Evaluating LLM-based Pseudonymization for Cross-Platform Identifiers

Yeong Su Lee, Hendrik Bothe and Michaela Geierhos
Research Institute CODE, University of the Bundeswehr Munich,
Neubiberg, Germany
e-mail: {yeongsu.lee | hendrik.bothe | michaela.geierhos}@unibw.de

***Abstract*—This paper addresses the impact of Large Language Model (LLM)-based pseudonymization on cross-platform identifier similarity and re-identification risk. Cross-platform profile linking enables organizations to analyze user behavior across multiple social networking platforms. However, this practice raises significant privacy and security concerns. Identifiers such as usernames and real names can be used to reconstruct user identities when aggregated across different online platforms. Therefore, addressing these challenges is highly relevant to current research on privacy-preserving data analytics, digital identity protection, and secure information management. Using Linktree profile connection datasets from Facebook, Instagram, LinkedIn, and Twitter/X, we analyze how pseudonymization affects identifier similarity and the potential recovery of original identifiers. Our findings demonstrate that pseudonymization significantly reduces identifier similarity, resulting in low re-identification success rates. The results highlight the potential of LLM-driven privacy-enhancing techniques for mitigating identity disclosure risks while preserving limited analytical utility. Consequently, this work contributes to ongoing research in privacy-aware data processing, secure cross-platform analytics, and trustworthy LLMs for online social networks.**

Keywords-pseudonymization; privacy-preserving data transformation; privacy-utility trade-off.

I. Introduction

Social networking platforms host vast amounts of user-generated data and have become important sources for digital analytics and sociotechnical research. Since individuals often have accounts on multiple platforms [1], researchers and organizations attempt to link user identities across different Social Networking Services (SNS) to better understand user behavior and digital presence [2]–[4]. However, cross-platform profile linking raises significant privacy concerns because identifiers, such as names and usernames, can be used to reconstruct user identities across platforms [5][6]. Recent research highlights how seemingly fragmented pieces of personal data can be combined to reconstruct digital identities. Studies on digital identity reconstruction demonstrate that data fragments from different platforms may collectively reveal sensitive personal information when linked together [5][7]. Similarly, prior work has shown that landing pages, such as Linktree, often act as central hubs that link multiple social media accounts. This facilitates cross-platform profile discovery and matching [8]. While these linkages enhance analytical capabilities, they also amplify privacy risks [9][10]. Consequently, there is a growing need for privacy-preserving mechanisms that enable social media analytics while protecting personal identifiers.

Previous studies have examined cross-platform profile linking [3][11] and the privacy risks associated with aggregating social media data [9][10]. However, existing studies have primarily focused either on improving matching accuracy or on anonymization techniques that remove identifiers entirely [4][12]–[14]. Traditional privacy protection approaches include pseudonymization techniques that aim to reduce the exposure of personal identifiers while maintaining data usability for analytical purposes [15]–[17]. Unlike anonymization, pseudonymization transforms identifiers while preserving the structural properties necessary for analytical tasks [18][19]. However, traditional pseudonymization methods, such as rule-based transformations or manual approaches, fail to optimize the trade-off between privacy and data utility, often damaging semantic and pragmatic meaning of the original texts [18]. Recent advances in LLMs introduce new opportunities for semantic pseudonymization, which may preserve the structural properties of identifiers while reducing their direct identifiability [19]–[21]. Researchers have explored mixed-methods approaches to pseudonymizing social media profile data, highlighting both the opportunities and challenges of automated pseudonymization techniques [19]. However, little systematic research has evaluated the impact of LLM-based pseudonymization on cross-platform identity matching tasks or analyzed the resulting trade-off between privacy protection and analytical utility. To address this gap, this study explores the following research question: How does LLM-based pseudonymization affect the privacy-utility trade-off in cross-platform social media profile matching? To answer this question, we use the datasets of linked social media profiles introduced in [5][7][8]. These datasets exploit publicly accessible Linktree landing pages that connect user accounts across multiple platforms. We extract person names and usernames from the linked profiles and evaluate their similarity across platforms before and after LLM-based pseudonymization.

This paper makes three main contributions. First, it empirically evaluates the use of LLM-based pseudonymization to transform social media identifiers in cross-platform analysis scenarios. Second, it analyzes the effect of pseudonymization on identifier similarity across platforms using real-world Linktree-based profile connection datasets. Third, it assesses the privacy implications of pseudonymization by performing simulated re-identification attacks and employing ranking-based metrics to quantify the privacy-utility trade-off.

The remainder of this paper is structured as follows: Section II reviews related work on cross-platform identity linkage and privacy-preserving identifier transformation. Section III describes the research design and experimental setup, followed by the evaluation results in Section IV. Finally, Section V

concludes the paper and outlines directions for future work.

II. Related work

Cross-platform profile linking, also known as user identity linkage, is the process of identifying accounts belonging to the same individual across multiple social networking platforms. Existing approaches can be broadly categorized into three groups: attribute-based, graph-based, and anchor-based methods [4][8][12][22]. Attribute-based and graph-based methods operate directly on data from social networking services. Attribute-based approaches compare profile attributes, such as usernames or display names, across platforms. Graph-based methods exploit structural information, such as friendship relations or interaction networks. In contrast, anchor-based approaches rely on external references that explicitly link accounts across platforms. Landing pages that aggregate links to multiple social media profiles, such as Linktree, serve as hubs that connect the accounts of the same individual. These pages also provide valuable ground-truth datasets for cross-platform profile matching. However, cross-platform profile linking also raises significant privacy concerns, as discussed extensively in [5]. Consequently, privacy-preserving data processing has become an important research topic in data services and applications. Privacy-preserving data processing commonly relies on two main approaches: anonymization and pseudonymization. Anonymization aims to remove or generalize identifying information so that individuals cannot be re-identified [23]. Typical anonymization techniques include data suppression, generalization, or aggregation [12]. Although anonymization can provide strong privacy protection, it often removes information necessary for analytical tasks.

In contrast, pseudonymization replaces identifying attributes with alternative identifiers while preserving the data's structural and analytical properties [16][17]. Since pseudonymized data can support analytical processing, this approach is often used in systems requiring both privacy protection and data usability. Traditional pseudonymization methods include rule-based transformations, hashing techniques, and cryptographic transformations that replace identifiers with encoded representations [16][17]. While these methods can effectively hide original identifiers, they often disrupt the semantic relationships necessary for similarity-based analysis [18]. Recent advances in LLMs provide new opportunities for semantic pseudonymization of textual identifiers. LLM-based methods can generate alternative identifiers that preserve structural characteristics, such as length and linguistic patterns, while reducing direct identifiability [19]–[21][24]–[26]. However, empirical studies evaluating how such transformations affect analytical tasks such as cross-platform profile matching remain limited. In the context of social networking platforms, direct identifiers such as display names and usernames, represent key attributes for cross-platform identity linkage. Therefore, protecting these identifiers while preserving their analytical usefulness remains a central challenge in privacy-preserving social media analytics. Recent work has explored pseudonymization approaches that transform such identifiers while retaining structural properties necessary for subsequent analysis, e.g., by combining LLM-based pseudonymization with lightweight encryption [19].

III. Research Design and Experimental Setup

This study relies on publicly available datasets that were introduced in prior research on cross-platform profile linking and privacy risks in social media environments. Specifically, we use datasets derived from Linktree landing pages, as well as datasets employed in studies on digital identity reconstruction [5][7][8]. These datasets are available upon request. Linktree landing pages often serve as hubs that connect various social media accounts belonging to the same individual. Thus, these pages provide explicit cross-platform references between user profiles and can be used as the ground truth for linking identities across platforms. The overall processing pipeline is shown in Figure 1.

A. Dataset

The dataset is derived from landing pages and contains links to user profiles on various social networking platforms, such as Facebook, Instagram, LinkedIn, and Twitter/X. Two identifiers were extracted from these linked profiles in these datasets for further analysis: the person name (display name) and the username (platform handle). These identifiers are commonly used in cross-platform profile matching because they are publicly visible and often contain information that links user accounts across platforms [11]. Prior to analysis, the identifiers were normalized by converting them to lowercase and removing platform-specific formatting elements to ensure consistent similarity comparisons across platforms. Table I (a) summarizes the number of collected profiles per platform, and Table I (b) reports the overlap of profiles across platform combinations based on shared Linktree identifiers.

TABLE I. Dataset statistics.

(a) Profiles per platform

Platform	Profiles
Facebook (FB)	4,907
Twitter/X (TW)	4,953
Instagram (IG)	4,566
LinkedIn (IN)	3,735

(b) Cross-platform profiles

Platform Combination	Shared Profiles
FB, IG, IN, TW	2,797
FB, IG, TW	3,530
FB, IN, TW	3,096
FB, IG, IN	2,874
IG, IN, TW	3,145
FB–IG	3,648
FB–IN	3,205
FB–TW	3,992
IG–IN	3,236
IG–TW	4,006
IN–TW	3,552

B. LLM-based Pseudonymization

To protect personal identifiers while maintaining analytical usability, the extracted identifiers were pseudonymized using LLMs. Specifically, the Llama language model [27] was used

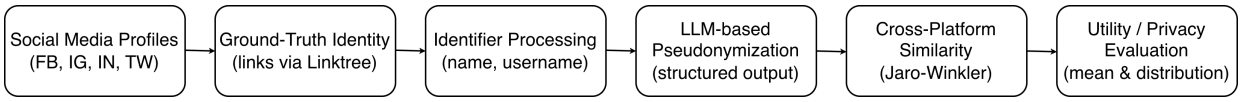

Figure 1. Overview of the proposed LLM-based pseudonymization and evaluation process. Social media profiles from Facebook, Instagram, LinkedIn, and Twitter/X are linked through Linktree-based ground-truth identities. User identifiers (names and usernames) are extracted and processed using an LLM-based pseudonymization approach with structured outputs. The resulting pseudonyms are then evaluated using cross-platform similarity analysis based on the Jaro-Winkler metric, followed by a utility and privacy assessment using similarity distributions and mean scores.

to generate alternative identifiers for both person names and usernames. The model was prompted to generate pseudonyms that removed direct personal references while maintaining the structural characteristics of the original identifiers. This approach reduces direct identifiability while preserving partial structural similarity, which may support analytical tasks such as cross-platform similarity analysis.

For prompting, Llama 3.3 with 70 billion parameters is employed via the Ollama Python library [28]. Table II shows an example of a query sent to the local model to process a dictionary-based input, along with the generated response. The `system` role specifies the model's behavior and provides instructions that guide the generation of responses to user requests.

TABLE II. AN EXAMPLE OF A PROMPT REQUEST AND RESPONSE.

Request	
role:	system
content:	Your task is to replace the name and username with a pseudonym in the given dict data. 1. Get the dict data! 2. Find out the items dict_data['name'] and dict_data['username'] from dict data and replace them with a pseudonym. 3. Keep the language and format of name and usernames for the pseudonyms! 4. Do not use the same pseudonym for different names and usernames! 5. Give the result back as dict format! 6. Do not include any explanation, example, transformation, or comments! Are you ready?
messages:	{'role': 'user', 'content': {'name': 'Renata Schmitt', 'username': 'renata.schmitt'}} {'role': 'assistant', 'content': {'name': 'Laris Lavati', 'username': 'laris.lavati'}}
parameter:	{temperature: 0.0, top_p: 0.9}
role:	user
content:	{"name": "Vlad Plavlkov", "username": "vld-plalkv"}
Response	
role:	assistant
content:	{"name": "Bryan Petrov", "username": "bryn-petrov"}

C. Evaluation Strategy

Before analyzing the similarity distributions, the consistency of the LLM-based pseudonymization process was verified. We implemented pseudonymization using structured, dictionary-based inputs and outputs. This ensured that the generated pseudonyms would follow a predefined format and that both identifiers (person name and username) would be produced consistently. This approach guaranteed that all identifiers were successfully transformed without formatting errors or incomplete outputs. Consequently, no entries were removed during preprocessing, and the number of identifiers in the pseudonymized dataset is the same as in the original dataset.

To evaluate the impact of pseudonymization, the Jaro-Winkler similarity metric was computed across platform pairs to measure identifier similarity [29]. We analyzed similarity distributions for both the original and the pseudonymized identifiers to assess the transformation's effect on cross-platform similarity patterns. The Jaro-Winkler metric was chosen because it effectively measures short textual identifiers, such as personal names. It emphasizes prefix similarity and tolerates minor spelling variations [30][31]. We computed the average similarity between identifiers across platform pairs to quantify the overall reduction in identifier similarity caused by pseudonymization. To evaluate the remaining privacy risk, a simulated re-identification attack was performed, comparing the pseudonymized identifiers against candidate pools of original identifiers and ranking them according to the Jaro-Winkler metric. These analyses allow us to evaluate the privacy-utility trade-off between reduced identifier similarity and the risk of identifier recovery.

These experiments compare the similarity distributions and average similarity scores of original and pseudonymized identifiers to quantify the effect of pseudonymization on cross-platform identifier similarity. Additionally, a re-identification experiment is conducted to evaluate the privacy risks associated with pseudonymized identifiers. In this experiment, the identifiers are compared against candidate pools of original identifiers from the same platform, and the candidates are ranked according to their similarity scores. The effectiveness of this attack is evaluated using ranking-based metrics, such as Mean Reciprocal Rank (MRR), Mean Rank (MeanR) and Median Rank (MedR), and top-k accuracy.

IV. RESULTS AND EVALUATION

This section presents the results and evaluation metrics, as described in Section III.

A. Average Similarity Scores

Table III shows the average Jaro-Winkler similarity between identifiers across platform pairs for both the original and the pseudonymized data. Similarity values for original usernames

range from 0.747 to 0.845, indicating the frequent reuse of identical or highly similar usernames across platforms. The highest similarity is observed between Instagram and Twitter (0.845). After pseudonymization, however, similarity decreases across all platform pairs. The average similarity for usernames drops by roughly 0.15–0.20, depending on the platform combination. A smaller reduction is observed for person names, where the average similarity decreases from 0.79–0.81 to 0.69–0.71. Overall, LLM-based pseudonymization weakens direct identifier similarity while preserving partial structural characteristics of the original identifiers. This reflects a privacy-utility trade-off: reduced similarity improves privacy protection, yet still allows for limited, similarity-based analysis.

TABLE III. AVERAGE JARO-WINKLER SIMILARITY BETWEEN PLATFORM PAIRS. UN:USERNAME, O:ORIGINAL, AND P:PSEUDONYM.

Identifier	FB-TW	FB-IG	FB-IN	IG-IN	IG-TW	TW-IN
UN (O)	0.792	0.822	0.747	0.763	0.845	0.751
UN (P)	0.611	0.633	0.595	0.616	0.645	0.597
Name (O)	0.811	0.805	0.785	0.812	0.809	0.797
Name (P)	0.693	0.706	0.691	0.711	0.702	0.687

B. Similarity Distribution

Figures 2 and 3 show the distribution of Jaro-Winkler similarity scores across different platform pairs, both using original and pseudonymized identifiers. A large proportion of username pairs exhibit very high similarity values in the original data. In particular, most username pairs fall into the highest similarity interval ($\geq$ 0.9). For instance, 61.8 % of username pairs on Instagram and Twitter, as well as 56.9 % on Facebook and Instagram, fall into this interval. These results confirm that many users reuse identical or highly similar usernames across platforms. However, compared to the original data, the proportion of very high similarity values ($\geq$ 0.9) decreases substantially across all platform pairs for the pseudonymized data. Instead, most pairs shift toward lower similarity intervals, particularly the $<$0.5 and 0.5–0.69 ranges. For instance, for the Facebook-Twitter pair the proportion of similarity values below 0.7 increases to more than 70 %. This indicates that LLM-based pseudonymization effectively reduces direct structural similarity between usernames.

Similar to usernames, original person names exhibit a strong concentration in the highest similarity interval ($\geq$ 0.9). Across all platform pairs, approximately half of the name pairs fall into this range. After applying LLM-based pseudonymization, the similarity distribution shifts toward lower similarity intervals. The proportion of highly similar names decreases significantly, while the proportion of medium similarity values (0.5–0.69) increases across most platform pairs. This indicates that the pseudonymization process removes direct name correspondences while preserving partial structural similarity.

These results indicate a clear trade-off between privacy and utility introduced by LLM-based pseudonymization. While original identifiers provide strong signals for cross-platform matching, pseudonymization substantially reduces the similarity of identifiers across platforms. For usernames, the proportion of highly similar identifiers (similarity $\geq$ 0.9) drops from about 47 % to 17 %, and the matching utility (recall with a threshold of 0.7) decreases from approximately 64 % to 24 %. A similar, albeit less pronounced, effect is observed for person names: highly similar identifiers decrease from approximately 51 % to 27 %, and recall decreases from about 68 % to 40 %. Overall, the results show that pseudonymization weakens direct identifier correspondence across platforms while preserving limited structural characteristics. While this reduction lowers matching performance, it significantly decreases the risk of direct identity reconstruction across social networking services.

C. Re-Identification Attack

To evaluate the privacy risks associated with pseudonymized identifiers, we conducted a re-identification (re-ID) attack simulation. In this scenario, an attacker observes a pseudonymized identifier and attempts to find its corresponding original identifier by searching a pool of identifiers from the same platform. Similarity scores are computed for each pseudonymized identifier (p_i) against all original identifiers (o_j) using the Jaro-Winkler metric. The original identifiers are then ranked according to their similarity scores. The position of the correct identifier in this list determines the success of the attack. Table IV summarizes the results of the simulated re-identification attack.

We evaluate the attack using ranking-based metrics commonly used in information retrieval [32]–[34]. The MRR metric measures how early the correct identifier appears in the ranked candidate list. Additionally, MeanR and MedR describe the average and median rank positions of the correct identifier, respectively. Top-k accuracy measures whether the correct identifier appears among the top k candidates ($k \in \{1, 3, 5, 10\}$). The candidate pool size differs from the total number of profiles because some Linktree landing pages reference multiple accounts from the same platform. In such cases, several social media profiles share the same Linktree identifier. These profiles are consolidated when constructing the candidate pool in order to avoid duplicate ground-truth mappings.

Overall, success rates remain relatively low across all platforms. For usernames, top-1 accuracy ranges from about 1.4 % to 2.6 %, indicating that directly reconstructing original usernames from pseudonymized identifiers is difficult. Person names have slightly higher re-identification rates, reaching 5.8 % on Twitter. Ranking-based metrics further confirm this observation. The relatively high mean and median ranks suggest that correct identifiers usually appear toward the end of the ranked candidate list. Although top-k accuracy increases with larger k, the overall success rates are low compared to the size of the candidate pools. These results suggest that LLM-based pseudonymization substantially reduces the risk of identifier reconstruction, though it does not eliminate the possibility of similarity-based re-identification.

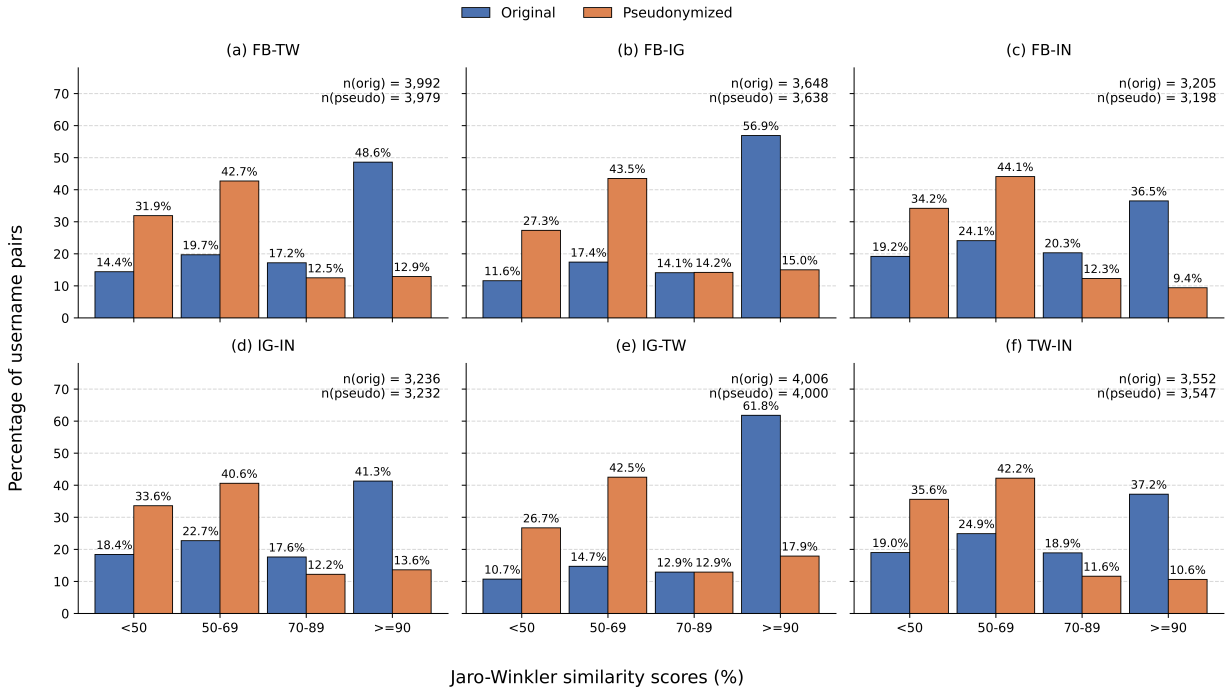

Figure 2. Similarity distribution of usernames. The slight reduction of numbers of pseudonymized usernames indicates that some pseudonymized usernames are duplicated.

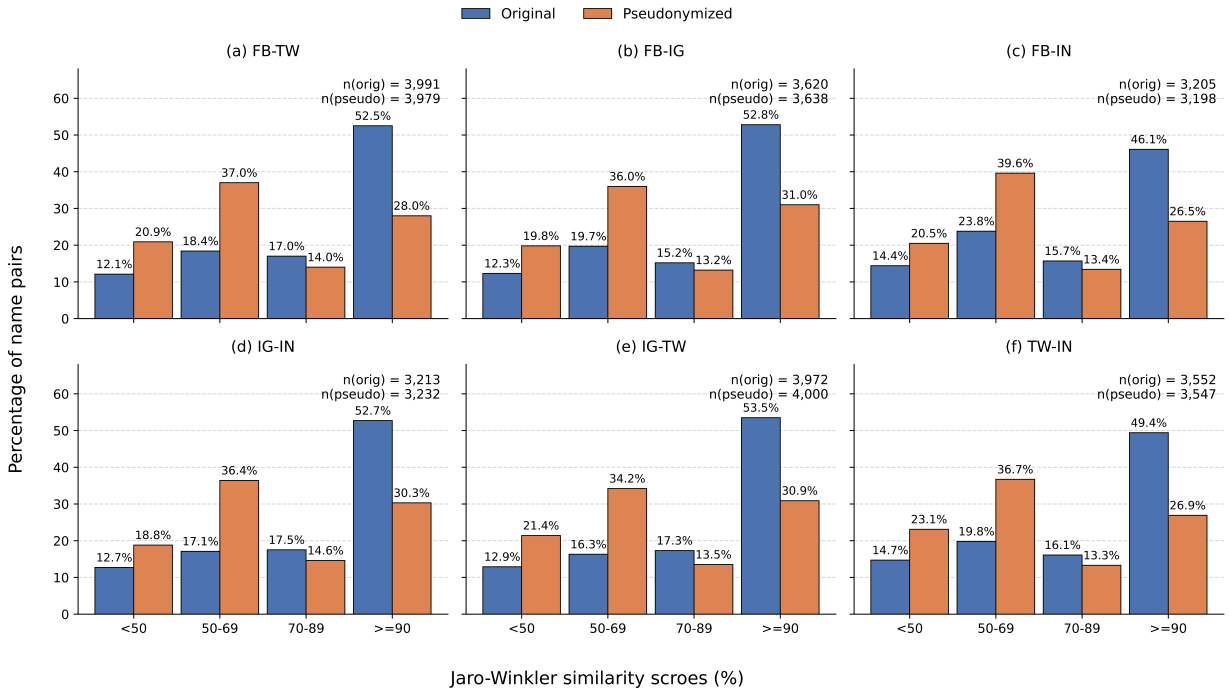

Figure 3. Similarity distribution of names. The slight difference of numbers of names from those of Table I indicates that some names in original and pseudonymized names are duplicated.

V. Conclusion and Future Work

This study examined the impact of LLM-based pseudonymization on the similarity of social media identifiers across platforms. Using Linktree-based profile connection datasets, we analyzed the impact of pseudonymization on two frequently used identifiers for cross-platform identity linkage: usernames and personal names. Our findings reveal that LLM-based pseudonymization significantly reduces the similarity of identifiers across platforms. This reduction

TABLE IV. RE-IDENTIFICATION ATTACK RESULTS

Platform	Identifiers	Pool	MRR	MeanR	MedR	Top-1	Top-3	Top-5	Top-10
Facebook	Username	4,169	0.035	1,329.96	965.0	2.10 %	3.39 %	4.31 %	6.06 %
	Name	4,168	0.059	1,145.24	762.5	3.84 %	6.05 %	7.50 %	9.21 %
Twitter	Username	4,646	0.035	1,525.25	1,193.0	2.12 %	3.60 %	4.24 %	5.68 %
	Name	4,646	0.086	1,085.69	572.0	5.76 %	8.93 %	10.71 %	13.58 %
Instagram	Username	4,149	0.041	1,326.86	996.0	2.57 %	3.91 %	5.31 %	6.43 %
	Name	4,115	0.078	1,066.99	639.5	5.19 %	8.29 %	9.90 %	12.29 %
LinkedIn	Username	3,697	0.026	1,115.44	837.0	1.45 %	2.44 %	3.32 %	4.29 %
	Name	3,697	0.035	1,137.68	848.0	1.98 %	3.56 %	4.61 %	5.95 %

is particularly pronounced for usernames, which often contain unique identifiers that are reused across multiple SNS. Person names exhibit a smaller decrease in similarity, suggesting that linguistic structure is partially preserved during pseudonym generation. These results highlight a privacy-utility trade-off. While pseudonymization weakens direct identifier correspondence and reduces the risk of identity reconstruction across platforms, partial structural similarity remains, which may support analytical tasks based on similarity. From an information systems perspective, this behavior is relevant to social media analytics scenarios in which identifiers must be transformed before analysis or data sharing in order to comply with privacy requirements.

Future research could expand upon this work in several ways. First, including additional profile attributes, such as biographies, profile descriptions, and user-generated content, would allow for a better evaluation of pseudonymization in richer social media contexts. Second, alternative baselines, prompting strategies, and model configurations could be explored to assess the impact of structural similarity preservation, reproducibility, and output consistency. Third, future studies could investigate the privacy-utility trade-off further by analyzing downstream analytical tasks and evaluating more advanced cross-platform linkage and re-identification attacks under stronger adversarial settings. Finally, future studies should address ethical limitations and residual risks, emphasizing that pseudonymization does not guarantee anonymization.

ACKNOWLEDGMENT

This research within the MuQuaNet project is partly funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union – NextGenerationEU.

REFERENCES

[1] DataReportal, *Digital 2026: Global overview report*, Feb. 2026.

[2] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – challenges in topic discovery, data collection, and data preparation", *International Journal of Information Management*, vol. 39, pp. 156–168, 2018, ISSN: 0268-4012. DOI: 10.1016/j.ijinfomgt.2017.12.002.

[3] M. Wang, W. Chen, J. Xu, P. Zhao, and L. Zhao, "User profile linkage across multiple social platforms", in *Web Information Systems Engineering – WISE 2020*, Z. Huang, W. Beek, H. Wang, R. Zhou, and Y. Zhang, Eds., Cham: Springer International Publishing, 2020, pp. 125–140, ISBN: 978-3-030-62005-9.

[4] C. Senette, M. Siino, and M. Tesconi, "User identity linkage on social networks: A review of modern techniques and applications", *IEEE Access*, vol. 12, pp. 171 241–171 268, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3500374.

[5] F. S. Bäumer, S. Schultenkämper, M. Geierhos, and Y. S. Lee, "Mirroring Privacy Risks with Digital Twins: When Pieces of Personal Data Suddenly Fit Together", *SN COMPUT. SCI.*, vol. 5, 2024. DOI: 10.1007/s42979-024-03413-z.

[6] S. Min, Y. Luo, K. Liu, Q. Gong, and Y. Chen, "From identification to obfuscation: A survey of cross-network mapping and anti-mapping methods", *Computers, Materials and Continua*, vol. 86, no. 2, pp. 1–23, 2025, ISSN: 1546-2218. DOI: 10.32604/cmc.2025.073175.

[7] S. Schultenkämper, F. S. Bäumer, B. Bellgrau, Y. S. Lee, and M. Geierhos, "From Digital Tracks to Digital Twins: On the Path to Cross-Platform Profile Linking", in *Enterprise Design, Operations, and Computing. EDOC 2023 Workshops*, T. P. Sales et al., Eds., Cham: Springer Nature Switzerland, 2024, pp. 158–171, ISBN: 978-3-031-54712-6. DOI: 10.1007/978-3-031-54712-6_10.

[8] S. Denisov and F. S. Bäumer, "The Only Link You'll Ever Need: How Social Media Reference Landing Pages Speed Up Profile Matching", in *Information and Software Technologies*, A. Lopata, D. Gudonienė, and R. Butkienė, Eds., Cham: Springer International Publishing, 2022, pp. 136–147, ISBN: 978-3-031-16302-9. DOI: 10.1007/978-3-031-16302-9_10.

[9] S. J. De and A. Imine, "Privacy Risk Analysis of Online Social Networks", *Synthesis Lectures on Information Security, Privacy, and Trust*, vol. 10, no. 1, 2020, ISSN: 1945-9742. DOI: 10.2200/s01056ed1v01y202009spt024.

[10] W. Wong, Z. Alomari, and Y. Liu, "Linkage Deanonymization Risks, Data-Matching and Privacy: A Case Study", in *8th International Conference on Cryptography, Security and Privacy (CSP)*, 2024. DOI: 10.1109/CSP62567.2024.00010.

[11] Y. Li, Y. Peng, Z. Zhang, H. Yin, and Q. Xu, "Matching user accounts across social networks based on username and display name", *World Wide Web*, vol. 22, no. 3, 2019, ISSN: 15731413. DOI: 10.1007/s11280-018-0571-4.

[12] A. Majeed and S. Lee, "Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey", *IEEE Access*, vol. 9, 2021, ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3045700.

[13] J. Zhao, B. Su, X. Rao, and Z. Chen, "A Cross-Platform Personalized Recommender System for Connecting E-Commerce

and Social Network", *Future Internet*, vol. 15, no. 1, 2023, ISSN: 1999-5903. DOI: 10.3390/fi15010013.

[14] M. G. Ayadi, H. Mezni, H. Elmannai, and R. I. Alkanhel, "Privacy-preserving cross-network service recommendation via federated learning of unified user representations", *Data & Knowledge Engineering*, vol. 158, p. 102 422, 2025, ISSN: 0169-023X. DOI: 10.1016/j.datak.2025.102422.

[15] European Commission, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, 2016.

[16] European Union Agency for Cybersecurity (ENISA), "Data pseudonymisation: Advanced techniques and use cases", European Union Agency for Cybersecurity (ENISA), Technical Report, Jan. 2021. DOI: 10.2824/860099.

[17] European Union Agency for Cybersecurity (ENISA), "Deploying pseudonymisation techniques: The case of the health sector", European Union Agency for Cybersecurity (ENISA), Technical Report, Mar. 2022. DOI: 10.2824/092874.

[18] E. Volodina, S. Dobnik, T. L. M. Tiedemann, and X. S. Vu, "Grandma Karl is 27 years old - research agenda for pseudonymization of research data", in *Proceedings - IEEE 9th International Conference on Big Data Computing Service and Applications, BigDataService 2023*, 2023. DOI: 10.1109/BigDataService58306.2023.00047.

[19] Y. S. Lee, H. Bothe, and M. Geierhos, "A mixed-methods approach to pseudonymizing users' social media profile data", in *2025 12th International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Nov. 2025, pp. 50–57. DOI: 10.1109/SNAMS67467.2025.11390950.

[20] O. Yermilov, V. Raheja, and A. Chernodub, "Privacy- and Utility-Preserving NLP with Anonymized Data: A case study of Pseudonymization", in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. DOI: 10.18653/v1/2023.trustnlp-1.20.

[21] T. Yang, X. Zhu, and I. Gurevych, "Robust utility-preserving text anonymization based on large language models", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 28 922–28 941, ISBN: 979-8-89176-251-0.

[22] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review", *SIGKDD Explor. Newsl.*, vol. 18, no. 2, pp. 5–17, Mar. 2017, ISSN: 1931-0145. DOI: 10.1145/3068777.3068781.

[23] A. Rossi, M. P. Arenas, E. Kocyigit, and M. Hani, "Challenges of protecting confidentiality in social media data and their ethical import", in *2022 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022, pp. 554–561. DOI: 10.1109/EuroSPW55150.2022.00066.

[24] D. Pissarra et al., "Unlocking the potential of large language models for clinical text anonymization: A comparative study", in *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, I. Habernal et al., Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 74–84.

[25] S. Sousa, M. Jantscher, M. Kröll, and R. Kern, "Large Language Models for Electronic Health Record De-Identification in English and German", English, *Information*, vol. 16, no. 2, Feb. 2025, ISSN: 2078-2489. DOI: 10.3390/info16020112.

[26] O. Dorémus et al., "Harnessing Moderate-Sized Language Models for Reliable Patient Data Deidentification in Emergency Department Records: Algorithm Development, Validation, and Implementation Study", *JMIR AI*, vol. 4, e57828, Apr. 2025, ISSN: 2817-1705. DOI: 10.2196/57828.

[27] A. Grattafiori et al., *The Llama 3 Herd of Models*, 2024. arXiv: 2407.21783 [cs.AI].

[28] Ollama, "Ollama", Accessed: May 29, 2026. [Online]. Available: https://ollama.com.

[29] RapidFuzz, "JaroWinkler GitHub Repository", Accessed: May 29, 2026. [Online]. Available: https://github.com/rapidfuzz/JaroWinkler.

[30] W. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage", in *Proceedings of the Survey Research Methods*, 1990, pp. 354–359.

[31] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks", in *Proceedings of the 2003 International Conference on Information Integration on the Web*, ser. IIWEB'03, Acapulco, Mexico: AAAI Press, 2003, pp. 73–78. DOI: 10.5555/3104278.3104293.

[32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, Web publication at informationretrieval.org.

[33] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data", in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[34] N. Craswell, "Mean reciprocal rank", in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds., Boston, MA: Springer US, 2009, pp. 1703–1703, ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9_488.