www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 13 Issue 09 September 2024, Page No. 26468-26476
ISSN: 2319-7242 DOI: 10.18535/ijecs/v13i09.4923

Advancing Data Anonymization Techniques for Secure and Privacy-Preserving Data Sharing in The Era of Big Data

Ravinder kaur [1], Sonia Rani [2*], Sagar Jambhorkar [3], Chitra Desai [4]

[1] Department of Computer Science, National Defence Academy, Pune, India

[2] Department of Computer Science, National Defence Academy, Pune, India

[3] Department of Computer Science, National Defence Academy, Pune, India

[4] Department of Computer Science, National Defence Academy, Pune, India

Abstract

In the era of big data, the secure sharing of sensitive information across various domains such as healthcare, finance, and social networks has become increasingly vital. Traditional data anonymization techniques often struggle to balance the competing demands of preserving privacy and maintaining data utility, particularly in complex and dynamic data-sharing environments. This paper presents a novel hybrid approach to data anonymization that integrates differential privacy with adaptive anonymization algorithms, specifically designed to enhance privacy protection while retaining the analytical value of the data. The proposed methodology tailor's anonymization strategies to the specific context of data sharing, effectively addressing the limitations of existing techniques. Extensive experiments conducted on diverse datasets, including healthcare and financial data, demonstrate the superior performance of this approach in reducing re-identification risks while maintaining high data utility. The findings suggest that these advancements in anonymization techniques provide a robust solution for secure and privacy-preserving data sharing, addressing the growing challenges posed by the increasing volume and sensitivity of data in modern digital ecosystems. The paper concludes with a discussion of the broader implications for cybersecurity and suggests future research directions to further enhance privacy-preserving technologies.

Keywords: Fake news, Machine learning algorithms, Natural language processing, Data mining, Python programming

1. Introduction

In today's data-driven landscape, the proliferation of big data has fundamentally transformed industries such as healthcare, finance, and social media. These sectors now routinely collect, process, and share vast amounts of sensitive information, necessitating robust privacy and security measures. However, the growing volume and complexity of this data have heightened concerns regarding the adequacy of traditional data protection techniques. Conventional anonymization methods, including k-anonymity, l-diversity, and t-closeness, have been widely employed to protect personal information by obfuscating identifiable details (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007). Despite their usefulness, these methods often fall short in balancing the trade-off between data privacy and utility, particularly in high-dimensional and diverse data-sharing environments where maintaining the analytical value of the data is crucial (Fung et al., 2010).

The rise of sophisticated cyber threats has underscored the limitations of existing anonymization techniques. High-profile data breaches and unauthorized access incidents highlight the vulnerabilities of traditional approaches, especially when dealing with complex datasets that span multiple domains and formats (Narayanan & Shmatikov, 2008). Moreover, the advent of big data analytics has introduced additional challenges, as the need for detailed and accurate data analysis often conflicts with the necessity of preserving individual privacy (Zhang et al., 2017; Li et al., 2018).

To address these challenges, recent research has focused on developing more advanced privacy-preserving mechanisms. Differential privacy has emerged as a leading framework, providing strong theoretical guarantees by introducing controlled noise into datasets, thereby safeguarding individual data points from re-identification (Dwork, 2008; Dwork & Roth, 2014). Despite its robustness, the application of differential privacy in real-world scenarios is often limited by the difficulty of maintaining data utility while achieving stringent privacy standards (Abadi et al., 2016; McMahan et al., 2017). In response to these limitations, this paper proposes a novel hybrid approach to data anonymization that integrates differential privacy with adaptive anonymization algorithms. This approach is specifically tailored to the context of data sharing, where the balance between privacy and utility is of paramount importance. By adapting anonymization techniques based on the sensitivity of the data and the specific requirements of the data-sharing scenario, this method offers enhanced privacy protection while retaining the data's analytical value.

Through extensive experimentation on real-world datasets from the healthcare and financial domains, we demonstrate the effectiveness of this hybrid approach in significantly reducing re-identification risks and maintaining high data utility. This research contributes to the ongoing discourse on privacy-preserving technologies by offering a viable solution to the complex challenges of secure data sharing in the era of big data.This paper is structured as follows: the next section reviews the existing literature on data anonymization and secure data sharing, highlighting current research gaps. The subsequent sections detail the methodology, experimental setup, and results, followed by a discussion of the broader implications of our findings. The paper concludes with suggestions for future research directions in this critical area of cybersecurity.

2. Literature Review

The protection of sensitive information in an increasingly data-driven world has led to the development and application of various data anonymization techniques. Among these, k-anonymity has been foundational. Proposed by Sweeney (2002), k-anonymity ensures that each record in a dataset is indistinguishable from at least k-1 other records concerning certain identifying attributes. However, k-anonymity has been criticized for its susceptibility to attacks such as homogeneity and background knowledge attacks, which can lead to re-identification of individuals even within anonymized datasets (Aggarwal, 2005).

To address the limitations of k-anonymity, l-diversity was introduced by Machanavajjhala et al. (2007). L-diversity extends k-anonymity by requiring that the sensitive attributes in each equivalence class (a group of records that are indistinguishable from each other) have at least l "well-represented" values. This approach reduces the risk of attribute disclosure but can still be vulnerable when the distribution of sensitive attributes is skewed (Li et al., 2007). In response to the shortcomings of l-diversity, t-closeness was proposed by Li et al. (2007). This technique requires that the distribution of a sensitive attribute in any equivalence class be close to the distribution of the attribute in the overall dataset, thus reducing the risk of both identity and attribute disclosure. Despite its improvements, t-closeness can be challenging to apply in practice, especially in high-dimensional data where maintaining closeness across all dimensions is difficult

(Meyerson & Williams, 2004). Recent advancements have shifted focus toward more robust privacy-preserving frameworks like differential privacy, which introduces controlled noise into datasets to prevent reidentification (Dwork, 2008). Differential privacy provides a strong theoretical foundation for privacy, ensuring that the output of any analysis is not significantly affected by the inclusion or exclusion of any single data point, thus protecting individual privacy (Dwork & Roth, 2014). This approach has seen extensive adoption in fields where privacy concerns are paramount, such as in the analysis of sensitive medical data (Abadi et al., 2016). The integration of differential privacy into real-world applications, however, has faced challenges, particularly in balancing privacy with the utility of the data (Zhang et al., 2017). To mitigate this issue, recent research has explored context-aware adaptive anonymization techniques, which tailor the level of anonymization based on the data-sharing context and the sensitivity of the data (Lee & Clifton, 2011). These adaptive techniques offer a more nuanced approach by adjusting the privacy parameters dynamically, considering factors like data sensitivity and the potential risk of re-identification (Fan et al., 2020). Composition attacks have also emerged as a significant challenge in the privacy domain, where an adversary can combine multiple datasets or use auxiliary information to compromise anonymized data (Ganta et al., 2008). The resilience of anonymization techniques against such attacks has become a critical area of research, with new methods being developed to strengthen the robustness of privacy-preserving frameworks (He et al., 2014).

Moreover, there is a growing recognition of the need for privacy-preserving data publishing frameworks that are scalable and capable of handling the complexities of big data environments. These environments often involve high-dimensional data, dynamic data streams, and the need for real-time processing, all of which pose significant challenges to traditional anonymization techniques (Fung et al., 2010). Newer frameworks such as federated learning combined with differential privacy are being explored to offer scalable solutions that protect privacy while allowing collaborative data analysis across decentralized datasets (Bonawitz et al., 2019). In conclusion, the limitations of traditional anonymization techniques and the evolving challenges of secure data sharing in complex environments underscore the necessity for innovative approaches. This paper contributes to this ongoing discourse by proposing a hybrid anonymization approach that integrates differential privacy with adaptive techniques tailored to specific data-sharing contexts, providing a stronger balance between privacy protection and data utility.

3. Methodology

This section outlines the methodology for developing and evaluating the proposed hybrid data anonymization approach. The methodology integrates differential privacy with adaptive anonymization algorithms, specifically tailored to address the limitations of existing techniques while enhancing both privacy and data utility.

3.1 Differential Privacy Framework

Differential privacy is leveraged as the foundational component of our methodology, providing a rigorous mechanism to protect individual data points from re-identification by introducing controlled noise into the dataset or query responses. The amount of noise is calibrated based on the sensitivity of the data, a measure of how much any single data point can influence the output of a query (Dwork & Roth, 2014).

In this research, we apply differential privacy using the Laplace mechanism for continuous attributes and the exponential mechanism for categorical attributes. The privacy budget (ϵ\epsilonϵ) is a crucial parameter in this framework, determining the trade-off between privacy protection and data utility. A smaller ϵ\epsilonϵ value indicates stronger privacy but at the cost of reduced data utility, and vice versa. To optimize this trade-off, the privacy budget is allocated adaptively across different attributes

depending on their sensitivity and importance to the analysis (Abadi et al., 2016).

3.2 Adaptive Anonymization Algorithms

To complement the differential privacy framework, adaptive anonymization algorithms are employed, which dynamically adjust the level of anonymization based on the specific context of data sharing. These algorithms consider various factors, including the type of data, the sensitivity of the attributes, the potential risk of re-identification, and the data utility requirements.

The adaptive process involves categorizing the data into different sensitivity levels based on predefined criteria, such as the frequency of occurrence of values and their potential for identifying individuals. For highly sensitive data, the algorithm applies stricter anonymization techniques, such as generalization and suppression, in conjunction with differential privacy. For less sensitive data, minimal anonymization is applied to preserve utility while still providing a baseline level of privacy protection (Fan et al., 2020). This context-aware approach draws on the concept of contextual integrity, emphasizing the importance of the specific context in which data is shared to ensure that privacy is protected without unnecessarily sacrificing data utility (Nissenbaum, 2004).

3.3 Implementation and Evaluation

The proposed hybrid approach was implemented and evaluated using real-world datasets from the healthcare and financial domains, where privacy concerns are particularly significant. The datasets were pre-processed to remove direct identifiers, such as names and social security numbers, before applying the anonymization techniques.

The evaluation process involved several key steps:

Baseline Comparison: The effectiveness of the proposed hybrid approach was compared with traditional anonymization techniques, including k-anonymity, l-diversity, and t-closeness (Sweeney, 2002; Machanavajjhala et al., 2007; Li et al., 2007). The comparison focused on key metrics, including the level of privacy protection, measured by the re-identification risk, and data utility, measured by the accuracy of data analysis tasks such as classification and clustering. Privacy and Utility Trade-off: The trade-off between privacy and utility was analyzed by varying the privacy budget (ϵ\epsilonϵ) and observing the impact on data utility. This analysis was critical in demonstrating the advantages of the adaptive approach, which dynamically adjusts anonymization levels to balance privacy and utility effectively. Statistical Analysis: To assess the significance of the results, statistical methods such as t-tests and ANOVA were employed. These techniques compared the performance of the proposed method with baseline methods across multiple datasets and scenarios. The results were validated using cross-validation techniques to ensure robustness.

4. Experiment and Results

Experiment Overview: Experiments were conducted on two real-world datasets from the healthcare and financial domains. The primary objectives were to assess privacy protection, evaluate data utility, and analyze the privacy-utility trade-off.

Privacy Protection Assessment: A simulated re-identification attack was conducted to evaluate the privacy protection of the anonymized datasets. The attack model assumed that the adversary had access to background knowledge, such as quasi-identifiers. The re-identification risk was quantified as the percentage of correctly re-identified records in the anonymized dataset.

Data Utility Assessment: The utility of the anonymized data was evaluated by training machine learning models, such as decision trees and logistic regression, on the original and anonymized datasets. Performance metrics included accuracy, precision, recall, and F1-score.

Privacy-Utility Trade-off Analysis: The trade-off was assessed by plotting re-identification risk against model accuracy for different values of the privacy budget (ϵ\epsilonϵ).

Experimental Findings: The proposed hybrid approach demonstrated a significant reduction in re-identification risk while maintaining high data utility, outperforming traditional methods across all evaluated metrics.

Table 1 Comparative Analysis of Anonymization Techniques

Data set	Anonymization Technique	Re-identification Risk (%)	Classification Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Healthcare Dataset	k-Anonymity (k=5)	18.5%	74.3%	72.1%	70.4%	71.2%
	l-Diversity (l=3)	14.2%	78.9%	76.5%	75.2%	75.8%
	t-Closeness (t=0.2)	12.1%	80.4%	78.9%	78.0%	78.4%
	Proposed Hybrid Approach	**6.3%**	**85.7%**	**84.2%**	**83.5%**	**83.8%**
Financial Dataset	k-Anonymity (k=5)	22.7%	70.5%	68.3%	67.1%	67.7%
	l-Diversity (l=3)	16.9%	74.6%	72.8%	71.5%	72.1%
	t-Closeness (t=0.2)	14.3%	77.3%	75.6%	74.8%	75.2%
	Proposed Hybrid Approach	**8.9%**	**82.1%**	**80.4%**	**79.7%**	**80.0%**

The results presented in the table 1 highlight the effectiveness of different anonymization techniques, including k-anonymity, l-diversity, t-closeness, and the proposed hybrid approach, across two datasets: a healthcare dataset and a financial dataset. The metrics evaluated include re-identification risk, classification accuracy, precision, recall, and F1-score.

a. Re-identification Risk

Re-identification risk is a critical measure of privacy protection, indicating the likelihood that an individual's identity can be inferred from the anonymized data.

- **k-Anonymity:** The re-identification risk is relatively high, with 18.5% for the healthcare dataset and 22.7% for the financial dataset. This suggests that k-anonymity, even with a moderate value of k=5, may not be sufficient to protect against re-identification, especially in the financial dataset where the risk is notably higher.
- **l-Diversity:** l-Diversity shows an improvement over k-anonymity, reducing the re-identification risk to 14.2% in the healthcare dataset and 16.9% in the financial dataset. However, this method still leaves a considerable risk, indicating that while it protects against attribute disclosure, it is not fully effective in high-risk scenarios.
- **t-Closeness:** t-Closeness further reduces the re-identification risk to 12.1% and 14.3% for the healthcare and financial datasets, respectively. This demonstrates its effectiveness in providing better protection by ensuring that the distribution of sensitive attributes within equivalence classes closely matches that of the entire dataset.
- **Proposed Hybrid Approach:** The hybrid approach significantly outperforms the other techniques, reducing the re-identification risk to just 6.3% in the healthcare dataset and 8.9% in the financial dataset. This indicates that the integration of differential privacy with

adaptive anonymization techniques offers a more robust defense against re-identification, addressing the limitations of the traditional methods.

b. Classification Accuracy

Classification accuracy measures how well machine learning models can correctly predict or classify data after anonymization.

- **k-Anonymity:** With accuracies of 74.3% for the healthcare dataset and 70.5% for the financial dataset, k-anonymity shows that it retains a reasonable level of utility. However, the relatively low accuracy suggests that the anonymization process distorts the data, affecting the performance of machine learning models.
- **l-Diversity:** l-Diversity improves accuracy to 78.9% for the healthcare dataset and 74.6% for the financial dataset. This improvement indicates that l-diversity, by ensuring a broader representation of sensitive attributes, helps preserve more of the data's analytical value.
- **t-Closeness:** Further improvement is seen with t-closeness, where the accuracy rises to 80.4% in the healthcare dataset and 77.3% in the financial dataset. This demonstrates its effectiveness in preserving the utility of the data while providing stronger privacy guarantees.
- **Proposed Hybrid Approach:** The highest classification accuracy is achieved by the hybrid approach, with 85.7% for the healthcare dataset and 82.1% for the financial dataset. This significant increase indicates that the adaptive methods used in this approach are effective in balancing privacy and utility, ensuring that the anonymized data remains highly useful for analytical purposes.

c. Precision, Recall, and F1-Score

These metrics provide a deeper understanding of the performance of the machine learning models:

- **Precision and Recall:** Across all techniques, there is a general trend that as privacy improves (i.e., re-identification risk decreases), precision and recall also improve, indicating better model performance with more accurate and complete predictions. The proposed hybrid approach consistently shows the highest precision and recall, indicating fewer false positives and negatives.
- **F1-Score:** The F1-score, which balances precision and recall, follows a similar pattern. The hybrid approach achieves the highest F1-scores (83.8% for the healthcare dataset and 80.0% for the financial dataset), suggesting that it offers the best overall performance in terms of both privacy protection and data utility.

d. Key Insights

- **Privacy-Utility Trade-off:** The results clearly illustrate the trade-off between privacy and utility. While traditional methods like k-anonymity and l-diversity provide some level of privacy protection, they often do so at the cost of data utility. The proposed hybrid approach, however, demonstrates that it is possible to achieve strong privacy protection with minimal compromise on data utility.
- **Effectiveness Across Datasets:** The hybrid approach's consistent performance across both healthcare and financial datasets highlights its versatility and effectiveness in different data-sharing contexts.
- **Superior Privacy Protection:** By combining differential privacy with adaptive anonymization, the hybrid approach offers significantly lower re-identification risks compared to traditional methods, making it a more robust solution for secure data sharing.

Overall, the proposed hybrid approach represents a significant advancement in the field of data anonymization, offering a practical solution that

balances the need for privacy protection with the preservation of data utility, making it suitable for use in sensitive domains like healthcare and finance.

5. Conclusion

The proposed hybrid approach to data anonymization, which integrates differential privacy with adaptive anonymization algorithms, demonstrates a robust and effective solution for secure data sharing in the era of big data. The results clearly indicate that this methodology significantly outperforms traditional techniques like k-anonymity, l-diversity, and t-closeness across multiple metrics, including re-identification risk and data utility. The hybrid approach achieves a substantial reduction in re-identification risk while maintaining high classification accuracy, precision, recall, and F1-scores. This balance between privacy protection and data utility is crucial for enabling secure and privacy-preserving data sharing, particularly in sensitive domains such as healthcare and finance. The results also underscore the versatility and effectiveness of this approach across different types of datasets, highlighting its potential for broad application.

This work not only addresses the inherent limitations of traditional anonymization techniques but also sets a new standard for the development of more advanced privacy-preserving technologies. The adaptive nature of the approach allows for a nuanced application of anonymization techniques, tailored to the specific context and sensitivity of the data, thereby ensuring optimal privacy without compromising the utility of the data. Looking ahead, future research will focus on extending this framework to accommodate emerging data types and more complex data-sharing scenarios. Additionally, there is potential to integrate other privacy-preserving techniques, such as homomorphic encryption or federated learning, to further enhance the robustness of this approach. By continuing to refine and expand this methodology, we can better meet the evolving challenges of data privacy in an increasingly data-driven world.

References

1. Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570.
2. Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3-es.
3. Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. IEEE 23rd International Conference on Data Engineering, 106-115.
4. Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (CSUR), 42(4), 1-53.
5. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. Proceedings of the 2008 IEEE Symposium on Security and Privacy, 111-125.
6. Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2017). Privacy-preserving deep computation model on cloud for big data feature learning. IEEE Transactions on Computers, 65(5), 1351-1362.
7. Li, J., Qiu, Z., Xiao, Y., & Zhao, H. (2018). Privacy-preserving data publishing: A survey on recent developments and future directions. ACM Computing Surveys (CSUR), 51(4), 1-36.
8. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.
9. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4), 211-407.

10. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.
11. McMahan, H. B., Moore, E., Ramage, D., & Hampson, S. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics, 1273-1282.
12. Aggarwal, G. (2005). k-Anonymity: An Enhanced Privacy Model. In Proceedings of the 21st International Conference on Data Engineering (ICDE), 96-105.
13. Meyerson, A., & Williams, R. (2004). On the complexity of optimal k-anonymity. In Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 223-228.
14. Lee, J., & Clifton, C. (2011). How much is enough? Choosing ε for differential privacy. In Proceedings of the 14th Information Security Conference (pp. 325-340). Springer, Berlin, Heidelberg.
15. Fan, L., Jin, H., & Wang, W. (2020). Differential privacy preservation in big data analytics: A survey. IEEE Communications Surveys & Tutorials, 22(2), 1126-1166.
16. Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 265-273). ACM.
17. He, X., Han, S., & Huang, T. S. (2014). Robust de-anonymization: Measuring the privacy risk of social network publishing. In Proceedings of the 23rd International Conference on World Wide Web (pp. 419-430). ACM.
18. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., & van Overveldt, T. (2019). Towards federated learning at scale: System design. In Proceedings of the 2nd SysML Conference, Stanford, CA, USA.
19. Nissenbaum, H. (2004). Privacy as contextual integrity. Washington Law Review, 79(1), 119-158.