anonymized clusters according to the desired l-diversity and t-closeness parameters in order to satisfy higher privacy preservation notions. In the following subsections, we present the design of each phase of the proposed model. Table 2 presents the mathematical symbols used in the proposed model's design, along with their significance.

Table 2. List of symbols.

Symbol	Quantity
G	OSN graph
RD	Raw OSN data
E	Set of edges (relationship between two users) in the graph
V	Set of vertices (users) in graph G
A	Set of attributes of each user/vertex
a^i_j	j^{th} attribute of i^{th} user/vertex
uid	Represents the user ID in raw OSN data
NR^{uid}	Number of reviews posted by uid
AY^{uid}	Active years of uid
NF^{uid}	Number of friends of uid
FS^{uid}	Number of fans of uid
VS^{uid}	Average vote score of uid

Table 2. *Cont.*

Symbol	Quantity
ES^{uid}	An elite score of *uid*
CS^{uid}	Compliment scores of *uid*
A^{uid}	Attribute set of *uid*
c	Number of clusters
C	Set of *c* clusters
A^{i}_{cent}	Attribute set of the centroid of i^{th} cluster
H^{uid}_{i}	The hybrid score value of user/vertex *uid* of i^{th} cluster
CM^{i}	Set of users/cluster members of i^{th} cluster
R	Total number of attributes in set A
$d^{uid,\ cent}$	The distance between the vertex *uid* and a centroid
e^{uid}	The eccentricity of vertex *uid*

3.2. Initial Phase

1. Preprocessing Algorithm

The initial phase normalizes the input raw OSN data and forms the initial clusters using the basic K-means clustering algorithm. The reasons for normalizing the OSN data and selecting the K-means clustering were disclosed in the previous section. The input dataset generally consists of a large number of users and their various attributes. The attributes represent each user's connection and behavior in the network. The raw data may have outliers or messy data; therefore, before performing the clustering, data normalization should be performed. The goal of the data normalization phase is to enhance the quality of OSN data via the statistical modeling of attributes. Algorithm 1 shows the process of data normalization that takes the raw OSN data and produces the user set V with its associated normalized set of attributes A.

Algorithm 1 Preprocessing

Input
f: raw OSN data
Output
V: Set of vertices
A: Set of attributes

```
while $(f\ not\ empty)$
    Acquire each profile $P \in f$
    $P \leftarrow f\ (uid,\ name)$
    if $(uid.attributes\ \neq NULL)$
        $NR^{uid} \leftarrow uid.review\_count$
        $AY^{uid} \leftarrow uid.difference\ (date,\ yelping\_since)$
        $NF^{uid} \leftarrow uid.size\ (friends)$
        $VS^{uid} \leftarrow uid.mean\ (useful,\ cool,\ funny)$
        $FS^{uid} \leftarrow uid.fans$
        $ES^{uid} \leftarrow uid.size\ (elite)$
        $RS^{uid} \leftarrow uid.average\_stars$
        $CS^{uid} \leftarrow uid.(compliment_{all})$
        $A^{uid} \leftarrow \left[NR^{uid},\ AY^{uid}, NF^{uid}, VS^{uid},\ FS^{uid},\ ES^{uid},\ RS^{uid},\ CS^{uid}\right]$
        $A \leftarrow \left[A; A^{uid}\right]$
        $V \leftarrow [V; uid]$
    end if
end while
return $(A,\ V)$
```

The novelty of Algorithm 1 is that it can be applied to any OSN data by adjusting the number of attributes and their value according to the properties of the input data. The statistical normalization of users and their corresponding attributes results in the removal of outliers or messy data without a special function. For example, we applied Algorithm 1 to the Yelp OSN dataset [41], which consists of a total of 18 attributes associated with each OSN user/vertex. After applying Algorithm 1 to these input datasets, we reduced the 18 preprocessed attributes to a total of 8 attributes. This helps to reduce the significant processing time without compromising data loss;

2. Initial Clustering (K-means Clustering Algorithm)

The normalized inputs of vertices (V) and attributes (A) fed to the clustering algorithm help in computing cluster centroids and their cluster members (CMs) accurately at the initial level only. The normalized attribute values also assist in the reduction of sensitive information. Next, we form the initial clusters to discover the cluster centroids from the input set of vertices using the K-means:

$$C = kmeans\ (V,\ c) \tag{2}$$

where c represents the number of clusters, generating initial clusters according to the mean of user attributes. Each cluster $C^i,\ i \in c$ has at least k users in order to satisfy k-anonymity in the network. At the initial level, the value of k is not the same for each cluster, i.e., each cluster may have a different number of users. Let us assume that we have a total of 100 vertices/users in the dataset, and we set the value of the clusters to 4; after applying Equation (2), this produces the outcome shown in Table 3. This shows that K-means clustering failed to achieve k-anonymity in the given OSN network. The K-means algorithm failed to achieve complete k-anonymity across all the clusters. Therefore, we further optimized each cluster using multiple graph properties—such as distance and eccentricity—to ensure that all clusters were of the same size. We will explore this process in the following section.

Table 3. Example of initial clustering.

Cluster Number	Number of Users
C^1	**29**
C^2	**20**
C^3	**21**
C^4	**30**

3.3. Cluster Optimization Phase

The goal of the proposed model is to achieve k-anonymity by utilizing clustering on preprocessed OSN data. However, as discussed earlier, using the simple K-means, we cannot achieve k-anonymity. The value of K indicates that every cluster has at least K anonymous users. A variation in cluster members leads to an information leakage problem. Let us assume that we have two clusters—C^1 and C^2—with K users and $p = K + 10$ users, respectively. In this way, all vertices in cluster C^1 are K-anonymous, and all vertices in cluster C^2 are p-anonymous. This variation in anonymity levels (different values of K for each cluster) results in sensitive information loss. Therefore, it is necessary to have clusters with an equal level of anonymity. To address this problem, we designed the cluster optimization phase to rearrange the clusters into clusters of the same size i.e., anonymized clusters with less IL. This rearrangement of clusters is possible by computing the score of each user in each cluster using two graph properties: the distance between the attributes of two vertices, and the eccentricity of each vertex. The number of users in the cluster should satisfy the constraint (n/k), where n is the number of users and k is the number of clusters. To the best of our knowledge, this is the first attempt to use multiple graph properties to ensure the privacy of all OSN graph elements.

The hybrid score is computed using distance and eccentricity graph properties for each vertex/user in the current cluster, and is sorted into the ordered matrix, as shown in Figure 2. Algorithm 2 presents the functionality of the cluster optimization phase that consists of two functions: hybrid score matrix computation, and cluster optimization.

1. Hybrid Score Matrix Computation

The initially formed clusters using K-means consist of a set of clusters with their centroid. To rearrange the clusters, we perform the computation of the hybrid score for each vertex/user. Algorithm 2 presents the process of computing the ordered hybrid matrix and cluster optimization.

Algorithm 2 Cluster Optimization

```
Inputs
$C:\ set\ of\ clusters\ with\ its\ centroid$
$c:\ number\ of\ clusters$
$A:\ set\ of\ attributes$
$n:\ total\ number\ of\ vertices\ in\ network$
Output
$SD$ : $sorted\ users\ list\ according\ to\ hybrid\ score$
$C:\ optimized\ clusters\ ensuring\ the\ K-anonymity$
Hybrid Score Matrix Computation:
$D \leftarrow ones(n,\ 2)$
$m = 1$
for $i = 1 : c$
for $j = 1 :\ size\ (C^i)$
$uid \leftarrow CM^i(j)$
$H_i^{uid} \leftarrow getScore\left(A^{uid},\ \ A_{cent}^i\right)$
$D(m,\ 1) \leftarrow uid$
$D(m,\ 2) \leftarrow H_i^{uid}$
$m \leftarrow m+1$
end For
end For
$SD \leftarrow Sort(D(:,\ 2),\ \text{“}ascending\text{”})$
Cluster Optimization:
for $i = 1 :\ n$
for $j = 1 : c$
if $\left((status\ (SD(i:,\ 1)) \neq assigned)\&\&\left(lenght\left(C^j\right) \leq |\frac{n}{c}|\right)\right)$
$C^j \leftarrow join\ (SD(i:,\ 1))$
$status\ (SD(i:,\ 1)) \leftarrow assigned$
end if
end for
end for
```

To calculate the hybrid score of each vertex/user, we use distance and eccentricity graph properties. To measure the distance property, we measure the distance from every user to its associated cluster centroid. For the eccentricity property, we measure the maximum connections of each vertex. In general, the distance property represents the number of edges between two vertices, and considers the minimum number of edges as its outcome. In short, the maximum closeness or similarity between the attributes of two vertices represents the distance property in the proposed scenario. On the other hand, the eccentricity property denotes each vertex's maximum number of connections in the network. In our case, we compute the eccentricity of each user/vertex from its attribute number of friends NF^{uid}. The inclusion of the eccentricity property ensures clustering balance and reliability, with minimal possibility of information loss (IL), as it enables grouping of the vertices according to their connections. We utilize both properties to produce a hybrid normalized score using a weight-based approach. Finally, we arrange all of the vertices into an ascending-order hybrid matrix, with the first column as a sorted list.

H_i^{uid} represent the hybrid score value of user/vertex *uid* of the i^{th} cluster, $size\ (C^i)$ represents the number of CMs of the i^{th} cluster. A^{uid} represents the attributes of the j^{th} user/vertex, and A^i_{cent} represents 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 i^{th} cluster and A^i_{cent}. Before that, we first obtain the vertex ID—i.e., *uid*—to obtain its corresponding set of attributes. Using the $getScore\ (.)$ function, we measure the two graph properties distance and eccentricity, as discussed above. The distance between A^{uid} and A^i_{cent} is computed by Equation (3):

$$d^{uid,\ cent} = \frac{\sum_{r=1}^{R} \left| a_r^{uid} - a_r^{cent} \right|}{R} \tag{3}$$

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 denotes the total number of attributes of each vertex in the network. Consider the absolute difference between the two attributes as the shortest distance between them; then, divide the aggregate distance of all of the attributes by the total number of attributes.

The second graph property—the eccentricity of each vertex A^{uid}—is computed from its value of NF^{uid} by Equation (4):

$$e^{uid} = \left(\frac{1}{NF^{uid}}\right) \times \lambda \tag{4}$$

The minimum value of e^{uid} represents the maximum eccentricity of the vertex. The symbol λ represents the scaling factor to normalize the eccentricity outcome. Compute this scaling factor by taking the mean of NF^{uid} of all of the vertices, using Equation (5):

$$\lambda = \frac{\sum_{i=1}^{n} NF^i}{n} \tag{5}$$

The scaling factor ensures the normalized eccentricity score of each vertex. Equation (6) computes the hybrid score of each vertex by using a weight-based technique:

$$H_i^{uid} = \left(w1 \times d^{uid,\ cent}\right) + \left(w2 \times e^{uid}\right) \tag{6}$$

where $w1$ and $w2$ represent the weights of each graph property. The value of both weight parameters should be $w1 + w2 = 1$. In this work, we assign equal weights to both distance and eccentricity parameters, i.e., $w1 = 0.5$ and $w2 = 0.5$. The two vertices with a minimum value of H_i^{uid} represent more closeness or similarity between them. This improves the probability of forming clusters with more similar vertices. Matrix D stores all vertices and their hybrid scores one-by-one; then, we sort the vertices in matrix D according to their hybrid score value, in ascending order, into matrix SD;

2. Cluster Optimization

As discussed above, the goal of the cluster optimization phase is to achieve the k-anonymity privacy preservation notion in OSN. For this purpose, we rearrange the clusters according to the sorted vertices matrix SD in Algorithm 2. Each user/vertex $SD(i :, 1)$ joins a current cluster C^j according to two constraints: (1) user/vertex status should not be "assigned", and (2) current cluster size should be less than or equal to n/k. These two constraints achieve clusters with almost similar sizes so as to attain k-anonymity with minimal IL. Once the current user in sorted matrix $SD(i :, 1)$ is satisfied by both constraints, its status is set to "assigned". In this manner, all of the vertices are divided into c clusters with equal probability, and each cluster has k users with maximum similarity.

This significantly reduces the chance of leaking sensitive information on nodes, edges, and their attributes. Therefore, the outcome of Algorithm 2 returns the optimized clusters, ensuring k-anonymity. Thus, the outcome shown in Table 4 is optimized further using the proposed clustering approach shown in Table 4, where k is set to 4 and n is set to 100.

Table 4. Example of optimized clustering.

Cluster Number	Number of Users
$\boldsymbol{C^1}$	**25**
$\boldsymbol{C^2}$	**25**
$\boldsymbol{C^3}$	**25**
$\mathbf{C^4}$	**25**

3.4. Privacy Preservation Phase

In the previous phase, the optimized clusters ensured k-anonymity for OSNs. However, k-anonymity does not guarantee the complete privacy preservation requirements via its limitation of attribute disclosure threat, background knowledge threat, and homogeneity threat [47]. l-Diversity addresses the problems of K-anonymity in [41]. However, l-diversity also suffers from various challenges, in that (1) preventing attribute disclosure threat is not sufficient, and (2) it is unnecessary to achieve. In [19], the authors state that t-closeness can be used to address the drawbacks of both k-anonymity and l-diversity. In this paper, our goal was to consider all privacy preservation notions using one common technique, called the one-pass algorithm. Therefore, this section presents the one-pass algorithm to extend the clusters, ensuring k-anonymity along with l-diversity and t-closeness. Before the one-pass algorithm, we give the standard definitions of l-diversity and t-closeness as follows:

> "*An equivalence class is said to have l-diversity if there are at least l well-represented values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity*" Definition of l-diversity [19].

> "*An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness*" Definition of t-closeness [19].

As per the definitions of l-diversity and t-closeness, l-diversity addresses the background knowledge and homogeneity attacks, but does not sufficiently address attribute disclosure. Meanwhile, t-closeness addresses all of the problems of l-diversity, but cannot deal with identity disclosure. k-anonymity effectively addresses the identity disclosure problem. This suggests that we should develop a common technique to extend the k-anonymized clusters with l-diversity and t-closeness. We designed the one-pass algorithm to post-process the optimized clusters in order to ensure the l-diversity and t-closeness security notions. Algorithm 3 shows the functionality of a one-pass algorithm that first ensures the t-closeness with predefined threshold value t, and then applies the l-diversity according to the entropy l-diversity technique. The one-pass algorithm thus achieves complete privacy preservation.

Algorithm 3 One-pass privacy preservation

```
Inputs
$C:\ optimized\ clusters\ ensuring\ the\ K-anonymity$
$c: number\ of\ clusters$
$A:\ set\ of\ attributes$
$l: desired\ diversity$
Output
$C:\ clusters\ ensuring\ the\ l-diversity\ and\ t-closeness$
for $i = 1:\ c$
    for $j =\ 1:\ \ lenght\left(C^i\right)$
        $uid \leftarrow C^i(j)$
        $temp \leftarrow getDist\ \left(A^{uid},\ A \leftarrow C^i\right)$
        $T(j,\ 1) \leftarrow uid$
        $T(j,\ 2) \leftarrow temp$
    end for
    $t \leftarrow mean(T)$
    Anonymize all users in clusters using $t$-closeness:
    for $i = 1:\ lenght\ (T)$
        if $(T(i,\ 2) < t)$
            $L1 \leftarrow join(T(i,\ 1))$
        else
            $L2 \leftarrow join(T(i,\ 1))$
        end if
    end for
    $C^i \leftarrow append\ (L1,\ L2)$ % returned the t-closeness anonymized cluster
    $LD^i \leftarrow getDiversity\ \left(C^i\right)$
end for
while $(diversity(LD) < 1)$ do
    $Max \leftarrow cluster\ with\ maximum\ diversity\ value\ from\ LD$
    $Min \leftarrow cluster\ with\ minimum\ diversity\ value\ from\ LD$
    $Temp \leftarrow Max + Min$
    $C \leftarrow C - \{Max,\ Min\} + Temp$
end while
```

Algorithm 3 presents the process in a very simplified manner to achieve l-diversity and t-closeness for the input k-anonymized clusters. As per the definition of t-closeness, for each cluster, we group the users according to their dynamically computed t-value. We calculate the t-value using the earth mover's distance (EMD) [14]. As shown in Algorithm 3, there is an equal distance between each user and all other members of the current cluster using function $getDist$ (.). 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 (7), as follows:

$$temp = \frac{1}{2}\sum_{j=1}^{q}\left|A^{uid} - A \leftarrow C^i(j)\right| \tag{7}$$

In this way, matrix T stores the EMD value for all of the cluster members. To anonymize the users in that cluster, we compute the t-value by the mean of all distances in matrix T. We have anonymous users that are satisfied and dissatisfied with the t-value. Finally, the users in both lists append to reform the cluster. This ensures the prevention of similarity attacks and attribute disclosure attacks in each cluster. Then, we extend the clusters to satisfy the l-diversity privacy notion using the entropy l-diversity concept [41].

For each t-closeness-anonymized cluster, we can compute its diversity using entropy and store its outcome in the matrix LD. The greedy algorithm ensures that each cluster satisfies the l-diversity. The process continues until all of the clusters achieve l-diversity.

Throughout the entire one-pass algorithm, because we do not eliminate any users from the cluster, the privacy notion of k-anonymity still exists for the optimized clusters.

The proposed clustering algorithm ensures the privacy preservation of vertices/users and their attributes, with minimal loss of sensitive information and minimal computational burden. After ensuring that the clusters meet all three privacy preservation notions—k-anonymity, l-diversity, and t-closeness—we further perform edge anonymization in OSNs. For edge anonymization, we use the algorithm proposed in [41] for the proposed model. As the CH node represents each cluster, the computation of super-edges among the clusters is applied to achieve edge anonymization. This approach anonymizes all of the edges for the weighted directed OSN graph. The outcome of the proposed model is clusters with k-anonymity, l-diversity, t-closeness, and anonymized edges.

4. Experimental Results

This section explains the outcomes of the experimental work for performance analysis of the proposed model with two state-of-the-art methods. We performed experiments using MATLAB on Windows 10 with an Intel I3 processor and 4 GB RAM. Each scenario was executed for 25 instances, and their performances were averaged. The first scenario was a hybrid swarm-intelligence-based OSN clustering method called PSO-GA [39], and the second method was l-diversity enhanced equi-cardinal (LECC) clustering [41] for privacy preservation. The reasons for selecting these methods were as follows: (1) both techniques are closely related to the proposed model because of their clustering approach, (2) both methods have recently been proposed to achieve OSN anonymization, (3) PSO-GA performs the clustering of OSNs to achieve anonymization, and (4) LECC performs clustering to ensure the k-anonymity and l-diversity using the distance graph properties. Furthermore, LECC performed edge anonymization similar to the proposed method. We introduced the threat of knowledge graphs for performance analysis of all methods.

4.1. Dataset and Performance Metrics

To analyze the efficiency of all methods, we performed experiments on a real-life Yelp dataset [48]. The Yelp dataset holds a consumer review set, where every user is attached to many other users and has information about those users' profiles. We used two files from this dataset for experimental analysis as friends and users. Those two files provided data on the user attributes and the edge data among the users. The user files consist of a user ID and 18 attributes of that user. To investigate the proposed method with PSO-GA and LECC—state-of-the-art methods—we measured three performance parameters: degree of anonymization (*DoA*), information loss *(IL*), and execution time (*ET*). The *ET* represents the average execution time for each scenario of 25 instances required to perform the OSN data anonymization. To compute the *DoA* of any user, we calculated the total number of assigned users in its cluster, i.e., user *DoA* is similar to the *DoA* of its cluster. Thus, *DoA* in Equation (8) is:

$$DoA = degree\ (C_{u_i}) \times i \tag{8}$$

where C_{u_i} denotes the degree of anonymization of user u_i that belongs to cluster C.

To compute the *IL* metrics, we used the formulation presented in [49]:

$$IL = \frac{SSE}{SST} \tag{9}$$

where *SSE* is the sum of squares within the cluster, and *SST* is the sum of squares among clusters.

4.2. Variations in Cluster Size

This section presents the performance analysis of variations in the cluster size, i.e., the number of clusters. We changed the cluster size from 20 to 100, and set the maximum number of users to 10,000. As we were working on a large dataset, we deemed that at least 20 clusters needed to be formed. This experimental study aimed to understand the effects of cluster size on the *DoA, IL*, and *ET* as performance metrics. The first observation of this outcome is that with the increase in the cluster size, the anonymization decreased. The primary reason for this is that the small number of clusters maintains the high number of K-anonymous users, but increasing the cluster size decreases the proportion of at least K-anonymous users in each cluster. Among the performances of the three methods, the proposed method outperformed both LECC and PSO-GA significantly. The LECC showed the second-best performance, with higher *DoA* for the input dataset compared to PSO-GA.

Figure 3 demonstrates the outcome of *DoA* for varying cluster size scenarios using PSO-GA, LECC, and the proposed method.

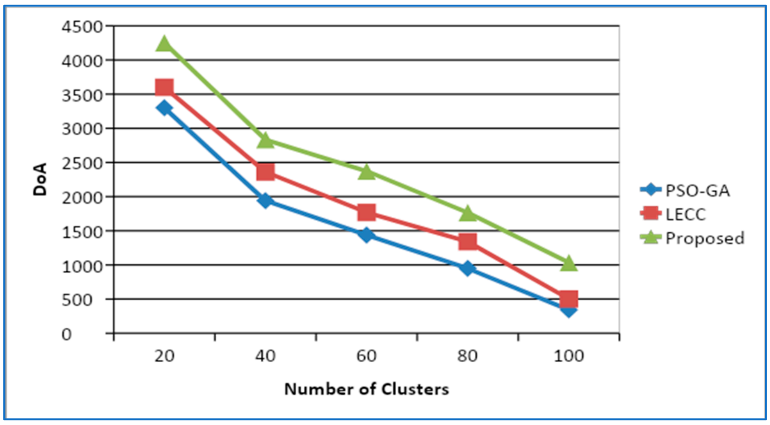

Figure 3. Effect of the number of clusters on *DoA* (degree of anonymization) performance.

The proposed model improved the *DoA* performance by 20% compared to the second-best method (LECC). The novel design of cluster optimization and the one-pass algorithm of the proposed model are the main reasons for the performance improvement. The clusters were optimized in the proposed model using the multiple graph properties, rather than just the one property in the LECC method. The cluster optimization step of the proposed model normalizes the OSN data and leads to efficient clustering. The one-pass algorithm for ensuring the l-diversity and t-closeness improved the *DoA* performance.

Figure 4 demonstrates another vital outcome, *IL*, using all three methods.

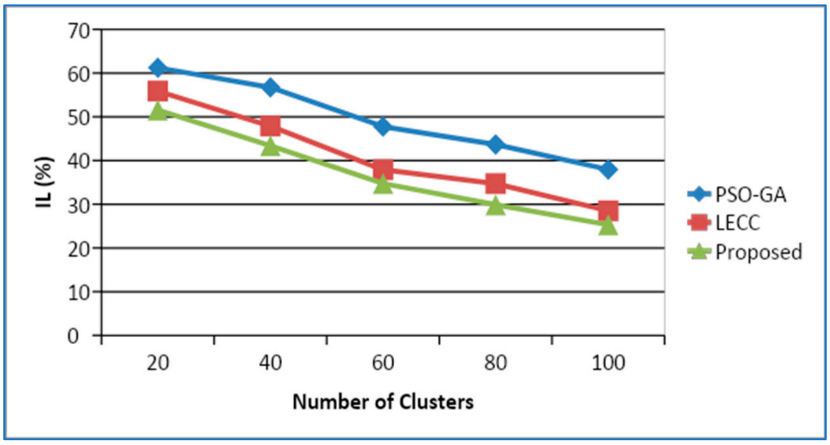

Figure 4. Effect of the number of clusters on *IL* (information loss) performance.

The outcome of *IL* with varying cluster sizes shows a similar trend to *DoA* performance, with contrasting effects. The increasing cluster size leads to minimal loss of sensitive information, while the higher number of clusters ensures a reduced number of at least K-anonymous users. Therefore, this ensures a minimal loss of sensitive information for a high number of clusters compared to a small number of clusters. Compared to the PSO-GA and LECC privacy preservation methods, the proposed anonymization approach reduced the *IL* ratio significantly, with improved *DoA*. The existing PSO-GA method mainly focuses on efficient cluster formation using a hybrid swarm intelligence model, without focusing on complete privacy preservation notions; thus, it showed the worst performances of all three methods for *DoA* and *IL*. The LECC method focuses on privacy preservation of the graph elements using clustering, but it relies on a single graph distance property for k-anonymity.

Furthermore, clusters ensuring k-anonymity can be extended to achieve the l-diversity notion in LECC, as k-anonymity and l-diversity suffer from more or less the same attribute disclosure that leads to *IL*. The proposed model effectively overcomes the challenges faced by LECC and PSO-GA by optimizing the clusters, using multiple graph properties for hybrid decision making and a one-pass algorithm to achieve t-closeness and l-diversity in order to reduce the *IL* caused in the LECC method.

PSO-GA takes a longer time for cluster formation compared to LECC and the proposed approach, because of its iterative optimization model with a longer convergence time. The LECC method achieved privacy preservation in the shortest time of all three techniques, as it relies on simple K-means clustering for ECC and LECC. Figure 5 shows a comparison of execution time with varying numbers of clusters.

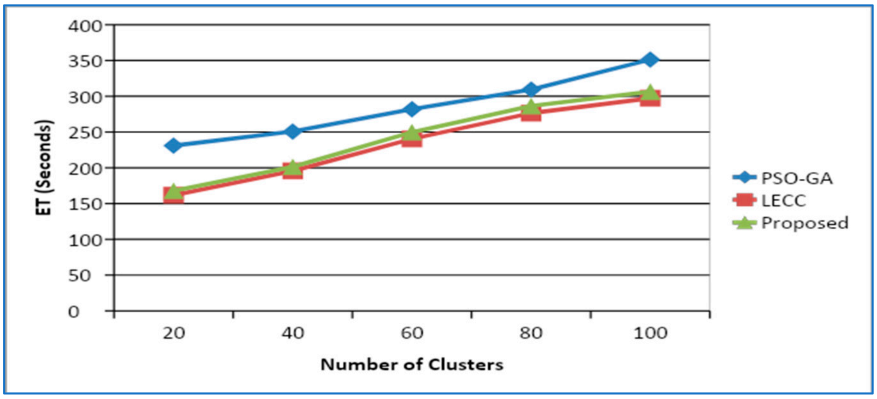

Figure 5. Effect of the number of clusters on *ET* performance.

The proposed model originated from the LECC approach, with the novel proposed data normalization. The inclusion of eccentricity graph properties and t-closeness functionality led to a slight increase in processing time compared to LECC. However, the proposed model significantly improved *IL* and *DoA* performances, and *ET* can be improved using parallel computing methods.

4.3. Variations in Density

This analysis aims to compare the performance of PSO-GA, LECC, and the proposed method based on the condition of the user density variation. We varied the user density from 2000 to 20,000 for that purpose, with a constant cluster size of 100. The clusters produced can be more significant if we give a greater number of users. Therefore, we ranged the user density from 2000 to 20,000, and conducted our investigations. Figures 6–8 demonstrate the outcome of this experiment for the parameters *DoA*, *IL*, and *ET*, respectively. Figure 6 indicates that *DoA* increased with the increase in user density.

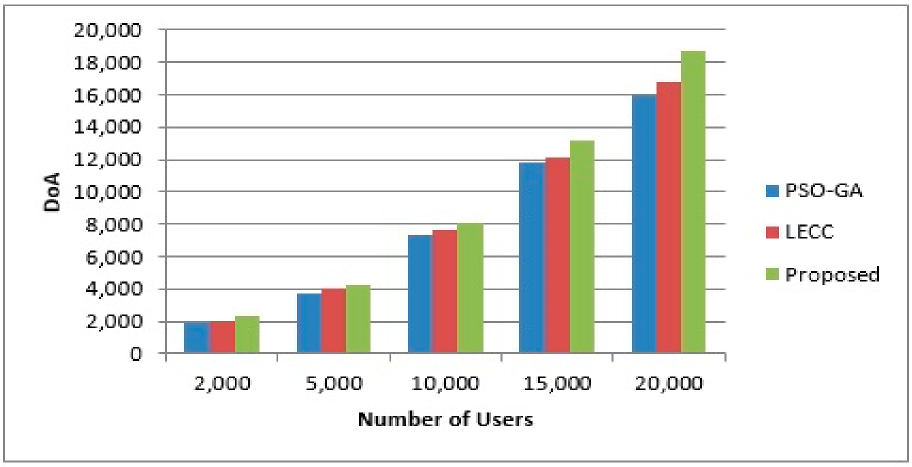

Figure 6. Effect of the number of users on *DoA* performance.

We observed that the increase in the *DoA* value was almost exponential, i.e., for 2000 users, the *DoA* outcome was approximately 2300, while for 20,000 users, the *DoA* outcome increased to approximately 18,000. The proposed model achieved the greatest *DoA* outcome for each user density scenario compared to the PSO-GA and LECC methods. The main reasons for this performance enhancement are given in the above section. In the proposed model, the initial phase ensures effective data normalization and initial cluster formation. Moreover, cluster optimization ensures the creation of the reliable and K-anonymous clusters using hybrid graph properties. Then, K-anonymous clusters are extended by providing the l-diversity and t-closeness privacy notions. Similarly, Figure 7 demonstrates the outcome of *IL* with varying user densities for each method.

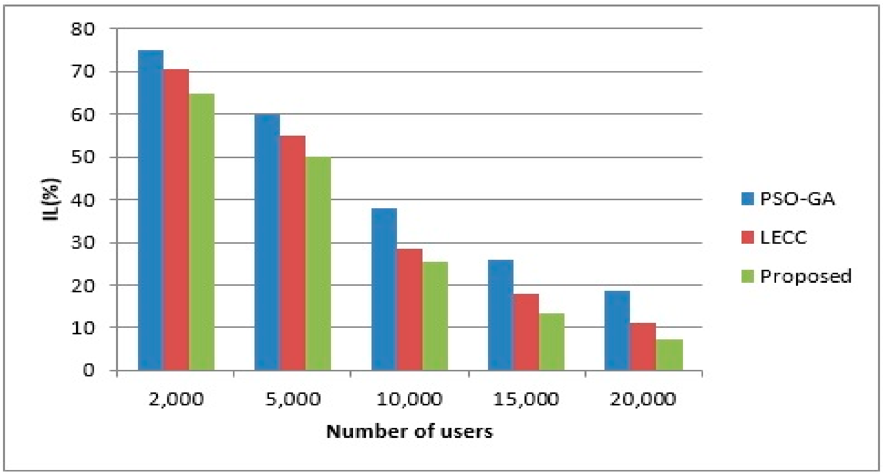

Figure 7. Effect of the number of users on *IL* performance.

The clusters formed are more meaningful due to the increase in the number of data points, leading to minimal *IL*. The proposed model achieved a reduction in *IL* by utilizing the multiple graph properties for cluster optimization, achieving privacy preservation of all of the elements of the OSN graph, as well as high-level privacy preservation, including t-closeness.

As shown in Figure 8, we observed a significant increase in execution time with increased user density. LECC showed the minimum *ET* compared to the proposed method and PSO-GA.

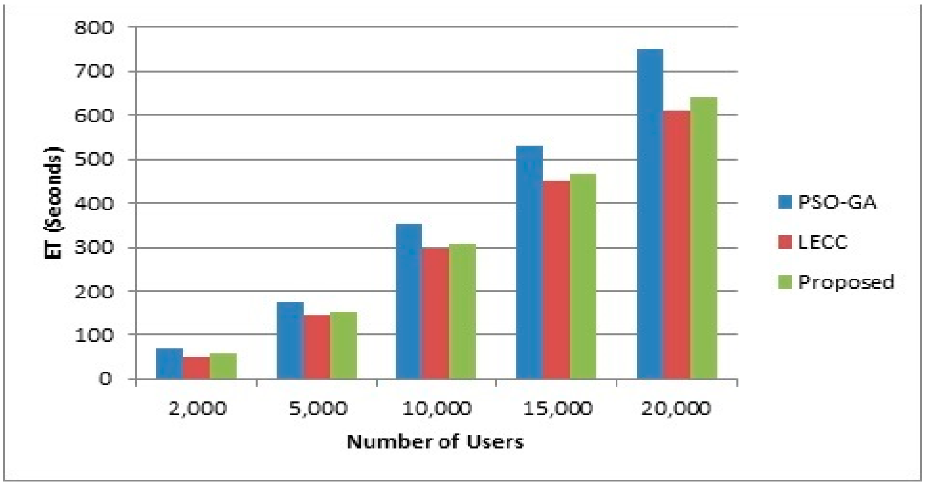

Figure 8. Effect of the number of users on *ET* (execution time) performance.

Table 5 presents the average outcomes for the parameters *DoA*, *IL*, and *ET*, using PSO-GA, LECC, and the proposed method.

Table 5. Comparative analysis of average performances.

	PSO-GA	LECC	Proposed
DoA	4861	5211	5880
IL	46.46	38.78	34.55
ET	330.29	272.31	283.85

4.4. Limitations

Although the experimental result show an improvement in performance over the existing methods, a few limitations of the proposed model need to be highlighted. Using an improved clustering mechanism, we achieved k-anonymity. As we kept the number of clusters fixed for the input dataset, this limits the scalability of the proposed model. We needed to manually adjust the total number of clusters according to the size of the dataset, leading to the erroneous and complex process of defining the correct number of clusters required to divide the input OSN. We investigated the proposed model on just one OSN dataset; in the future, we will extend our work to different OSN datasets. We did not test the security of the proposed model, and this needs to be investigated further with additional attacks similar to knowledge graph threats.

5. Conclusions and Future Directions

A novel anonymization model was proposed for OSNs to ensure minimal loss of sensitive structural information and a high degree of anonymity. The proposed model performs in three phases—namely, initial, cluster optimization, and privacy preservation. All three phases address the challenges of state-of-the-art techniques. The normalization of input OSN data using the statistical approach further improves the functionality of the proposed model. The hybrid score computation of each vertex leads to more reliable cluster formation than any single graph property. The one-pass algorithm not only achieved protection against well-known attacks such as attribute disclosure, similarity attack, etc., but also reduced the *IL*. The experimental results prove the effectiveness of the proposed model utilizing the Yelp OSN dataset and tested for various metrics (i.e., *IL*, *DoA*, and *ET*). Experiments were conducted to evaluate the efficiency of the proposed model by varying cluster sizes and user density. The average outcome of the proposed model improved the *DoA* by approximately 20% and reduced the *IL* by 10% compared to state-of-the-art methods. Several suggestions can be made for future research, such as (1) to improve the proposed model by dynamic clustering rather than fixed clustering, (2) to evaluate the performance of the proposed model using other OSN datasets, and (3) to evaluate the performance of the proposed model by introducing the other attacks aside from knowledge graph attacks.

Author Contributions: Conceptualization, A.S. and A.P.; methodology, R.G.; investigation, R.G.; resources, A.S., A.P. and R.G.; validation, R.G.; writing—original draft preparation, R.G.; writing—review and editing, A.S., A.P., R.J. and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Research Support Fund of Symbiosis International (Deemed University)".

Data Availability Statement: The Yelp dataset is open source. Available online: http://snap.stanford.edu/data/ (accessed online on 6 Jane 2021).

Acknowledgments: We would like to thank Symbiosis International (Deemed University) for providing research facilities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Novak, E.; Li, Q. A survey of security and privacy in online social networks. *Coll. William Mary Comput. Sci. Tech. Rep.* **2012**, 1–32.
2. Gangarde, R.; Sharma, A.P.A. DigitalCommons @ University of Nebraska-Lincoln Bibliometric Survey of Privacy of Social Media Network Data Publishing. *Libr. Philos. Pract.* **2019**, *3617*, 1–21.
3. Mishra, N.; Pandya, S. Internet of Things Applications, Security Challenges, Attacks, Intrusion Detection, and Future Visions: A Systematic Review. *IEEE Access* **2021**, *9*, 59353–59377. [CrossRef]
4. Poovarasan, G.; Susikumar, S.; Naveen, S. International Journal of Engineering Technology Research & Management. *Academia. Edu.* **2020**, *4*, 131–134.
5. Maple, C. Security and privacy in the internet of things. *J. Cyber Policy* **2017**, *2*, 155–184. [CrossRef]
6. Li, C.; Palanisamy, B. Privacy in Internet of Things: From Principles to Technologies. *IEEE Internet Things J.* **2019**, *6*, 488–505. [CrossRef]
7. Yang, Y.; Wu, L.; Yin, G.; Li, L.; Zhao, H. A Survey on Security and Privacy Issues in Internet-of-Things. *IEEE Internet Things J.* **2017**, *4*, 1250–1258. [CrossRef]
8. Sadeghian, A.; Zamani, M.; Shanmugam, B. Security threats in online social networks. In Proceedings of the 2013 International Conference on Informatics and Creative Multimedia, ICICM 2013, IEEE Computer Society, Kuala Lumpur, Malaysia, 4–6 September 2013; pp. 254–258.
9. Jaber, K.M.; Institute of Electrical and Electronics Engineers. Jordan Section; Institute of Electrical and Electronics Engineers. Region 8; Institute of Electrical and Electronics Engineers. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*; IEEE: Piscataway, NJ, USA, 2019; ISBN 9781538679425.
10. Prasad, M.R.; Kumar, S. Advance Identification of Cloning Attacks in Online Social Networks. *Int. J. Eng. Technol.* **2018**, *7*, 83. [CrossRef]
11. Devmane, M.A.; Rana, I.N.K. Privacy Issues in Online Social Networks. *Int. J. Comput. Appl.* **2012**, *41*.
12. Ali, S.; Islam, N.; Rauf, A.; Din, I.U.; Guizani, M.; Rodrigues, J.J.P.C. Privacy and security issues in online social networks. *Futur. Internet* **2018**, *10*, 114. [CrossRef]
13. Jamshidi, M.B.; Lalbakhsh, A.; Alibeigi, N.; Soheyli, M.R.; Oryani, B.; Rabbani, N. Socialization of Industrial Robots: An Innovative Solution to improve Productivity. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 832–837. [CrossRef]
14. Jamshidi, M.B.; Alibeigi, N.; Rabbani, N.; Oryani, B.; Lalbakhsh, A. Artificial Neural Networks: A Powerful Tool for Cognitive Science. In Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 1–3 November 2018; pp. 674–679. [CrossRef]
15. Jamshidi, M.; Lalbakhsh, A.; Talla, J.; Peroutka, Z.; Hadjilooei, F.; Lalbakhsh, P.; Jamshidi, M.; Spada, L.; La Mirmozafari, M.; Dehghani, M.; et al. Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment. *IEEE Access* **2020**, *8*, 109581–109595. [CrossRef]
16. Revathi, S.; Suriakala, M. An intelligent and novel algorithm for securing vulnerable users of online social network. In Proceedings of the 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 15–16 February 2018; pp. 214–219. [CrossRef]
17. Garcia, D. Leaking Privacy and Shadow Profiles in Online Social Networks. *Sci. Adv.* **2017**, *3*, e1701172. [CrossRef] [PubMed]
18. Friedland, G.; Choi, J. Semantic computing and privacy: A case study using inferred geo-location. *Int. J. Semant. Comput.* **2011**, *5*, 79–93. [CrossRef]
19. Ninghui, L.; Tiancheng, L.; Venkatasubramanian, S. t-Closeness: Privacy beyond k-anonymity and ℓ-diversity. In Proceedings of the International Conference on Data Engineering, Istanbul, Turkey, 15 April 2007; pp. 106–115. [CrossRef]

20. Majeed, A.; Lee, S. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access* **2021**, *9*, 8512–8545. [CrossRef]
21. Hay, M.; Miklau, G.; Jensen, D.; Towsley, D.; Li, C. Resisting structural re-identification in anonymized social networks. *VLDB J.* **2010**, *19*, 797–823. [CrossRef]
22. Zheng, X.; Cai, Z.; Li, Y. Data Linkage in Smart Internet of Things Systems: A Consideration from a Privacy Perspective. *IEEE Commun. Mag.* **2018**, *56*, 55–61. [CrossRef]
23. Zheleva, E.; Getoor, L. Preserving the Privacy of Sensitive Relationships in Graph Data. *Lect. Notes Comput. Sci. Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinform.* **2008**, *4890 LNCS*, 153–171. [CrossRef]
24. Sun, C.; Yu, P.S.; Kong, X.; Fu, Y. Privacy preserving social network publication against mutual friend attacks. In Proceedings of the IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013, IEEE Computer Society, Dallas, TX, USA, 7–10 December 2013; pp. 883–890. [CrossRef]
25. Cheng, J.; Fu, A.W.C.; Liu, J.; Association for Computing Machinary. K-isomorphism: Privacy preserving network publication against structural attacks. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD' 10), New York, NY, USA, 6–10 June 2010; pp. 459–470.
26. Zhang, Y.; O'Neill, A.; Sherr, M.; Zhou, W. Privacy-preserving network provenance. *Proc. VLDB Endow.* **2017**, *10*, 1550–1561.
27. Gangarde, R.; Sharma, A.; Pawar, A. Research opportunities in privacy of online social network data publishing. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 5095–5101.
28. Cai, Z.; He, Z. Trading private range counting over big IoT data. In Proceedings of the International Conference on Distributed Computing Systems(ICDCS), Dallas, Texas, USA, 7–9 July 2019; pp. 144–153. [CrossRef]
29. Cai, Z.; Zheng, X. A Private and Efficient Mechanism for Data Uploading in Smart Cyber-Physical Systems. *IEEE Trans. Netw. Sci. Eng.* **2020**, *7*, 766–775. [CrossRef]
30. Casas-Roma, J.; Herrera-Joancomartí, J.; Torra, V. A survey of graph-modification techniques for privacy-preserving on networks. *Artif. Intell. Rev.* **2017**, *47*, 341–366. [CrossRef]
31. Yan, Z.; Feng, W.; Wang, P. Anonymous Authentication for Trustworthy Pervasive Social Networking. *IEEE Trans. Comput. Soc. Syst.* **2015**, *2*, 88–98. [CrossRef]
32. Feng, W.; Yan, Z.; Xie, H. Anonymous Authentication on Trust in Pervasive Social Networking Based on Group Signature. *IEEE Access* **2017**, *5*, 6236–6246. [CrossRef]
33. Ghayvat, H.; Pandya, S.N.; Bhattacharya, P.; Zuhair, M.; Rashid, M.; Hakak, S.; Dev, K. CP-BDHCA: Blockchain-based Confidentiality-Privacy preserving Big Data scheme for healthcare clouds and applications. *IEEE J. Biomed. Health Inform.* **2021**, *2194*, 1–12. [CrossRef] [PubMed]
34. Liu, Q.; Wang, G.; Li, F.; Yang, S.; Wu, J. Preserving Privacy with Probabilistic Indistinguishability in Weighted Social Networks. *IEEE Trans. Parallel Distrib. Syst.* **2017**, *28*, 1417–1429. [CrossRef]
35. Siddula, M.; Li, L.; Li, Y. An Empirical Study on the Privacy Preservation of Online Social Networks. *IEEE Access* **2018**, *6*, 19912–19922. [CrossRef]
36. Qu, Y.; Yu, S.; Gao, L.; Zhou, W.; Peng, S. A hybrid privacy protection scheme in cyber-physical social networks. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 773–784. [CrossRef]
37. Liu, P.; Xu, Y.X.; Jiang, Q.; Tang, Y.; Guo, Y.; Wang, L.E.; Li, X. Local differential privacy for social network publishing. *Neurocomputing* **2020**, *391*, 273–279. [CrossRef]
38. Shao, Y.; Liu, J.; Shi, S.; Zhang, Y.; Cui, B. Fast De-anonymization of Social Networks with Structural Information. *Data Sci. Eng.* **2019**, *4*, 76–92. [CrossRef]
39. Yazdanjue, N.; Fathian, M.; Amiri, B. Evolutionary algorithms for k-anonymity in social networks based on clustering approach. *Comput. J.* **2021**, *63*, 1039–1062. [CrossRef]
40. Qian, J.; Li, X.Y.; Zhang, C.; Chen, L.; Jung, T.; Han, J. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. *IEEE Trans. Dependable Secur. Comput.* **2019**, *16*, 679–692. [CrossRef]
41. Siddula, M.; Li, Y.; Cheng, X.; Tian, Z.; Cai, Z. Anonymization in online social networks based on enhanced equi-cardinal clustering. *IEEE Trans. Comput. Soc. Syst.* **2019**, *6*, 809–820. [CrossRef]
42. Zhao, P.; Huang, H.; Zhao, X.; Huang, D. P^3: Privacy-Preserving Scheme Against Poisoning Attacks in Mobile-Edge Computing. *IEEE Trans. Comput. Soc. Syst.* **2020**, *7*, 818–826. [CrossRef]
43. De Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1–5. [CrossRef] [PubMed]
44. Cai, Y.; Zhang, S.; Xia, H.; Fan, Y.; Zhang, H. A Privacy-Preserving Scheme for Interactive Messaging over Online Social Networks. *IEEE Internet Things J.* **2020**, *7*, 6817–6827. [CrossRef]
45. Gao, T.; Li, F. Protecting Social Network with Differential Privacy under Novel Graph Model. *IEEE Access* **2020**, *8*, 185276–185289. [CrossRef]
46. Qu, Y.; Yu, S.; Zhou, W.; Chen, S.; Wu, J. Customizable Reliable Privacy-Preserving Data Sharing in Cyber-Physical Social Networks. *IEEE Trans. Netw. Sci. Eng.* **2021**, *8*, 269–281. [CrossRef]
47. Machanavajjhala, A.; Kifer, D.; Gehrke, J.; Venkitasubramaniam, M. ℓ-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **2007**, *1*, 3. [CrossRef]

48. Stanford Large Network Dataset Collection. Available online: http://snap.stanford.edu/data/ (accessed on 13 January 2020).
49. Domingo-Ferrer, J.; Mateo-Sanz, J.M. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 189–201. [CrossRef]