KMFC-GWO: A Hybrid Fuzzy-Metaheuristic Algorithm for Privacy Preserving in Graph-based Social Networks

Saeideh Memarian
Departamento Tecnología Electrónica, Universidad de Sevilla, Sevilla, Spain
Email: saemem@alum.us.es

Andreea M. Oprescu
Departamento Tecnología Electrónica, Instituto de Ingeniería Informática, Universidad de Sevilla, Sevilla, Spain
Email: aoprescu@us.es

Betsaida Alexandre-Barajas
Departamento Tecnología Electrónica, Universidad de Sevilla, Sevilla, Spain
Email: balexandre@us.es

Gloria Miró-Amarante
Departamento Tecnología Electrónica, Universidad de Sevilla, Sevilla, Spain
Email: mmiro@us.es

M. Carmen Romero-Ternero
Departamento Tecnología Electrónica, Instituto de Ingeniería Informática, Universidad de Sevilla, Sevilla, Spain
Email: mcromerot@us.es

Abstract- **In recent years, the proliferation of social networks has been remarkable, providing a rich source for data mining endeavours. However, a significant challenge lies in safeguarding the privacy of individuals while sharing these databases publicly. Current approaches such as K-anonymity, L-diversity, and T-closeness, are commonly employed for data anonymization in social networks. However, these techniques entail considerable information loss due to random alterations in the graph-based datasets. To address these limitations, this paper introduces a new anonymization technique called KMFC-GWO, which combines K-Member Fuzzy Clustering with Grey Wolf Optimizer. This integrated method is designed to strengthen the anonymized graph against a range of threats, including identity, attribute, link disclosure, and similarity attacks, while significantly reducing information loss. Within the KMFC-GWO framework, K-member fuzzy c-means clustering is utilized to create well-balanced clusters, each meeting the K-anonymity requirement. Subsequently, the Grey Wolf Optimizer is applied to optimize cluster formation and effectively anonymize the social network graph. The objective function is carefully crafted to minimize both clustering error and information loss, while ensuring adherence to predefined anonymity criteria.**

Index Terms **- privacy preserving, K-anonymity, L-diversity, T-closeness, fuzzy clustering, Grey Wolf Optimizer (GWO), Graph-based GWO.**

I. Introduction

Sharing social network data with data miners necessitates a careful balance between privacy protection and knowledge retention. Anonymizing techniques, such as K-anonymity (KA) and its extensions like L-diversity (LD) and T-closeness (TC), are commonly employed to mitigate privacy risks while preserving data utility [1-3]. KA aims to group users into clusters with at least *K* members to prevent identity disclosure, though it doesn't safeguard against attribute or link disclosure. LD addresses attribute disclosure by ensuring each cluster contains diverse attribute values, while TC focuses on maintaining the global attribute distribution within clusters to mitigate similarity attacks.

Privacy threats in social network data publishing encompass different attacks such as identity, attribute, and link disclosure [3]. Identity disclosure exposes a user's identity, while attribute disclosure reveals sensitive user attributes and link disclosure unveils sensitive relationships between users. LD and TC complement KA by addressing attribute and similarity attacks respectively, enhancing overall privacy protection in anonymized datasets [4].

This research concentrates on protecting social network data publication from a range of threats, including revealing identities, disclosing attributes/links, and potential similarity attacks. To address these challenges, we introduce a hybrid anonymization approach called KMFC-GWO, which combines K-member Fuzzy Clustering (KMFC) with Grey Wolf Optimizer (GWO). The main objective of KMFC-GWO is to fortify the privacy of graph-based social networks while reducing information loss. Our approach employs a modified variant of fuzzy c-means (FCM), referred to as KMFC, to create well-balanced clusters containing a minimum of K members in each cluster, thereby fulfilling the KA criterion. Furthermore, an optimization problem is formulated and solved using GWO, to satisfy the LD and TC conditions.

ISBN:978-84-09-62140-8

II. Proposed KMFC-GWO Algorithm

In this paper, we present a hybrid fuzzy-metaheuristic graph-based privacy-preserving approach for social networks, termed KMFC-GWO, which combines KMFC with GWO, offering a highly effective solution for both balanced clustering and anonymization in social networks. Our method addresses the optimization problem of privacy preservation within graph-based social networks by integrating KMFC, considering KA, LD, and TC criteria. Subsequently, GWO is leveraged to tackle this optimization challenge, aiming to maximize anonymity while minimizing information loss in the published graph. Specifically, the objective function of GWO is formulated to minimize information loss and uphold anonymity conditions (KA, LD, and TC). The proposed KMFC-GWO approach effectively safeguards the anonymized social network graph against identity, attribute disclosure, and similarity attacks by enforcing these three constraints.

Consider a social network represented as a graph, where the edges (referred to as G) have dimensions $N \times N$, and the attribute data matrix (referred to as A) has dimensions $N \times M$. Here, N denotes the number of users, and M denotes the number of personal characteristics associated with each user. Each user, denoted as i (where i ranges from 1 to N), possesses a vector of edges indicating connections with other users and a vector of features represented by M. The graph matrix, G, is binary: $G_{ij}=1$ signifies the existence of an edge between users i and j, while $G_{ij}=0$ denotes the absence of such an edge. Furthermore, A_{im} represents the m-th personal characteristic of user i. The aim of the anonymization procedure is to alter the original data to generate a modified edge graph (denoted as G_{new}) and a modified data matrix (denoted as A_{new}).

In the KMFC-GWO approach, the K-anonymity requirement is met through KMFC forming C clusters, each comprising a minimum of K users. Then, the GWO algorithm is utilized to address the anonymity criteria of LD and TC, while simultaneously reducing information loss in the published social network graph.

A. Satisfying KA using KMFC

As mentioned above, to satisfy the KA condition, we apply a KMFC technique on the FCM. Our KMFC method utilizes a customized version of the FCM algorithm to construct balanced clusters with at least K members, satisfying the KA condition. To achieve this purpose, first, we generate an initial clustering using FCM, and then, the generated clusters are balanced to satisfy the KA condition.

The FCM algorithm was initially introduced by Bezdek in 1981 [5]. Unlike traditional clustering techniques such as c-means and k-means, FCM assigns a degree of membership for each data point to every cluster. The objective of FCM is to minimize the total distance between the data points and the cluster centroids. Its primary aim is to partition N data points into C distinct clusters. Each data point can be characterized by two feature vectors: binary edges and personal attributes.

Generally, the traditional KA-based clustering methods suffer from two drawbacks:

- *Number of clusters*: Selecting the best value for the number of clusters C is a challenging issue.
- *Unbalanced clusters*: In certain clusters, the sample count may be lower than K, while in others, there could be significantly more members than K.

To handle the first problem, we consider the number of clusters in such a way that minimizes the Cernability AVG (*CAVG)* criterion [4], as formulated in Eq. (1). The *CAVG* metric (where $CAVG \geq 1$) quantifies the degree of cluster balance, with values closer to 1 indicating higher balance among clusters. In the case of K-member clustering, *CAVG* tends to approximate 1, i.e., $CAVG \approx 1$.

$$CAVG = \frac{N}{C \times K} \quad (1)$$

Since N and K are fixed parameters, we consider $C \approx N/K$. To have somewhat a free level for clustering, we have set $C=0.9 \times N/K$ in our simulations to have around 10% more samples on average in each cluster.

Furthermore, to satisfy the KA condition, we present a revision phase on the initial clustering results of the FCM algorithm. Our proposed KMFC algorithm is provided in Algorithm 1.

B. Satisfying LD and TC using GWO

Once users are clustered into C balanced clusters (each containing at least K users) to meet the KA requirement, the GWO algorithm is applied to further anonymize the social network graph, adhering to the requirements of LD and TC. This anonymization process is conducted simultaneously on both the graph matrix G and the attribute matrix A. At the graph level, anonymization involves randomly adding or removing edges between users, while at the attribute level, it entails randomly altering the values of user features.

GWO is a swarm intelligence algorithm characterized by its balanced exploration and exploitation capabilities. It was initially introduced in 2014 by Mirjalili et al. [6], drawing inspiration from the hunting behaviour of grey wolves. The algorithm commences by randomly initializing a population of grey wolves. In each iteration, the fitness of the current population is evaluated using a customized fitness function tailored to the particular application. Subsequently, the entire population undergoes adjustments through two phases: attacking prey (exploitation) and searching for prey (exploration). These main steps of GWO including random generation of initial population, objective function evaluation, and population updating, are described in the following:

- *Generation of initial population*: As seen in Fig. 1, a feasible solution *SOL* can be represented as a matrix of $N \times (N+M) = N \times F$, where each row i represents the edge and attribute modifications in G and A matrices. In the case of edge modification, $SOL_{i,j}=1$ if the edge between nodes i and j is modified (adding a new edge between nodes i and j or removing a previously connection link between nodes i and j). Furthermore, for the attribute modification, $SOL_{i,N+m}=1$ if the value of the personal feature m is randomly changed for the data of user i.

Algorithm 1. Proposed clustering revision in KMFC algorithm.

Inputs: Initial clustering solution of FCM, graph matrix G, attribute matrix A, and parameters *C* and *K*

Output: Revised Clusters

Begin

1 Divide the initial clusters into three sets (UL, BL, and OL):
- UL (Underload Clusters): with less than *K* samples
- BL (Balanced Clusters): with the number of samples between *K* and 1.1×*K*
- OL (Overload Clusters): with more than 1.1×*K* samples

2 Consider all clusters within BL as final revised clusters.

3 Revising clusters within OL and UL:

for1 $k \in$ UL

for2 $k' \in$ OL

Calculate Euclidian distance between the centroid of clusters *k* and *k'*: $d_{s_k,s_{k'}}$

end for2

Sort the samples of cluster *k'* according to their Euclidian distance to the center of cluster *k*

while *number of samples within cluster k be equal to K*

Transfer the nearest sample of cluster *k'* to cluster *k*

end while

4 Revise the current state of all BL, UL, and OL clusters

end for1

5 Generate the final revised BL clusters, each with at least *K* samples

End

6 Generating the clusters of KMFC algorithm

	1	*2*	*…*	*N*	*N+1*	*N+2*	*…*	*F=N+M*
1	1	0	…	0	1	1	…	1
2	0	1	…	1	1	0	…	0
⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋱	⋮
N	1	0	…	0	1	0	…	0

Fig. 1. Representation of a feasible solution in GWO.

- *Objective function evaluation*: There is a trade-off between the information loss (due to the modifications in G and A) and the anonymity level. The more changes in the G and A, the higher anonymity level, by accepting more information loss. To minimize the information loss within the published social network data, we present an objective function to minimize the summation of the information loss of the modified G and A, denoted by IL_G and IL_A, while satisfying the constraints of KA, LD and TC.
- *Population updating*: Grey wolves exhibit a hierarchical social structure comprising various levels, including alpha, beta, delta, and omega [6]. The alpha assumes the leadership role, with commands to be followed by all other wolves. The beta acts as an advisor to the alpha, supporting the alpha's directives and offering feedback. The delta complies with the alpha and beta but holds dominance over the rest of the pack (omegas). During each iteration of the algorithm, the fitness function assesses the quality of all grey wolves, which are then sorted based on their fitness values, ranging from best to worst.

After reaching the maximum iterations of GWO, the global best solution is considered as the final anonymized graph- based social network, which is ready to be sent to the data miners for further processing.

III. Conclusion

In this paper, a novel anonymization method (KMFC-GWO) has been presented. It combines a K-Member fuzzy clustering with a metaheuristic-driven optimization algorithm to enhance the resilience of anonymized graph-based social networks against various threats while minimizing information loss of the published attribute and graph matrices. By presenting a K-member variant of the fuzzy c-means clustering algorithm to achieve K-anonymity and GWO to further optimize the anonymity conditions, the proposed framework effectively anonymizes graph-based social networks. As a future research direction, experiments conducted on real-world datasets from prominent social networks such as Facebook or Twitter, could be included to assess the effectiveness of the proposed KMFC-GWO algorithm.

Acknowledgment

This work was partially funded by Grant PID2022-141045OB-{C41,C42,C43} funded by MCIN/AEI /10.13039 /501100011033/ and FEDER A way of making Europe in ARTIFACTS Project: generAtion of Reliable syntheTIc health data for Federated leArning in seCure daTa Spaces.

References

[1] Gangarde, R., Sharma, A., Pawar, A., Joshi, R., & Gonge, S. (2021). Privacy preservation in online social networks using multiple-graph-properties-based clustering to ensure k-anonymity, l-diversity, and t-closeness. *Electronics*, *10*(22), 2877.

[2] Gangarde, R., Sharma, A., & Pawarq, A. (2023). Enhanced clustering based OSN privacy preservation to ensure k-anonymity, t-closeness, l-diversity, and balanced privacy utility. *Computers, Materials & Continua*, *75*(1), 2171-2190.

[3] Panda, B. S., Naveen Kumar, M., & Patro, S. (2023, April). Apply Rough Set Methods to Preserve Social Networks Privacy—A Review. In *Proceedings of 3rd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2022* (pp. 427-436). Singapore: Springer Nature Singapore.

[4] Langari, R. K., Sardar, S., Mousavi, S. A. A., & Radfar, R. (2020). Combined fuzzy clustering and firefly algorithm for privacy preserving in social networks. *Expert Systems with Applications*, *141*, 112968.

[5] Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York.

[6] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in engineering software*, *69*, 46-61.