2024 International Conference on Smart Systems for Electrical, Electronics, Communication and Computer Engineering (ICSSEECC) | 979-8-3503-7817-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSSEECC61126.2024.10649423

Privacy-Preserving Social Network Clustering Using Differential Privacy

K Sathiyapriya
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
spk.cse@psgtech.ac.in

R Kavin Aravindhan
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
kavin.aravindhan@gmail.com

Kireshvanth B
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
kiresh20122002@gmail.com

Yadav Ranganathan
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
yadavranganathan@gmail.com

Hardik P
Department of Computer Science and Engineering
PSG College of Technology
Coimbatore, India
hardikprabhu13@gmail.com

***Abstract*— In the contemporary landscape of online social networks, preserving users' privacy while applying clustering techniques is a pivotal concern. This paper explores the integration of differential privacy into social network clustering to strike a balance between privacy and clustering effectiveness. The principal objective is to safeguard users' personal information, and the study meticulously examines the trade-offs inherent in differential privacy parameters and their influence on clustering performance. This paper provides a comprehensive insight into the nuanced interplay between user privacy, data utility, and clustering effectiveness within the distinctive dynamics of social networks, offering a promising solution for the field.**

Keywords - Social Networks, Differential Privacy, Network Clustering, User Privacy, Privacy Parameters, Data Privacy.

I. Introduction

In the digital age, online social networks have transformed communication and information exchange, yet significant concerns about user privacy and data security persist. This paper addresses these issues by focusing on safeguarding user privacy during social network clustering. The methodology revolves around social network clustering, organizing users based on features to uncover patterns and correlations without predefined categories. Simultaneously, it prioritizes privacy preservation, integrating differential privacy into clustering to protect user data while maintaining the utility and effectiveness of the clustering process.

The primary focus includes introducing privacy mechanisms, evaluating their impact, and assessing trade-offs between privacy and utility. This approach not only aims to mitigate privacy risks but also promises more accurate and personalized recommendations, with potential applications in e-commerce and recommendation systems. The following sections delve into methodologies, results, and the broader impact of this work.

II. Literature Survey

In the era of ever-expanding social networks, the rapid generation of data presents unparalleled opportunities to delve into the intricacies of human interactions and behaviors. However, this data-driven era is accompanied by a paramount concern—privacy. This literature survey embarks on an exploration of privacy-preserving social network clustering.

Jain et al. [1] underscore the impact of big data on modern life across diverse spheres, emphasizing the need for privacy in extensive data production. Jiang et al.'s survey [2] delves into the convergence of differential privacy principles with social network analysis, recognizing its growing applications and privacy preservation in information sharing. Chen and Zhang [3] contribute to the technological intricacies of leveraging differential privacy for social network clustering, shedding light on the nuanced interplay between privacy preservation and meaningful clustering insights.

In the realm of privacy-preserving clustering, Wang and Li [4] introduce an ingenious approach based on differential privacy, striking a harmonious balance between accurate clustering results and user privacy. Rodriguez and K. Gupta [5] navigate the forefront of research, investigating the integration of differential privacy into social network clustering paradigms. Gupta et al. [6] investigate novel techniques to enhance privacy in social network clustering through the lens of differential privacy. Kim and Park's comprehensive survey [7] provides a panoramic view of privacy-preserving techniques, offering a contextual umbrella for the proposed project within the broader privacy landscape.

Davis et al. [8] emphasize differential privacy's significance in social network clustering, reviewing recent research and technological developments. Zheng and Xiong's [9] survey offers insights into balancing accuracy and privacy in differentially private clustering. Sharma et al. [10] focus on integrating differential privacy into clustering strategies, while Smith and Brown [11] highlight its role in maintaining user confidentiality in social network analysis. Patel and Gupta [12] provide a comprehensive review of recent advancements in differential privacy for social network clustering, discussing emerging trends and challenges.

Wang and Chen's survey [13] extends the scope beyond social networks, delving into privacy-preserving techniques in big data clustering. This broader perspective sets the stage for Liu and Yang's [14] innovative framework, proposing a novel approach to privacy-preserving social network clustering through the lens of differential privacy, effectively balancing accuracy and user privacy. Meanwhile, Zhang and Wu's [15] exploration focuses on the challenges and opportunities associated with integrating differential privacy into social network analytics, offering valuable insights for researchers and practitioners in the field.

In conclusion, the literature review highlights the importance of privacy in extensive data production, particularly in social network clustering. Innovative techniques such as Laplace noise with varied hyperparameters and Laplace equation for matrix transformation enhance privacy. However, balancing privacy with meaningful clustering insights poses a challenge, as higher privacy parameter values may compromise clustering utility. The use of selected features for K-Means clustering is also explored, offering insights into imparting sensitivity to these features.

III. Methodology

A. System Architecture

The proposed system integrates differential privacy with clustering to achieve robust privacy-preserving clustering in social networks. At its core is a modified version of K-Means clustering adapted for social network data, which replaces the original feature matrix with a noisy one and incorporates Laplacian-distributed noise within the differential privacy module. This controlled introduction of noise reduces re-identification risks and enhances the security of sensitive data, while still allowing for the extraction of valuable social insights from the clustered data.

However, adding Laplacian noise can impact clustering accuracy, potentially reducing the quality of insights. Balancing privacy and utility is challenging, as strong privacy measures can compromise clustering performance. The architecture of the proposed system, shown in Figure 1, illustrates the interplay between privacy and utility.

B. K-Means Clustering

K-Means clustering [16][17] is a tool for the analysis of social network data, aimed at unveiling underlying patterns by categorizing network users into K clusters, where K is either user-defined or determined through optimization, providing valuable insights into community structures, behaviors, and connections. The primary objective of K-Means is to minimize the within-cluster sum of squares, expressed as:

$$W(K) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} || x_j - c_i ||^2 \qquad (1)$$

Where:

- $W(K)$ denotes the within-cluster sum of squares for Kclusters.
- K signifies the number of clusters.
- n_i represents the number of data points in cluster i.
- x_j is an individual data point in cluster i.
- c_i stands for the centroid of cluster i.

The K-Means clustering architecture in social network analysis begins by importing the feature matrix and initializing cluster assignments based on the number of active features in each node. Cluster centroids are computed, and subsequent assignments are made accordingly. Convergence is checked, and if not reached, centroids are recalculated, followed by cluster reassignment.

C. Differential Privacy

Differential privacy protects sensitive data by adding noise, ensuring individual privacy in social network analysis. It proactively addresses privacy concerns through mathematical notation, with sensitivity as a key aspect. It is formally expressed as:

$$\Delta f = max(| f(D) - f(D') |) \qquad (2)$$

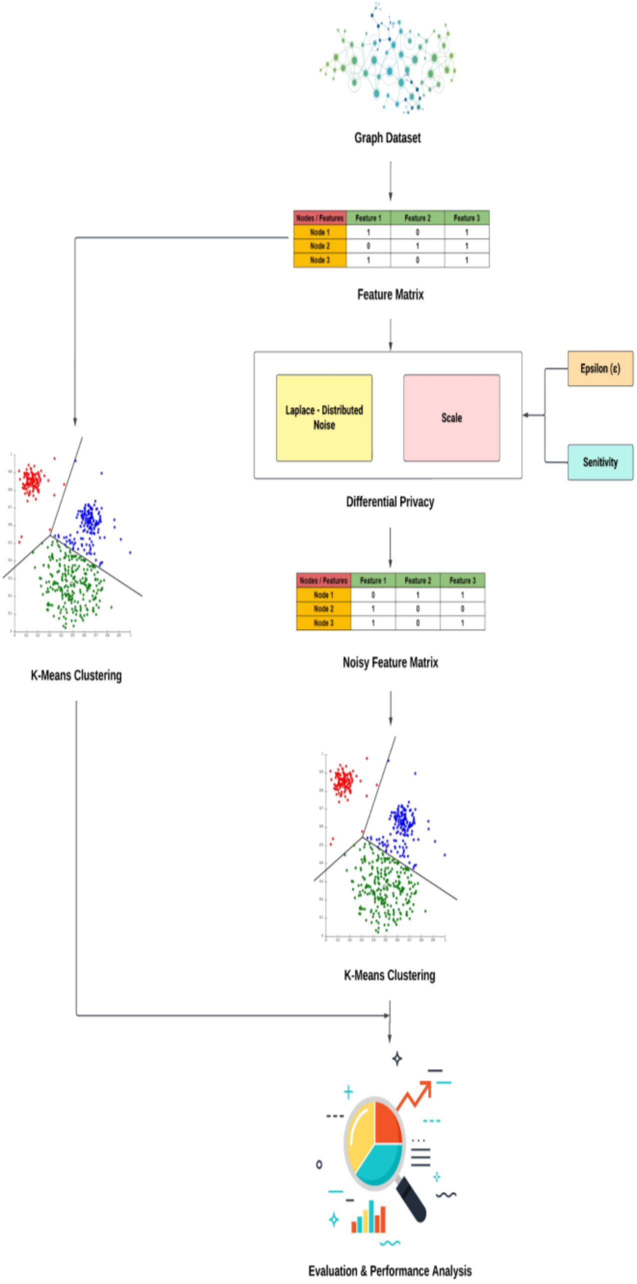
Here, Δf represents the sensitivity of function f, D represents the initial dataset, and D' signifies a neighboring dataset that differs by a single data point. Additionally, the Laplace mechanism [18] is instrumental in introducing noise to a function's output, thereby ensuring differential privacy.

The Laplace mechanism is mathematically represented as:

$$f(D) + Laplace(\frac{\Delta f}{\varepsilon}) \qquad (3)$$

Where $Laplace(\frac{\Delta f}{\varepsilon})$ represents the addition of Laplace noise to the output of the function Δf, with the scale parameter (b).

The paper proposes integrating Laplace noise into a feature matrix to protect individual data while preserving statistical properties, adhering to differential privacy principles. It introduces create_noise_matrix with parameters for the original matrix, sensitivity list, and epsilon, and an add_laplace_noise function to compute noise scale and enhance data privacy.



Nodes / Features	Feature 1	Feature 2	Feature 3
Node 1	1	0	1
Node 2	0	1	1
Node 3	1	0	1

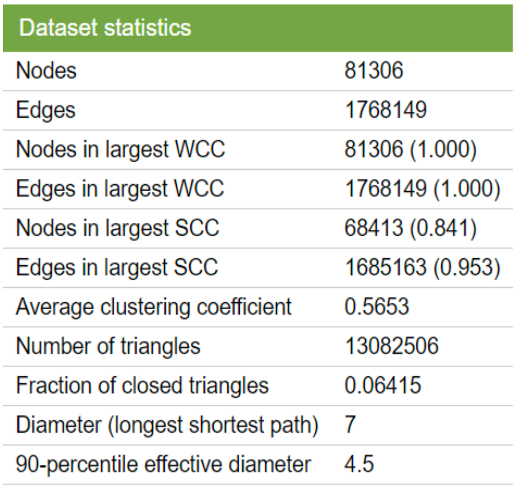
Nodes / Features	Feature 1	Feature 2	Feature 3
Node 1	0	1	1
Node 2	1	0	0
Node 3	1	0	1

Fig. 1 Overall Architecture of Privacy-Preserving Clustering

IV. Results

A. Dataset

The dataset comprises various files, encompassing individual Twitter user profiles, follower and friend lists, and a subgraph representing interactions within the Twitter network. Twitter, a prominent social media and microblogging platform, enables real-time communication and the sharing of concise messages known as "tweets."



Dataset statistics	
Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653
Number of triangles	13082506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

Fig. 2 Dataset Statistics

B. Adjusted Rand Index

The Adjusted Rand Index (ARI) [20][22] quantitatively assesses the similarity between two sets of clusters, considering both alignment and divergence. The formula for the ARI is expressed as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]} \tag{4}$$

Where $RI\ (Rand\ Index)$ measures the clustering consistency, $E[RI]$ denotes its expected value for random clustering & $max(RI)$ represents its upper bound.

C. Silhouette Score

The Silhouette Score [19] measures the separation and cohesion of clusters, providing insights into clustering effectiveness. The formula for the Silhouette Score for an individual data point, $S(i)$ is:

$$S(i) = \frac{b(i) - a(i)}{\max(\mathrm{a(i),b(i)})} \tag{5}$$

Where $a(i)$ is the average distance between a data point and others in the same cluster (cohesion), while $b(i)$ is the smallest average distance to points in a different cluster (separation). The overall Silhouette Score for the entire dataset, symbolized as □, is computed as the average of the Silhouette Scores for all data points, where N represents the total number of data points in the dataset:

$$S = \frac{1}{N} \sum_{i=1}^{N} S(i) \tag{6}$$

D. Misclustered Points

"Misclustered Points" in clustering analysis refer to data points inaccurately assigned to clusters that do not align with their inherent characteristics. These misclassifications highlight discrepancies in clustering algorithm performance or parameter settings.

E. Davies-Bouldin Index

The Davies-Bouldin Index [23] is a valuable metric for evaluating clustering quality, measuring both cluster compactness and separation. The formula for the Davies-Bouldin Index, denoted as DB, is expressed as:

$$DB = \frac{1}{N}\sum_{i=1}^{N} \max_{j \neq i}\left(\frac{S_i+S_j}{d(c_i,c_j)}\right) \qquad (7)$$

Where N is the total clusters, S_i & S_j are the average distances within clusters i and j to their centroids, assessing compactness and $d(c_i, c_j)$ is the separation between the centroids of clusters i and j.

F. Accuracy

The Accuracy is calculated as follows:

$$Accuracy = \frac{s-m}{s} \times 100\% \qquad (8)$$

Where s is the total cluster labels & m is the total misclustered labels.

G. K-Means Clustering Analysis

Comparing clustering results from both matrices reveals the impact of privacy-preserving techniques on the privacy-utility trade-off. Visualizing clusters (Figures 3 & 4) highlights differential privacy's influence on results [24].

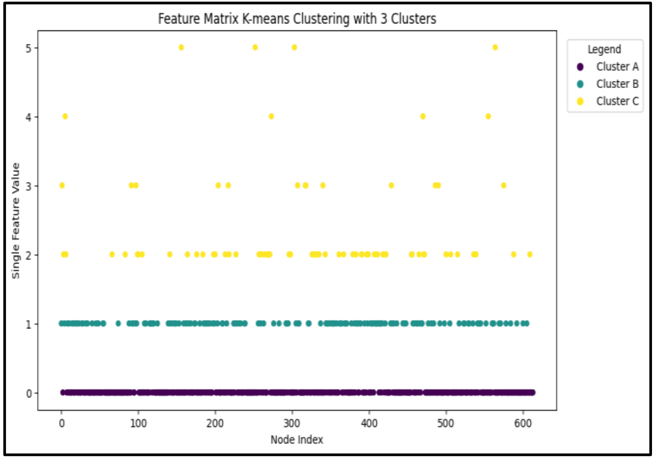


Fig. 3 K-Means Clustering of Feature Matrix

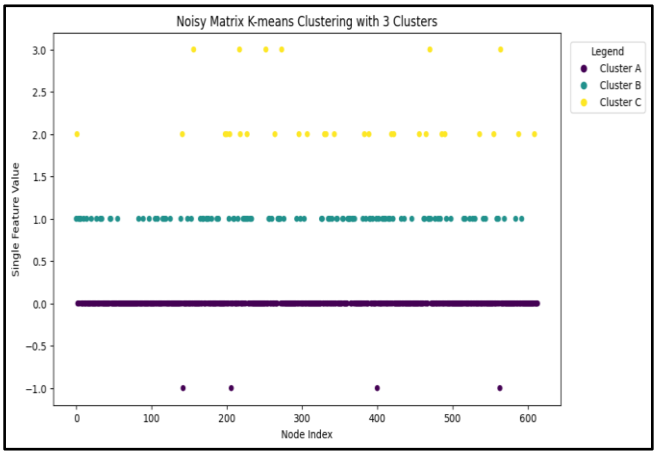


Fig. 4 K-Means Clustering of Noisy Matrix

H. Silhouette Score Comparison

Silhouette Score quantifies privacy-cluster integrity trade-offs for feature and noisy matrices, while Figure 5 visually compares clustering outcomes.

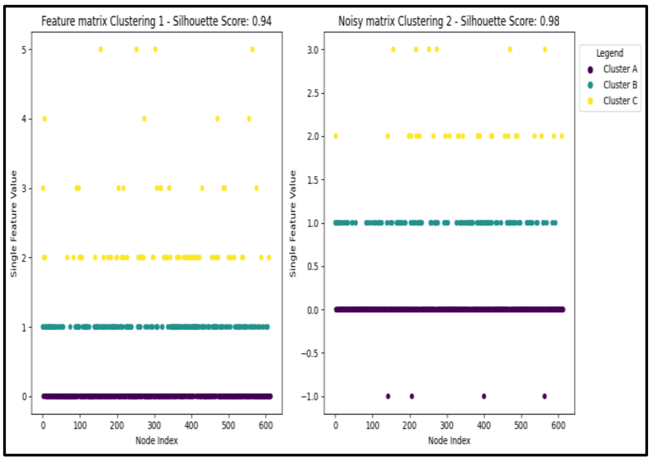


Fig. 5 Silhouette Score Comparison

I. Davies-Bouldin Score Comparison

The Davies-Bouldin Score assesses clustering quality for original and noisy matrices (Clustering 1 and Clustering 2), while Figure 6 facilitates intuitive comparison of clustering assignments and their impact on cluster quality.

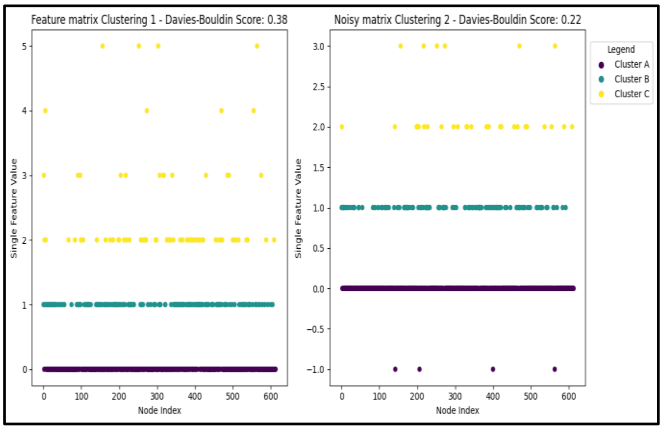


Fig. 6 Davies-Bouldin Score Comparison

J. Epsilon Tuning

Differential privacy in K-means clustering was evaluated through systematic parameter (ε) fine-tuning, assessing key parameters (ε, Adjusted Rand Index, Silhouette Score, Misclustered points, Accuracy), summarized in Tables 1, 2, and 3.

Table 1. Epsilon Tuning (ε = 0.1 to ε = 1)

Epsilon	Adjusted Rand Index	Silhouette Score	Misclustered Points	Davies-Boudlin Index	Accuracy
0.1	0.01	0.52	412	0.56	34.52%
0.2	0.01	0.53	402	0.56	34.52%
0.3	0.01	0.56	432	0.57	29.64%
0.4	0.01	0.56	432	0.55	29.64%
0.5	0.01	0.6	399	0.55	35.02%
0.6	0.03	0.59	478	0.53	22.15%
0.7	0.06	0.66	458	0.52	25.41%
0.8	0.06	0.67	355	0.57	42.18%
0.9	0.07	0.77	477	0.55	22.31%
1	0.01	0.85	435	0.39	29.15%

Table 2. Epsilon Tuning (ε = 1.1 to ε = 2)

Epsilon	Adjusted Rand Index	Silhouette Score	Misclustered Points	Davies-Boudlin Index	Accuracy
1.1	0.12	0.85	508	0.36	17.26%
1.2	0.16	0.87	426	0.34	30.62%
1.3	0.23	0.89	549	0.32	10.59%
1.4	0.36	0.88	227	0.28	63.03%
1.5	0.41	0.9	488	0.32	20.52%
1.6	0.48	0.92	131	0.25	78.66%
1.7	0.46	0.95	208	0.22	66.12%
1.8	0.47	0.96	496	0.26	77.26%
1.9	0.47	0.98	139	0.26	18.57%
2	0.49	0.98	500	0.17	78.50%

Table 3. Epsilon Tuning (ε = 2.1 to ε = 3)

Epsilon	Adjusted Rand Index	Silhouette Score	Misclustered Points	Davies-Boudlin Index	Accuracy
2.1	0.53	0.96	197	0.25	67.91%
2.2	0.54	0.98	118	0.31	80.53%
2.3	0.46	0.98	137	0.3	77.68%
2.4	0.46	0.98	135	0.27	78.01%
2.5	0.51	0.98	129	0.32	78.99%
2.6	0.52	0.99	135	0.2	78.01%
2.7	0.46	0.98	137	0.31	77.68%
2.8	0.48	0.98	191	0.36	68.89%
2.9	0.52	0.98	196	0.33	68.07%
3	0.52	0.99	504	0.24	17.91%

K. Epsilon vs Adjusted Rand Index

The analysis investigates Epsilon's impact (2.0 to 3.0) on privacy and clustering fidelity, assessing noise levels and privacy guarantees, with ARI measuring alignment and Figure 7 illustrating privacy-accuracy trade-offs.

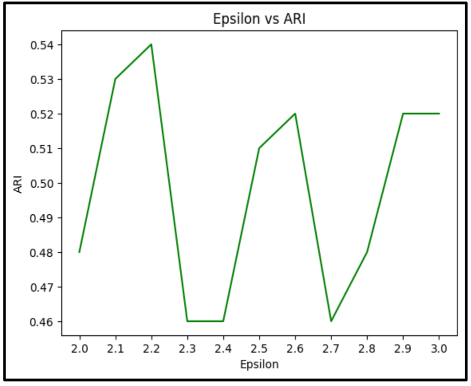


Fig. 7 Plot of Epsilon (ε) vs ARI

L. Epsilon vs Misclustered Points

Incrementally adjusting Epsilon (2.0 to 3.0) unveils trade-offs, notably in misclustered points, depicted in Figure 8, offering insights into privacy's impact on clustering accuracy.

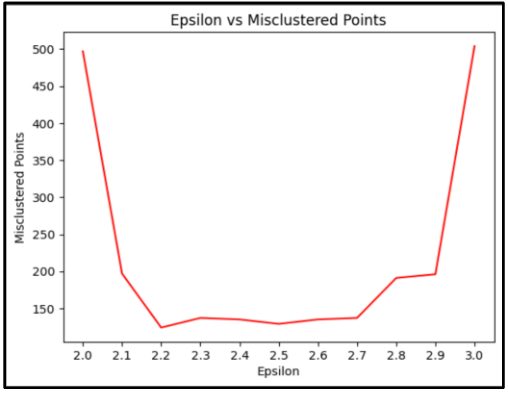


Fig. 8 Plot of Epsilon (ε) vs Misclustered Points

M. Epsilon vs Silhouette Score

Epsilon values (2.0 to 3.0) are adjusted to explore privacy-clustering trade-offs, with higher Epsilon potentially improving clustering alignment. Figure 9 visually depicts this relationship between Epsilon and Silhouette Scores, offering insights into privacy's impact on clustering quality.

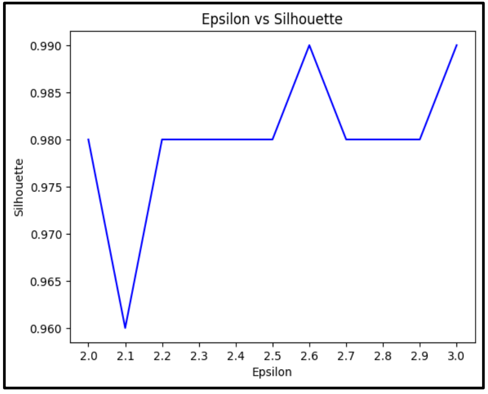


Fig. 9 Plot of Epsilon (ε) vs Silhouette Score

N. Epsilon vs Davies-Bouldin Index

The analysis explores privacy technique trade-offs via Epsilon variation (2.0 to 3.0), with Figure 10 illustrating its impact on the Davies-Bouldin Index, revealing differential privacy's influence on clustering quality in social networks.

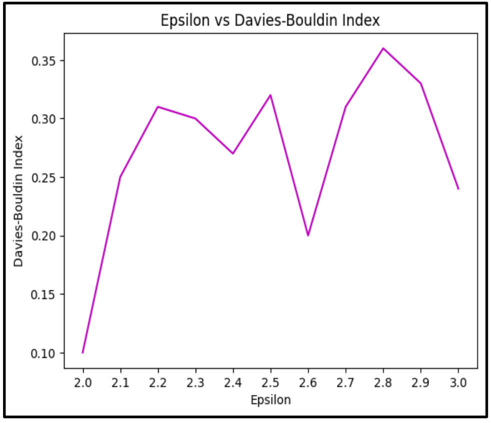


Fig. 10 Plot of Epsilon (ε) vs Davies-Bouldin Index

O. Epsilon vs Metrics Score

Systematic exploration of Epsilon (2.0 to 3.0) assesses privacy's impact on clustering accuracy via ARI, Silhouette Score, and Davies-Bouldin Index, depicted in Figure 11, aiding optimal Epsilon selection aligned with project objectives.

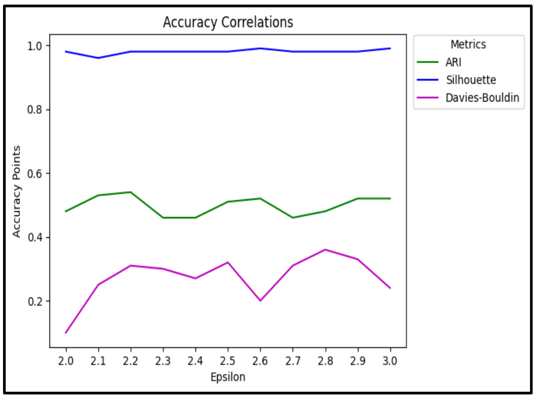


Fig. 11 Plot of Epsilon (ε) vs Metrics Score

V. CONCLUSION

In conclusion, this paper introduces a robust solution at the intersection of privacy preservation and social network analysis. The application of K-means clustering to the feature matrix, followed by its extension to a noisy feature matrix generated through the addition of Laplace noise using differential privacy methods, constitutes a fundamental aspect of our approach. The systematic variation of the privacy parameter, epsilon, ranging from 0.1 to 3 with increments of 0.1, provides insights into the dynamic relationship between privacy and evaluation metrics. Notably, at an epsilon value of 2.2, the clustering accuracy peaks at 80.53%, signifying a harmonious balance between privacy and clustering effectiveness. The visualization of the epsilon-evaluation metric relationship in the results section offers a comprehensive understanding of the trade-offs involved, enabling the manipulation of the privacy parameter to tailor results to specific needs.

The meticulous evaluation procedures conducted throughout this research affirm the precision of Laplace noise integration and the reliability of core functions, thereby upholding the integrity of privacy-preserving mechanisms. This work contributes to the evolving landscape of privacy in social network analysis, presenting a nuanced perspective on the delicate interplay between privacy considerations and clustering outcomes. The findings presented pave the way for future research focused on optimizing the privacy-utility trade-off within the domain of social network analysis.

VI. FUTURE ENHANCEMENTS

The successful integration of privacy-preserving methodologies with social network clustering addresses immediate concerns and lays the foundation for secure social network analysis. Future work extends to larger datasets across diverse domains, with a focus on enhancing sensitivity matrices and exploring advanced privacy techniques beyond differential privacy. This pursuit aims to safeguard user privacy while extracting valuable insights, harmonizing privacy, data analysis, and machine learning in social networks.

REFERENCES

[1] P. Jain, M. Gyanchandani, and N. Khare, "Differential Privacy: Its Technological Prescriptive Using Big Data", Journal of Big Data, vol. 5, no. 15, 2018.

[2] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of Differential Privacy in Social Network Analysis: A Survey".

[3] X. Chen, Y. Liu, and Z. Zhang, "Advancements in Differential Privacy Techniques for Social Network Clustering: A Literature Survey".

[4] Y. Wang and X. Li, "Privacy-Preserving Clustering with Differential Privacy", ACM Transactions on Knowledge Discovery from Data, vol. 12, no. 1, pp. 7:1-7:25, Jan. 2018.

[5] M. Rodriguez, K. Gupta, and S. Patel, "Exploring the Frontiers: A Survey on Privacy-Driven Social Network Clustering Using Differential Privacy".

[6] R. Gupta, S. Patel, and M. Lee, "Enhancing Privacy in Social Network Clustering through Novel Differential Privacy Techniques", Journal of Privacy Research, vol. 9, no. 1, pp. 45-68.

[7] J. Kim and H. Park, "Privacy-Preserving Data Mining: A Survey on Recent Advances", IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1067-1080, Jun. 2020.

[8] E. Davis, J. Yang, and K. Smith, "Privacy-Driven Social Network Clustering: A Contemporary Survey on Differential Privacy Approaches".

[9] S. Zheng and L. Xiong, "Differentially Private Clustering: A survey", IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 11, pp. 2085-2098, Nov. 2019.

[10] P. Sharma, Q. Li, and S. Chen, "Differential Privacy in the Social Network Sphere: An In-Depth Literature Survey on Clustering Strategies", Journal of Privacy Research, vol. 7, no. 2, pp. 123-145, 2010.

[11] J. Smith and A. Brown, "Enhancing Privacy in Social Network Analysis: A Differential Privacy Perspective," IEEE Transactions on Knowledge and Data Engineering, 2020.

[12] R. Patel and S. Gupta, "Advancements in Differential Privacy for Social Network Clustering: A Comprehensive Review," IEEE Transactions on Emerging Topics in Computing, 2021.

[13] Q. Wang and L. Chen, "Privacy-Preserving Techniques in Big Data Clustering: A Survey," IEEE Transactions on Big Data, 2019.

[14] M. Liu and H. Yang, "A Novel Framework for Privacy-Preserving Social Network Clustering Using Differential Privacy," IEEE Transactions on Information Forensics and Security, 2022.

[15] Y. Zhang and X. Wu, "Differential Privacy in Social Network Analytics: Challenges and Opportunities," IEEE Transactions on Network Science and Engineering, 2023.

[16] S. Na, L. Xumin, and G. Yong, "Research on K-means Clustering Algorithm: An Improved K-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

[17] Y. Li and H. Wu, "A Clustering Method Based on K-Means Algorithm," Physics Procedia, vol. 25, pp. 1104-1109, 2012, doi: 10.1016/j.phpro.2012.03.206.

[18] H. Jiang, J. Pei, D. Yu, J. Yu, B. Gong, and X. Cheng, "Applications of Differential Privacy in Social Network Analysis: A Survey," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 1, pp. TBD, January 2023.

[19] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 2020, pp. 747-748, doi: 10.1109/DSAA49011.2020.00096.

[20] Warrens, M.J., van der Hoef, H. Understanding the Adjusted Rand Index and Other Partition Comparison Indices Based on Counting Object Pairs. J Classif 39, 487–509 (2022).

[21] Fowlkes, E., Mallows, C.: A method for comparing two hierarchical clusterings. Journal of the American Statistical Association 78, 553–569 (1983).

[22] J. M. Santos and M. Embrechts, "On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification," in Artificial Neural Networks – ICANN 2009, 2009, vol. 5769, pp. TBD, ISBN: 978-3-642-04276-8.

[23] J. C. Rojas Thomas, M. S. Peñas and M. Mora, "New Version of Davies-Bouldin Index for Clustering Validation Based on Cylindrical Distance," 2013 32nd International Conference of the Chilean Computer Science Society (SCCC), Temuco, Chile, 2013, pp. 49-53, doi: 10.1109/SCCC.2013.29.

[24] Sieranoja, S., Fränti, P. Adapting k-means for graph clustering. Knowl Inf Syst 64, 115–142 (2022).

[25] https://snap.stanford.edu/data/ego-Twitter.html