International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

A Hybrid Probabilistic Privacy Preserving Based Community Detection Model on Online Social Networking Data

[1]Shamila. M, [2]G. Rekha, [3]K. Vinuthna Reddy

[1]Research scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur, A.P, India

[2]Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Aziz Nagar, Hyderabad, Telangana, India.

[3]Associate Professor, Department of CSE, Neil Gogte Institute of Technology, Uppal, Hyderabad, Telangana, India.

Abstract— Privacy preserving plays a vital role on the online social networking sites due to high dimensionality and data size. Community detection is used to find the social relationships among the node edges and links. However, most of the conventional models are difficult to process the community structure detection due to high computational time and memory. Also, these models require contextual weighted nodes information for privacy preserving process. In order to overcome these issues, an advanced probabilistic weighted based community detection and privacy preserving framework is developed on the large social networking data. In this model, a filter based probabilistic model is developed to remove the sparse values and to find the weighted community detection nodes and its profiles for privacy preserving process. Experimental results show that the filter based probabilistic community detection framework has better efficiency in terms of normalized mutual information, Q, rand index and runtime (ms).

Keywords- Social network dataset, privacy preserving, community detection**.**

I. INTRODUCTION

J.A.Barnes coined the term "social network" in 1954 to describe a social structure made up of nodes connected by edges that represent one or more types of interdependency. OSNs come in a variety of shapes and sizes to meet the needs of the user. Facebook, Linkedin, Youtube, Instagram, Twitter, and Tumblr are some of the most commonly used services that users use on a daily basis. Despite the fact that the services provided by various OSNs are designed for various purposes such as networking, microblogging, video sharing, and so on, they all share a common set of core features. The number of OSN users and the services provided by OSNs are growing at an unprecedented rate. OSNs allow users to form virtual friendships with both known friends and strangers who share a common interest. Users are motivated to share a lot of information in the OSN for a variety of reasons, including the desire to connect with others, increase the number of connections, gain fame, the inherent trust they have in their affinity group, and herding behaviour. Today's online world thrives on the concept of sharing, and everyone, whether intentionally or unintentionally, contributes some information to the online world. Intentional sharing occurs when a user shares whatever they know and believe is appropriate to share with a group in a specific context. The user's online interactions are captured, aggregated, and correlated with other available information to create a profile of the user in unintentional sharing. In today's web, sharing is a highly encouraged feature. While sharing personal information can be beneficial, it also exposes users to a variety of privacy risks. As a result, online social media applications should provide services that strike the right balance between the user's utility and privacy needs. Online Social Networks (OSNs) have emerged as a transformative technological development over the past decade, captivating the interest of millions of users worldwide since their inception. Notable OSNs like Facebook, Twitter, LinkedIn, Pinterest, MySpace, among others, have witnessed unprecedented popularity due to their ability to cater

to diverse user interests and behaviors [1]. In the realm of safeguarding privacy in OSNs, extensive research has been conducted, although the concept of privacy remains multifaceted. Westin's privacy theory examines how individuals safeguard their privacy by selectively controlling others' access to their personal information. Whether at an individual, group, or institutional level, privacy is asserted when individuals or organizations exercise authority over when, how, and to what extent their information is shared with others.Within the context of online social networks, social entities, or nodes, are interconnected through links, symbolizing relationships or communication between them. The exponential growth in the popularity of OSNs has transformed them into vibrant platforms where millions of users freely express their thoughts, beliefs, and ideas [2]. Activities such as advertising, blogging, review collection, and political awareness campaigns thrive within this dynamic digital environment. OSNs like Facebook, WhatsApp, Twitter, Flickr, and Instagram have seamlessly integrated into our daily lives, fostering connectivity on a global scale. Notably, the essence of social networks lies in their constant evolution and expansion, with the primary objective being the proliferation of connections. This expansion not only enhances the utility of OSN services but also facilitates the rapid dissemination of information—a mission evident in popular platforms like Facebook, aiming to "bring the world closer together," and LinkedIn, aspiring to "connect the world's professionals for enhanced productivity and success."The formation of new connections within OSNs signifies the emergence of social relationships among users who share common interests, backgrounds, or real-life connections. Analyzing the intricate factors influencing the growth of social networks is a challenging endeavor. A simplified approach to this complex task involves examining the connections between specific nodes, giving rise to the link prediction problem. This problem revolves around determining the most probable associations between disconnected nodes in a social network based on the current state of nodes and connections. Link prediction holds immense significance across all social media platforms, where the goal is to foster the creation of as many connections as possible [3].Prominent examples of link prediction features can be observed in LinkedIn's "People You May Want to Hire," Facebook's "People You May Know," and Google+'s "You May Know" recommendations. These features offer users valuable suggestions, such as job opportunities on LinkedIn and potential friends on Facebook and Google+. Remarkably, the "People You May Know" feature on Facebook has been particularly effective in establishing a "significant fraction" of connections among users who were previously disconnected.It is worth noting that OSNs like Facebook typically require users to register using their real names and email addresses, reinforcing the importance of authenticity and accountability within these digital ecosystems. OSN services allow users to maintain virtual connections with their real-life friends. Users on dating sites such as Match.com, Tinder, and others are not required to reveal their true identities. In those accounts, pseudo-anonymity is used to manage user privacy. A Facebook user could be linked to other OSNs, launching a linking attack. Even if the user has a real identity in the linked OSN, the adversary can gather additional information about the user by linking the profiles. It's possible that information shared in one context isn't appropriate to reveal in another[4]. The user's profile pictures, attributes, or friends list could be used to identify and link them to other OSN sites. Users on online social networks (OSN) tend to be more trusting of spam messages when they receive them from people in their friend lists. These unwanted messages are typically distributed to a large audience when a spammer gains access to an email list. The task is made simpler when a third-party agency (TPA) unintentionally shares user attributes, including email addresses. Even if email addresses are not directly disclosed, basic profile information is often shared, making it relatively easy to deduce email addresses by cross-referencing external data sources. Spamming is often used to promote products but can also be combined with more malicious activities like phishing, putting users at risk. The persistence of cyberbullying remains a significant concern in the realm of OSN, necessitating proactive measures to prevent such messages from being posted.In the domain of online social networks, link prediction and recommendation have emerged as crucial components of the ecosystem. Over the past few decades, this area of research has gained increasing importance. An essential step in link prediction is computing the similarity score (or proximity score) between pairs of disconnected nodes. This score is employed to determine whether these nodes should be recommended for connection, thereby influencing whether they become linked.Companies engaged in data mining take steps to encrypt their data before sharing it with third parties. There is a need for the development of privacy tools that are agnostic to specific tasks and applications, capable of safeguarding data containing sensitive numerical and nominal attributes while minimizing processing time. Ensuring that traditional Privacy-Preserving Data Mining (PPDM) models remain robust against various privacy attacks and that data utility is preserved is essential. Additionally, the development of an adversarial model to assess the risk of data disclosure within the context of traditional PPDM is required.

While generating K-anonymous tables is a common sanitization technique, maintaining L-diversity throughout

sensitive attributes is crucial to prevent homogeneity attacks within equivalence classes. However, preserving diversity alone does not suffice to protect sensitive attributes from proximity and divergence breaches. Thus, in this research, it is decided not to alter quasi-identifying attributes while preserving sensitive information. In cases where vertically partitioned data is available, parties with sensitive attributes can share them for mining after applying the proposed transformation. The data miner consolidates the transformed tables and conducts mining tasks before sharing the results with stakeholders. Furthermore, an adversarial model is proposed to assess the risk of data disclosure in the context of the transformation-based method. The primary objective is to extract valuable information from extensive databases while upholding privacy standards. This entails factoring in external noise, data exchange, consolidation, and elimination.As businesses collect and process increasingly large volumes of personal data, the complexity of their information security systems grows. Scalability, heterogeneity, evolution, evaluation, collective intelligence, and privacy pose common challenges in the field of community detection in social media data[5-7].

Scalability: The volume of online social media content is expanding rapidly, with billions of nodes and connections. Managing and processing this expanding network require exponentially more resources. Traditional community detection methods are designed for smaller networks, often dealing with tens of thousands of nodes.

Heterogeneity: Raw social media networks exhibit diverse edges and vertices, often represented using hypergraphs or k-partite graphs. These complex structures are not well-suited for many community detection algorithms, leading to the extraction of simplified network forms that capture only a portion of the interactions.

Evolution: Social media networks are highly dynamic, and community detection methods need to account for their evolving nature. Traditional approaches assume static networks, but incorporating time dynamics is essential for accurate analysis.

Evaluation: Assessing the performance of community detection methods is challenging due to the absence of reliable ground-truth data. Current evaluation methods often rely on manual inspection, which is labor-intensive and may not be comprehensive, particularly for small networks.

Collective Intelligence: Collective intelligence, derived from user-generated content like comments, reviews, and ratings, is valuable in various contexts. Efficiently extracting intelligence from such data is a complex task.Privacy: Privacy is a significant concern in social media, with platforms like Facebook and Google frequently at the center of privacy debates. Protecting user data and ensuring privacy is a critical consideration in the community detection process. Anonymization alone does not guarantee privacy protection. A secure and trustworthy system is dangerous when private information is involved. As a result, much valuable information is withheld due to security concerns. The majority of existing community detection studies focus on benchmark datasets, with only a few examining the crawled dataset. The strength of a specific network, as well as its community, sub-community, and overlapping community detection, were not thoroughly examined. Despite extensive studies on the community structure of real networks, the literature survey reveals that there is no consensus on a definition for a community. As a result, assessing the quality of communities discovered by various community detection algorithms is difficult. There are also issues with using community detection for network data analysis due to the scalability of existing algorithms. Data generated by social media sites is noisy, dispersed, unstructured, and dynamic. Because of the scalability, heterogeneity, and security of the social network, detecting communities is a difficult and time-consuming task. The proposed work is motivated by a desire to conduct research on community detection in social networks and to develop an efficient community detection algorithm by converting the social network analysis problem into a graph partition problem because networks are represented as graphs. On a sample twitter network of a sportsperson, various community detection approaches such as community-based, sub-community-based, and overlapping community detection were tested, and two new approaches for overlapping community detection were proposed[8].

II. RELATED WORKS

Privacy-preserving techniques can be described as a collaborative effort among multiple parties who possess sensitive databases and wish to engage in data mining while keeping their data confidential. The objective is to collectively obtain results from the database without exposing critical information to other involved parties [9]. This concept of safeguarding data privacy in the context of data mining is often referred to as Privacy-Preserving Data Mining (PPDM), encompassing three primary aspects:

Input Privacy: This aspect focuses on protecting the confidentiality of input data.

Output Privacy: It concerns the safeguarding of the mined results, ensuring that they are only shared with the intended recipients.

Minimizing Discrepancy: Minimizing the difference between the results obtained from mining the original data and those obtained from mining transformed data.

Traditionally, two common methodologies for mining distributed databases are employed:

Centralized Approach: In this method, data is centralized to a single location before the mining process occurs.

Distributed Approach: Here, replicas of the data are created at each site, and subsequently, all replicas are transferred to a common central location [10].

The choice between these approaches involves a trade-off. The centralized approach tends to be faster, while the distributed approach excels in terms of accuracy and preserving privacy. It is crucial to note that there is often a trade-off between accuracy and privacy in PPDM algorithms and models. The success of these algorithms can be assessed based on various criteria, including accuracy, speed, data utilization, cost, resilience against privacy attacks, and the level of privacy achieved. Different algorithms or solutions may perform optimally in specific contexts but may fall short in others. There is no one-size-fits-all solution in the field of PPDM, as each approach comes with its own set of advantages and trade-offs [11].

The methodology employed in this study aims to ensure the privacy of classification rules through the addition of dummy transactions into the database. Initially, classification rules are extracted from the dataset, and subsequently, sensitive rules are identified by the data owner. These identified sensitive rules are then prioritized based on the number of attributes they involve and arranged in descending order. By scanning the database, the transactions supporting these sensitive rules are identified, and modifications are made to protect them. It's important to note that not all attributes of a sensitive rule need to be altered; only a specific set of attributes is modified.However, a drawback of this algorithm is the introduction of false transactions, which increases the dataset's size, thereby extending the time required for the transformation process and reducing the utility of the resulting dataset[12-15]. To address this, an alternative technique was introduced, involving the addition of noise during the preservation of privacy in classifiers based on decision trees. This noise is applied by shuffling or altering attribute values in tuples based on the class label of the leaf in the decision tree. This approach aims to maintain the similarity between the transformed dataset and the original dataset, but it does come at the cost of reduced utility in the resulting classification model due to the added noise.In another approach, researchers proposed a technique for safeguarding extracted knowledge while minimizing side effects. This technique incorporates randomization and k-anonymization. The process is divided into two parts, with randomization applied to the dataset using an attribute transitional probability matrix in the first part and subsequent application of k-anonymity to the randomized data. However, it should be noted that k-anonymity can sometimes lead to issues with dataset homogeneity.Furthermore, a data modification technique was introduced to protect sensitive classification rules, focusing on a category known as associative classification rules. This technique involves the deletion of selected tuples from the original dataset and reducing the support of sensitive rules below a minimum support threshold. The evaluation of this methodology considers the utility of the transformed dataset, the loss of non-sensitive rules, and the emergence of new rules. Removing tuples that meet sensitive rule criteria can significantly impact the knowledge derived from the transformed dataset compared to the original dataset[16].Privacy preservation has become a critical concern, particularly in the context of personal information and sensitive data. Various methods, including fuzzy-based approaches and weight algorithms, have been proposed to address this challenge. These methods leverage techniques such as infrequent pattern mining and multiple support measures to protect privacy.Additionally, researchers have explored link prediction techniques based on keyword matching and text similarity in different networks. These techniques calculate node pair similarity and have demonstrated success in various scenarios. However, concerns remain regarding the privacy controls and data leakage in online social networks (OSNs). Privacy controls in OSNs are currently insufficient, leading to potential privacy breaches. Furthermore, cyberbullying detection solutions face limitations as they rely on alerts and may not prevent the spread of harmful content.While numerous network topological property indices and link prediction techniques have been proposed, there is a need to explore the correlation between these indices and link prediction methods further. Additionally, combining node structure and profile information for link prediction is an area that requires more investigation, as existing techniques often rely on threshold values that may be challenging to determine in dynamic social networks.In the field of Privacy-Preserving Data Mining (PPDM), several techniques have been developed to transform data while maintaining privacy. These methods aim to strike a balance between privacy preservation and data utility. Some approaches involve item-restriction designs to limit noise addition and dataset removal. Additionally, the use of the SIF-IDF approach and tree-based structures have been explored to improve performance and reduce database scans. Finally, it is important to acknowledge the widespread sharing of personal information on social networking sites, which poses privacy risks. This highlights the need for robust privacy controls and data protection measures in the context of online social networks[17-19].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

III. Filtered Based Online Social Networking Data Privacy Preserving Approach

In this work, a data filter based community detection based privacy preserving is developed on the online social networking datasets. Initially, input data is filtered using the pre-processing algorithm. Since, each social networking input data contains numerical attributes, it is necessary to implement a hybrid data filtering approach in order to filter the sparse values in the data. In the second phase, each filtered data is given to the privacy preserving model using the density community clustering approach. In this framework, a probabilistic clustering approach is developed on the filtered dataset in order to detect the communities in the dataset. Here, each cluster represents the single community and the data points within the cluster represents the intra-cluster variance. The objects between the clusters represents the inter-cluster variation due to change in the data attributes. Finally, privacy preserving approach is implemented on the sensitive attribute which contains the user's profile information. The overall proposed framework is presented in figure1. Instead of traditional k-anonymization algorithms, proposed privacy preserving approach anonymize the profile sensitive attribute using the privacy preserving approach

Algorithm 1: PPDM OSN Graph data pre-processing

Step 1: Load the input datasets (MD) containing multiple sources of data.
Step 2: For each training data (D[i]) in the input datasets:
Step 3: Begin a loop to process each training data.
Step 4: For each record (I[r]) in the current training data:
Step 5: Begin a loop to process each record.
Step 6: For each attribute in the current record:
Step 7: Begin a loop to process each attribute.
Step 8: Check if the attribute (Aτ[I]) is of type Continuous and is not empty (φ).
Step 9: If the attribute meets the conditions, perform the following:
Step 10: Replace the attribute value (Aτ[I]) using Equation (1):
Aτ[I] = (Aτ[I] - (Mx(Aτ) + Mn(Aτ)) / 2) / (Max_c(Aτ) - Min_c(Aτ)) // Equation (1)
Step 11: End the conditional check.
Step 12: If the attribute (Aτ[I]) is of type Categorical and is not empty (φ):
Step 13: Perform the following:
Step 14: Replace the attribute value (Aτ[I]) using Equation (2):
Aτ[I] = Σ[F(Aτ[i] / c_m) - F(c_m)] / (Mx * Prob(Aτ[i] / c_m)) // Equation (2)
Step 15: End the conditional check.
Step 16: End the loop for processing attributes.
Step 17: End the loop for processing records.
Step 18: End the loop for processing training data.
Step 19: End the loop for processing input datasets.

In the algorithm 1, each data instance from the distributed data source is pre-processed using the min-max measure and max probabilistic measure. In each data source, each instance is pre-processed by using numerical or nominal type of attribute. If the attribute type is continuous, then the equation (1), is used to fill the missing value of the numerical attributes. Similarly, in case of nominal or categorical attributes, conditional probability of the attribute is used to replace the sparsity values.

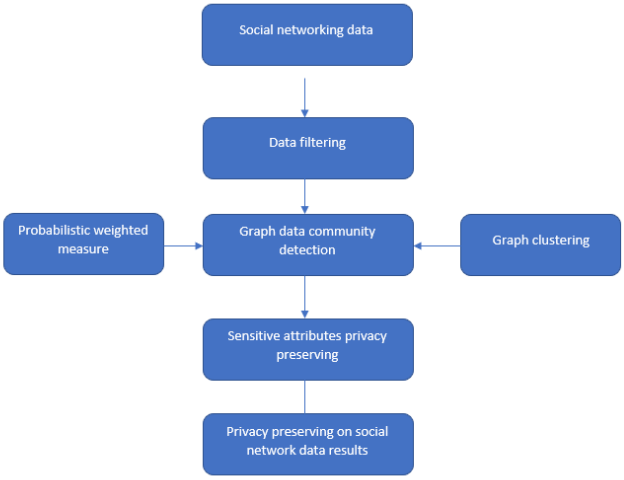

Figure 1: Proposed framework

Algorithm 2: Weighted Probabilistic Community Detection

Step 1: Probabilistic weighted measure for community detection
Initialize an empty list "nodes_with_weight"
for each node in the social networking graph:
Compute the centralized mean weighted measure (lambda1) using the formula:
lambda1 = (M_A1 - M_A2) / (2 * sqrt(min(sigma_A1, sigma_A2)))
# Where M_A1 and M_A2 are calculated as follows:
M_A1 = CalculateAverageAttributeA1(node)
M_A2 = CalculateAverageAttributeA2(node)
Add (node, lambda1) to the "nodes_with_weight" list
Step 2: Maximizing the weighted probabilistic measure for community detection
Initialize an empty dictionary "community_membership"

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

for each community in the network:
Initialize an empty list "community_nodes"
for each node in the social networking graph:
Compute the probability of node's attribute A1 given community (Prob(A1 / Cm))
Compute the probability of node's attribute A2 given community (Prob(A2 / Cm))
Compute the maximized weighted probabilistic measure (lambda2) using the formula:
lambda2 = max(Prob(A1 / Cm), Prob(A2 / Cm)) / (2 * |N| * |Cm|)
# Where |N| is the number of nodes in the network, and |Cm| is the number of nodes in the community
Add (node, lambda2) to the "community_nodes" list
Sort the "community_nodes" list in descending order of lambda2
Select the top node as the community leader (or representative)
Add the community leader to the "community_membership" dictionary with the community as the value
Result: "community_membership" dictionary contains the mapping of nodes to their respective community leaders
Functions:
Function CalculateAverageAttributeA1(node):
Calculate the average of attribute A1 with respect to class samples for the given node
Return the computed average value
Function CalculateAverageAttributeA2(node):
Calculate the average of attribute A2 with respect to class samples for the given node
Return the computed average value

Input: Dataset D
Output: Privacy-enabled dataset D'

1. Read the input dataset D.
2. For each numerical attribute A in dataset D, apply algorithm 1 for data filtering.
3. If the dataset contains a sensitive attribute in the list SA:
For each attribute s in SA:
- Apply privacy-preserving data transformation on the attribute:
- Use non-linear data transformation functions with dynamic permutation matrices Q and R.
- Calculate R1, R2, and R3 based on SK (Secret Key) and Q.R.
- Compute H[i] for each byte in P[i] using bitwise XOR:
$H[i] = R1 \oplus R2 \oplus R3$
4. The privacy-enabled dataset D' is obtained by applying k-anonymization on D.

5.Iterate over each instance Oi within the K-nearest neighbor (KNN) objects KNN[]:
6. For every instance Oj in IPG[] where $i \neq j$:
Calculate the Chebyshev distance Nm^k among the KNN objects.
7. For each Chebyshev distance object within the local skyline modeling, identify the k nearest objects from the sorted list Nm^k[].
Employ local density estimation probability on the refined local skyline objects.
8. In the context of the MapReduce (MR) framework, for each reducer:
Determine the closest density objects using the proposed probabilistic KNN approach. - Compute Distc, mean^k, and mean^kappa as follows:
$Distc = mean^k + \kappa * \sqrt{((1 / (N-1)) * \sum(\varphi^k - mean^k)^2)}$
$\varphi^k = max\_j \{ KNN\_i(Dist\_\psi) \}$
$mean^k = (1 / N) * \sum(\varphi^k)$
9. Calculate prior estimation probability -κ.
10. Compute the proposed local density estimation (PLDE) for each instance:
$PLDE(v\_i) = (1 / \eta) * e^{(-(\|log(v\_p) - Dist\_c\|^2) / (2 * \sigma^2))} * \sum(T * e^{(-(v\_q^2) / (Dist\_e^2))})$
11. Filter all the k-nearest neighbor objects using local kernel density estimation.
12. Done

The above steps outlines an algorithm for "Privacy-Preserving Data Mining (PPDM) Community Classification" on a dataset D. The algorithm begins by reading the input dataset and, for each numerical attribute A, applies a data filtering algorithm (referred to as "algorithm 1"). If the dataset contains sensitive attributes in the list SA, privacy-preserving data transformations are applied to these attributes. These transformations involve non-linear functions with dynamic permutation matrices Q and R, resulting in a new representation H[i] for each byte in P[i].Subsequently, the algorithm anonymizes the dataset D using k-anonymization, ensuring privacy preservation. It then proceeds to perform K-nearest neighbor (KNN) analysis on the dataset, calculating Chebyshev distances between instances. The algorithm identifies the k-nearest neighbors for each instance and applies local density estimation to these neighbors.Within the MapReduce framework, the algorithm utilizes a probabilistic KNN method to identify nearest density objects. It calculates Distc, mean^k, and mean^kappa based on the kth nearest neighbor instances, involving statistical measures and probabilities. The algorithm further computes a local density estimation (PLDE) for each instance, considering various factors, such as distances and densities.Finally, the algorithm

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

filters the k-nearest neighbor objects using local kernel density estimation, ultimately producing a privacy-enabled dataset D'. This comprehensive approach ensures that community classification tasks can be performed while preserving privacy through data transformations and probabilistic methods.

EXPERIMENTAL RESULTS

In the Java programming environment, we conducted experiments utilizing third-party graph and similarity libraries to simulate experimental results. Our aim was to assess the performance of the proposed model, and we utilized four distinct datasets: Yelp, Football, Zachary, and Dolphin datasets. In this experimental study, we employed various metrics to evaluate the outcomes. These metrics include Q, NMI (Normalized Mutual Information), Variation of Information (VI), and Rand Index. These metrics were employed to assess the results obtained from the training datasets.

The Q metric is computed as follows:

$Q = \Sigma (e_{i \cdot i} - a_i^2)$

The NMI (Normalized Mutual Information) between P and Q is calculated as:

$NMI(P \mid Q) = (e(P) + e(Q) - e(P, Q)) / ((e(P) + e(Q)) / 2)$

Here, X represents the original values, and Y represents the predicted communities. e(P) and e(Q) denote the entropy values associated with their respective communities.

Sample Yelp dataset and its user's profile link

	A	B	C	D	E	F	G	H	I	J
1	business_id	date	review_id	stars	text	type	user_id	cool	useful	funny
2	9yKzy9PApeiPPOUJEtnvkg	26-01-2011	fWKvX83p0-ka4JS3dc6E5A	5	My wife	review	rLtl8ZkDX5vH5nAx9C3q5Q	2	5	0
3	ZRJwVLyzEJq1VAihDhYiow	27-07-2011	IjZ33sJrzXqU-0X6U8NwyA	5	I have no	review	0a2KyEL0d3Yb1V6aivbIuQ	0	0	0
4	6oRAC4uyJCsJl1X0WZpVSA	14-06-2012	IESLBzqUCLdSzSqm0eCSxQ	4	love the gy	review	0hT2KtfLiobPvh6cDC8JQg	0	1	0
5	_1QQZuf4zZOyFCvXc0o6Vg	27-05-2010	G-WvGaISbqqaMHlNnByodA	5	Rosie,	review	uZetl9T0NcROGOyFfughhg	1	2	0
6	6ozycU1RpktNG2-1BroVtw	05-01-2012	1uJFq2r5QfJG_6ExMRCaGw	5	General	review	vYmM4KTsC8ZfQBg-j5MWkw	0	0	0
7	#NAME?	13-12-2007	m2CKSsepBCoRYWxiRUsxAg	4	Quiessen	review	sqYN3lNgvPbPCTRsMFu27g	4	3	1
8	zp713qNhx8d9KCJJnrw1xA	12-02-2010	riFQ3vxNpP4rWLk_CSri2A	5	Drop	review	wFweIWhv2fREZV_dYkz_1g	7	7	4
9	hW0Ne_HTHEAgGF1rAdmR	12-07-2012	JL7GXJ9u4YMx7Rzs05NfiQ	4	Luckily, I	review	1ieuYcKS7zeAv_U15AB13A	0	1	0
10	wNUea3IXZWD63bbOQaOw	17-08-2012	XtnfnYmnJYi71yIuGsXIUA	4	Definitely	review	Vh_DlizgGhSqQh4qfZ2h6A	0	0	0
11	nMHhuYan8e3cONo3Porn	11-08-2010	jJAIXA46pU1swYyRCdfXtQ	5	Nobuo sho	review	sUNkXg8-KFtCMQDV6zRzQg	0	1	0
12	AsSCv0q_BWqIe3mX2JqsO	16-06-2010	E11jzpKz9Kw5K7fuARWfRw	5	The	review	#NAME?	1	3	1
13	e9nN4XxjdHj4qtKCOPq_vg	21-10-2011	3rPt0LxF7rgmEUrznoH22w	5	Wonderful	review	C1rHp3dmepNea7XiouwB6Q	1	1	0
14	h53YuCIIDfEFSJCQpk8v1g	11-01-2010	cGnKNX3I9rthE0-TH24-qA	5	They	review	UPtysDF6cUDUxq2KY-6Dcg	1	2	0
15	WGNIYMeXPyoWav1APUq	23-12-2011	FvEEw1_OsrYdvwLV5Hrliw	4	Good tatto	review	Xm8HXE1JHqscXe5BKf0GFQ	1	2	0
16	yc5AH9H71xJidA_J2mChLA	20-05-2010	pfUwBKYYmUXeiwrhDluQcw	4	I'm 2	review	JOG-4G4e8ae3lx_szHtR8g	1	1	0
17	Vb9FPCEL6Ly24PNxLBaAFw	20-03-2011	HvqmdqWcerVWO3Gs6zbrOw	2	Was it	review	ylWOj2y7TV2e3yYeWhu2QA	0	2	0
18	supigcPNO9IKo6olaTNV-g	12-10-2008	HXP_0UI-FCmA4f-k9CqvaQ	3	We went	review	SBbftLzfYYKItOMFwOTIJg	3	4	2

1	JkeCKyEaQlbLd9uZYl4DjA::LiiLii C.::http://www.yelp.com/user_details?userid=JkeCKyEaQlbLd9uZYl4DjA
2	cs91PAsv6esdWAaSkzm2lg::Jan Ellen T.::http://www.yelp.com/user_details?userid=cs91PAsv6esdWAaSkzm2lg
3	cMgGj2FXHEbzdNZdLN_EwA::Saki U.::http://www.yelp.com/user_details?userid=cMgGj2FXHEbzdNZdLN_EwA
4	KXJbnHT4PDS1JZNCFKdmMg::stephanie h.::http://www.yelp.com/user_details?userid=KXJbnHT4PDS1JZNCFKdmMg
5	Tpmvufw1eea1DrjLAY2jLg::Theodore J.::http://www.yelp.com/user_details?userid=Tpmvufw1eea1DrjLAY2jLg
6	L22W35q3Ci3TytpA2LW34g::Doug H.::http://www.yelp.com/user_details?userid=L22W35q3Ci3TytpA2LW34g
7	xiJ6e1qN3Sqv4dS4D8CpNg::Amelia M.::http://www.yelp.com/user_details?userid=xiJ6e1qN3Sqv4dS4D8CpNg
8	pu96s510jutWeFOuofgY2g::Ty G.::http://www.yelp.com/user_details?userid=pu96s510jutWeFOuofgY2g
9	zcOlcoYhVgEgWxRpTVjUJA::Steve K.::http://www.yelp.com/user_details?userid=zcOlcoYhVgEgWxRpTVjUJA
10	_NH7Cpq3qZkByP5xR4gXog::Chris M.::http://www.yelp.com/user_details?userid=_NH7Cpq3qZkByP5xR4gXog
11	9YLxlEqpEjiugcA1NCbQnw::Christina W.::http://www.yelp.com/user_details?userid=9YLxlEqpEjiugcA1NCbQnw
12	PiSHysV8QdhgzU7QIRn1Kg::Cheri W.::http://www.yelp.com/user_details?userid=PiSHysV8QdhgzU7QIRn1Kg
13	PB3OGUgRSajFx18EBUwwEQ::Lord H.::http://www.yelp.com/user_details?userid=PB3OGUgRSajFx18EBUwwEQ
14	e9ATa_PlOWiyTS4QhdsJKA::Derek F.::http://www.yelp.com/user_details?userid=e9ATa_PlOWiyTS4QhdsJKA
15	urQ0TUF3uhc-oOHzb3F5tQ::Yuan D.::http://www.yelp.com/user_details?userid=urQ0TUF3uhc-oOHzb3F5tQ
16	DUJLlGEHUerY8feKhjO85A::Ryan H.::http://www.yelp.com/user_details?userid=DUJLlGEHUerY8feKhjO85A
17	ti5uWqAxf7pOkKs7csMlDA::Tiffany R.::http://www.yelp.com/user_details?userid=ti5uWqAxf7pOkKs7csMlDA
18	JssCk4of-CQ7j81ZLrZMZg::Jennifer C.::http://www.yelp.com/user_details?userid=JssCk4of-CQ7j81ZLrZMZg
19	dMEMCWkbmi2hzr24_9J-ZA::Michael P.::http://www.yelp.com/user_details?userid=dMEMCWkbmi2hzr24_9J-ZA
20	Y_v3O9q_-vSOk25cOEiIkQ::Dan S.::http://www.yelp.com/user_details?userid=Y_v3O9q_-vSOk25cOEiIkQ

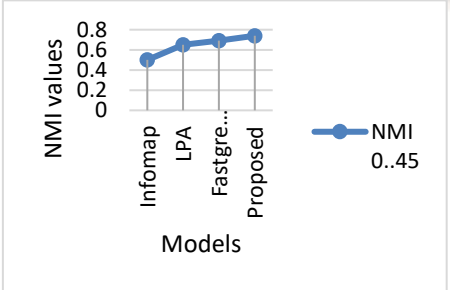

Figure 2 illustrates a comparative analysis of the proposed model against existing models using Yelp data.

Figure 2 illustrates a comparative analysis between our current probabilistic model and traditional models using the Yelp dataset. As depicted in Figure 2, our proposed NML demonstrates superior efficiency compared to conventional models when applied to the Yelp dataset. The NLM value in this context serves as an indicator of the quality of both inter and intra-community detection processes within the Yelp dataset.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

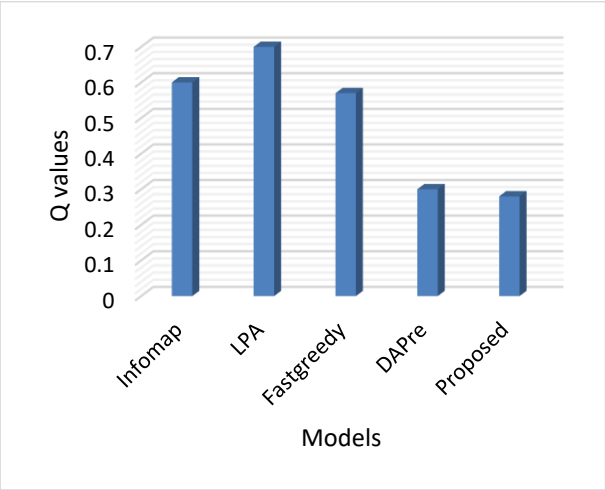

Figure 3 presents a comparative analysis of the Q value for our proposed model in contrast to existing models using dolphin data.

Figure 3, it offers a similar comparative analysis, but this time with the Dolphin dataset. Figure 3 reveals that our proposed Q value exhibits greater efficiency compared to traditional models when applied to the Dolphin dataset. The Q value in this case reflects the quality of the inter and intra-community detection process within the Dolphin dataset.

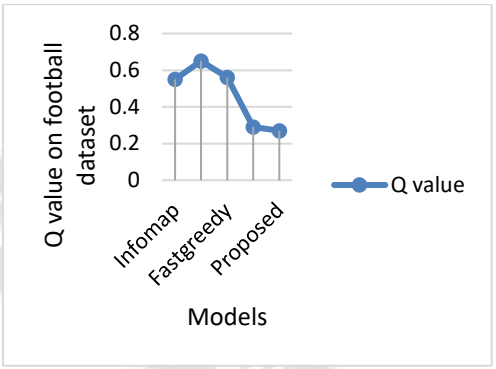

Figure 4 illustrates a comparative analysis of the Q value of our proposed model in contrast to existing models, using football data as the basis for evaluation.

Figure 4 is also used for a comparative analysis, this time involving the Football dataset. Once again, the proposed Q value outperforms traditional models when applied to the Football dataset, indicating better efficiency in the inter and intra-community detection process within this dataset.

Table 1 presents a comparison of the performance of our proposed runtime (measured in milliseconds) against conventional models across various Dolphin network data samples under the condition where the threshold (T) is set to 0.5.

TestData	DAPre	Infomap	LPA	Fastgreedy	Proposed
#1	5457.73	5602.01	5256.92	5905.25	4150.31
#2	5522.57	5808.21	5512.65	5031.87	3889.65
#3	5267.81	5034.47	5152.21	5671.64	4103.64
#4	4654.16	4590.52	5743.77	5467.65	4034.57
#5	4907.8	4697.92	5697.26	4590.71	4006.92
#6	5255.06	5783.48	4806.17	5753.15	4194.32
#7	5576.42	5187.76	5584.53	5680.31	3985.07
#8	5463.97	5509.66	5200.4	5330.78	4004.18
#9	5398.41	5008.92	4854.44	5109.32	3930.54
#10	4675.31	5708.5	5875.32	4570.44	4040.29
#11	5683.32	5489.67	5400.73	4707.45	3889.79
#12	5794.14	5892.72	5789.68	5277.21	3821.02
#13	5130.36	4955.01	5246.21	4741.8	4220.31
#14	5277.45	5652.51	4740.57	5285.34	4151.39
#15	5053.29	4992.08	5737.81	5882.44	4223.02

Table 2 illustrates the performance of the proposed model in terms of counting local patterns on various Yelp data samples, as compared to conventional models.

TestData	DAPre	Infomap	LPA	Fastgreedy	Proposed
#1	20334	11735	21072	13188	29256
#2	22143	15635	16544	10690	26578
#3	17977	11498	17887	14858	29769
#4	11267	12477	19834	13798	29614
#5	10706	19568	13386	14714	27749
#6	13320	15893	14324	17121	28893
#7	18316	18646	17336	10275	29380
#8	22271	21100	18217	18174	25863
#9	23570	15000	13250	13549	25061
#10	11644	12693	13353	11547	26848
#11	14015	17529	12325	20362	26773
#12	15167	15064	17419	11624	25519
#13	12622	18761	19716	18376	29503
#14	18197	10373	14146	23863	27306
#15	16177	12112	20599	10818	25653

Table 3 illustrates the performance of the proposed model in terms of counting local patterns compared to conventional models across various samples of football data.

TestData	DAPre	Infomap	LPA	Fastgreedy	Proposed
#1	5255.59	5552.48	5739.05	5796.98	3853.59
#2	5188.98	5922.72	5728.26	5332.06	3950.78
#3	5465.57	4743.73	5330.89	5787.95	4107.79
#4	5395.68	5317.58	5122.1	5122.45	3847.42
#5	4581.96	5501.57	5097.4	5707.32	4143.73
#6	4653.26	5111.33	5553.15	5275.48	3847.41
#7	4750.36	5291.44	4871.21	4970.34	3844.06
#8	5678.76	5518.11	5099.13	5728.01	3853.61
#9	5476.22	4617.5	5616.15	5351.27	4013.81
#10	5240.38	5743.09	5253.83	5426.25	4200.27
#11	5556.68	5043.22	4715.02	5163.9	3897.82
#12	5120.32	5378.32	4928.66	5709.21	3799.88
#13	4570.56	4564.81	4913.83	5330.53	3810.43
#14	4683.93	5631.18	4599.53	5581.28	4187.64
#15	5342.86	5076.49	4980.97	4822.49	4245.36

Additionally, we present a table comparing the runtime in milliseconds (ms) of our proposed model with conventional models. Across the board, it is evident from the table that our probabilistic privacy-preserving model outperforms conventional models in terms of runtime efficiency. This observation holds true for all three datasets: Yelp, Dolphin, and Football.

CONCLUSION

Traditional methods for protecting privacy in social networking datasets primarily rely on data perturbation techniques, as opposed to data transformation methods. This choice is driven by the considerable computational memory and time required when dealing with real-time social networking data. Additionally, conventional privacy-preserving approaches typically employ fixed metrics and static measures for community detection on social networking datasets.In our research, we introduce a probabilistic framework that employs a filtering approach to enhance the efficiency of community detection. Unlike traditional methods, our framework incorporates dynamic metrics, resulting in improved computational runtime, NMI, and Q rates compared to existing models, as confirmed by our experimental results.Looking ahead, our future work will involve the utilization of optimized meta-heuristic techniques for both local and global methods, with the aim of expanding the search space within the privacy-preserving process. This will further enhance the effectiveness of our approach, particularly when applied to datasets.

REFERENCES

[1]J. Zhou, Z. Cao, X. Dong, N. Xiong, and A. V. Vasilakos, “4S: A secure and privacy-preserving key management scheme for cloud-assisted wireless body area network in m-healthcare social networks,” Information Sciences, vol. 314, pp. 255–276, Sep. 2015, doi: 10.1016/j.ins.2014.09.003.

[2]E. Bandara et al., “A blockchain empowered and privacy preserving digital contact tracing platform,” Information Processing & Management, vol. 58, no. 4, p. 102572, Jul. 2021, doi: 10.1016/j.ipm.2021.102572.

[3]R. G. Pensa and G. Di Blasi, “A privacy self-assessment framework for online social networks,” Expert Systems with Applications, vol. 86, pp. 18–31, Nov. 2017, doi: 10.1016/j.eswa.2017.05.054.

[4]S. Nicolazzo, A. Nocera, D. Ursino, and L. Virgili, “A privacy-preserving approach to prevent feature disclosure in an IoT scenario,” Future Generation Computer Systems, vol. 105, pp. 502–519, Apr. 2020, doi: 10.1016/j.future.2019.12.017.

[5]S. Kavianpour, A. Tamimi, and B. Shanmugam, "A privacy-preserving model to control social interaction behaviors in social network sites," Journal of Information Security and Applications, vol. 49, p. 102402, Dec. 2019, doi: 10.1016/j.jisa.2019.102402.

[6]S. Beg et al., "A privacy-preserving protocol for continuous and dynamic data collection in IoT enabled mobile app recommendation system (MARS)," Journal of Network and Computer Applications, vol. 174, p. 102874, Jan. 2021, doi: 10.1016/j.jnca.2020.102874.

[7]S. Aghaalizadeh, S. T. Afshord, A. Bouyer, and B. Anari, "A three-stage algorithm for local community detection based on the high node importance ranking in social networks," Physica A: Statistical Mechanics and its Applications, vol. 563, p. 125420, Feb. 2021, doi: 10.1016/j.physa.2020.125420.

[8]A.-T. Tran, T.-D. Luong, J. Karnjana, and V.-N. Huynh, "An efficient approach for privacy preserving decentralized deep learning models based on secure multi-party computation," Neurocomputing, vol. 422, pp. 245–262, Jan. 2021, doi: 10.1016/j.neucom.2020.10.014.

[9]P. Barsocchi et al., "COVID-19 & privacy: Enhancing of indoor localization architectures towards effective social distancing," Array, vol. 9, p. 100051, Mar. 2021, doi: 10.1016/j.array.2020.100051.

[10]L. Bahri, B. Carminati, and E. Ferrari, "Decentralized privacy preserving services for Online Social Networks," Online Social Networks and Media, vol. 6, pp. 18–25, Jun. 2018, doi: 10.1016/j.osnem.2018.02.001.

[11]P. Liu et al., "Local differential privacy for social network publishing," Neurocomputing, vol. 391, pp. 273–279, May 2020, doi: 10.1016/j.neucom.2018.11.104.

[12]X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," Medical Image Analysis, vol. 65, p. 101765, Oct. 2020, doi: 10.1016/j.media.2020.101765.

[13]J. Yang, C. Fu, and H. Lu, "Optimized and federated soft-impute for privacy-preserving tensor completion in cyber-physical-social systems," Information Sciences, vol. 564, pp. 103–123, Jul. 2021, doi: 10.1016/j.ins.2021.02.028.

[14]I. Kayes and A. Iamnitchi, "Privacy and security in online social networks: A survey," Online Social Networks and Media, vol. 3–4, pp. 1–21, Oct. 2017, doi: 10.1016/j.osnem.2017.09.001.

[15]E. Ezhilarasan and M. Dinakaran, "Privacy preserving and data transpiration in multiple cloud using secure and robust data access management algorithm," Microprocessors and Microsystems, vol. 82, p. 103956, Apr. 2021, doi: 10.1016/j.micpro.2021.103956.

[16]"Privacy Preserving Technique in Data Mining by Using Chinease Remainder Theorem | SpringerLink." https://link.springer.com/chapter/10.1007/978-3-642-32112-2_50 (accessed May 07, 2021).

[17]D. Aruna Kumari and T. Gunasekhar, "A Reconstruction Algorithm using Binary Transform for Privacy-Preserving Data Mining," Indian Journal of Science and Technology, vol. 9, no. 17, May 2016, doi: 10.17485/ijst/2016/v9i17/93122.

[18]X. Zheng, Z. Cai, G. Luo, L. Tian, and X. Bai, "Privacy-preserved community discovery in online social networks," Future Generation Computer Systems, vol. 93, pp. 1002–1009, Apr. 2019, doi: 10.1016/j.future.2018.04.020.

[19]Y. Zhao, S. K. Tarus, L. T. Yang, J. Sun, Y. Ge, and J. Wang, "Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives," Information Sciences, vol. 515, pp. 132–155, Apr. 2020, doi: 10.1016/j.ins.2019.10.019.