Article Type: Data Anonymization

SD*k*A: Synthetic Data Integrated *k*-Anonymity Model for Data Sharing with Improved Utility

Abdul Majeed, and Seong Oun Hwang, *Gachon University, Seongnam, 13120, South Korea*

Abstract—This paper presents a novel anonymity model named “SDkA” that establishes the synergy between synthetic data (SD) and k-anonymity (k-A) model as a privacy-utility enhancer in data sharing scenarios. The SDkA introduces a data-level solution for data quality, diversity, and quantity enrichment and controls higher modifications in the data. Specifically, SD produced by a conditional GAN, and an automated method for fixing commonly encountered quality-related problems, are introduced to address size, quantity, and diversity issues. By introducing these enhancements in the data, a better defense against diverse privacy attacks is accomplished, compared to standard k-A model. Following this, the selective generalization concept is introduced as an optimization in the conventional k-A model to prevent unnecessary generalization, thereby further fortifying the k-A model to yield enhanced utility. Detailed experiments prove that SDkA effectively defends against diverse types of attacks and significantly improves the utility of the anonymized data, compared to the k-A model.

***Index Terms**– Synthetic data, privacy, k-anonymity, utility, data generalization*

In today's highly digitized era, data sharing (DS) has become vital for knowledge discovery and mining purposes, which in return can contribute to improving the quality of real-world services. However, the existence of personal information in the data is a major obstacle, and data owners are often reluctant to share personal data with third parties or analysts. To prevent privacy issues from arising, data owners under legal obligations (e.g., Europe's GDPR) are required to employ privacy-enhancing technologies (PETs) like masking, anonymization, obfuscation, and differential privacy. Unlike the others, anonymization is widely used to address privacy issues while improving data utility in DS. The k-anonymity (*k*-A) model and its improved versions are mostly used to address privacy issues in DS [1]. However, *k*-A-based anonymization methods face two major issues in the context of data-driven products and personalized services[1]:

- All personally identifiable parts of the data are removed as standard practice, impacting the development of personalized services.
- Remaining information in the data is heavily modified/anonymized, seriously hindering the knowledge discovery process.

Although many enhancements have been made in *k*-A to preserve utility without compromising privacy, it is still challenging to resolve the privacy-utility trade-off. Recently, synthetic data (SD) (i.e., a coarse form of real data (RD)) has become one of the leading PETs, and it has already contributed to solving diverse privacy issues. According to Gartner, companies can use SD to avoid 70% of fines imposed for privacy violations[2]. In a recent study, SD outperformed two widely adopted privacy solutions for DS in terms of utility [2]. Considering the potential of SD from diverse viewpoints, it is necessary to explore SD use in conjunction with anonymization methods. However, the synergy of SD with *k*-A has been explored in a limited way (e.g., parameter satisfaction only [3], or ensuring *k*-A in SD creation [4]). To bridge this gap, we implement a new model where SD is linked with *k*-A to improve defense as well as data utility. We especially enhance diversity

XXXX-XXX © 2025 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

[1] https://standards.ieee.org/industry-connections/activities/synthetic-data/

[2] https://www.ciodive.com/news/IT-predictions-Gartner-2022/608517/

in sensitive information via SD to increase privacy protection. Conversely, we lower the generalization degree where diversity becomes high due to SD addition to augment utility. Our major contributions are as follows.

1) We analyze the poor defense against diverse privacy threats and low utility guarantees offered by the well-known privacy model named k-A when facing poor-quality data, and we explore opportunities to devise a novel model called Synthetic Data k-Anonymity (SDkA) by linking SD with k-A to address the above-cited problems in k-A.
2) We implement a practical method to address the data quality, diversity, and scarcity issues to augment the defense level of k-A against diverse attacks, which is limited to two attacks only [5].
3) By improving data quality, we optimize the k-A model by lessening unneeded generalizations to increase data utility, which is very low in k-A due to wide-interval generalization or suppression.
4) Through SDkA, we highlight how small adjustments in data can accomplish the privacy guarantees of multiple privacy models while yielding significantly better utility for downstream tasks.

Experiments were performed on benchmark datasets, and results are compared with k-A and other baseline methods to justify the SDkA's feasibility in DS. The data owners can integrate SDkA in the DS component of their data life cycle by sequentially implementing all five steps. SDkA can act as a PET for personal DS with analysts for knowledge discovery. We further demonstrate the efficacy of SDkA in enterprise information systems to address real-world challenges along with three potential use cases in **Appendix A.**

Background and Related Work

This section discusses the background and related state-of-the-art (SOTA) studies.

Overview of the data model

Tabular data is the most widely used form of data in the privacy literature. $T = \{X_1, X_2, \dots, X_N\}$ denotes a real dataset in table form where X_i denotes the entire i'th record encompassing basic plus sensitive information. Consider a scenario in which T contains m attributes; then $m - 1$ attributes are regarded as basic attributes (a.k.a. quasi-identifiers (QIDs)), one of them being the sensitive attribute (SA).

Overview of k-A and its extensions

The example of k-A on a sample of six records is in **Appendix B**. The k-A model simply ensures that each class must encompass at least k records with identical QID values. However, this notion was proven weak due to over-reliance on QIDs only while ignoring SAs. As a result, two more models (ℓ-diversity [6] and t-closeness [7]) that consider SAs alongside QIDs were proposed to rectify the k-A model. In these models, SA values are bounded by parameters ℓ and t, respectively. However, certain data characteristics can undermine the applicability of these models, and anonymized data is not secure from privacy threats. Also, information loss is hefty from these models due to over-generalization or suppression. Besides, they offer protection for a limited array of attacks at the expense of data utility.

Analysis of related SOTA studies

Of late, many studies have been proposed that rectify the previous developments and yield better utility and privacy results. We classify the existing developments into the following two broad categories.

- Algorithmic solutions: These solutions introduce modified algorithms (or relax parameters) to yield greater privacy and higher utility. Srijayanthi and Sethukarasi [8] introduced a feature selection concept to reduce overhead in the clustering process when yielding k-A datasets. Similarly, the clustering concept was modified to pick records farthest from cluster centers to reduce frequent updating [9].
- Data solutions: These solutions alter the structure of the data to make the computations faster without losing guarantees of privacy and utility. Mehta and Rao [10] devised a data-level solution to sort columns based on the number of unique values in order to accomplish k-A in a short time.

In algorithmic solutions, the anonymity model architecture, parameters, objectives, optimization procedure, etc. are modified while the data remains unchanged. In contrast, data solutions modify data structure, clean data, curate more data, etc. while the underlying anonymity model remains unchanged. The proposed model belongs to the category of perturbative methods where actual values of raw data are modified to generalized values to preserve privacy [11].

The Proposed SDkA Model

In this section, we introduce the proposed SDkA model and explain its key modules, with Fig. 1 showing the workflow, and its five key components discussed below.

1- Applying SDC practices to the data

At the outset, common statistical disclosure control (SDC) practices are applied to prepare data for anonymization, including 1) attribute classification into QIDs, direct identifiers (DIDs), non-sensitive attributes (NSAs), and SAs, 2) DID removal, 3) QID grouping, 4) restructuring the SAs (e.g., moving SAs to the end), and 5) deciding whether to retain/delete NSAs. Below, we discuss the purpose of each step.

This article has been accepted for publication in IT Professional. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MITP.2025.3634142

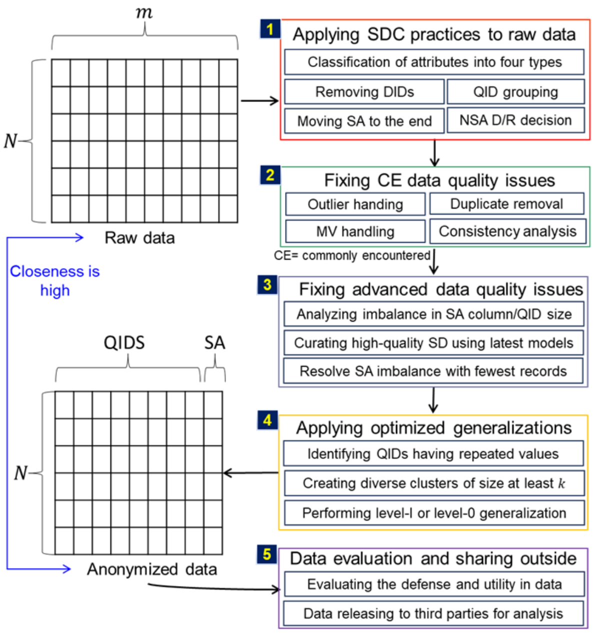

m
N
Raw data
1 Applying SDC practices to raw data
Classification of attributes into four types
Removing DIDs
QID grouping
Moving SA to the end
NSA D/R decision
2 Fixing CE data quality issues
Outlier handing
Duplicate removal
MV handling
Consistency analysis
CE= commonly encountered
Closeness is high
3 Fixing advanced data quality issues
Analyzing imbalance in SA column/QID size
Curating high-quality SD using latest models
Resolve SA imbalance with fewest records
QIDS
SA
N
4 Applying optimized generalizations
Identifying QIDs having repeated values
Creating diverse clusters of size at least k
Performing level-I or level-0 generalization
Anonymized data
5 Data evaluation and sharing outside
Evaluating the defense and utility in data
Data releasing to third parties for analysis

FIGURE 1. Workflow of the proposed five-step SD*k*A model.

1) Attribute classification → To pay attention to identity-revealing attributes, and to protect SAs.
2) DID removal → To lower the chances of explicit identity disclosure, which often leads to SA disclosure as well.
3) QID grouping → To execute relevant computing techniques (e.g., distances for clustering), and to satisfy privacy definitions (k-A, (α,k)-A, etc.).
4) Moving SAs to the end → To give systematic structure to the data, making interpretation easier (e.g., QID connections with SAs).
5) Decisions about NSA deletion/removal → To either augment utility (if retained) or remove NSAs to lower computing complexity and avoid privacy issues (when combined with QIDs).

With the help of the above-cited processes, initial preprocessing is performed to accomplish SDC criteria.

2- Resolving commonly encountered data quality issues

After SDC-based pre-processing, the data is further improved by fixing the commonly encountered data quality issues. To do so, we employed common practices used in the machine learning literature. It is vital to note there can be multiple problems in the data that require detection before resolution. Also, detection methods can vary depending on data composition and structure (all numerical attributes, all categorical attributes, and/or mixed). In this work, we consider a mixed dataset encompassing both numerical QIDs (NQIDs) and categorical QIDs (CQIDs), respectively. To fix commonly encountered problems, we adopted a set of detection and resolution techniques from our recent work [12]. It is worth noting that fixing commonly encountered problems or deleting incomplete parts of the data offers very little or no assistance in the context of anonymity models. Quantity, diversity, and scarcity improvements are imperative to advance protection status or improve utility, as discussed in the next step.

3- Amalgamating SD with RD to resolve advanced issues (e.g., size/imbalance)

To rectify k-A and to yield better privacy-utility requirements, we devised the idea of linking SD with RD. The rationale for using SD is to enhance diversity in the SAs, which is crucial for lowering the posterior probability of the attacker [13]. In some cases, the use of SD is imperative owing to greater differences between SA values. For example, in the Kaggle stroke prediction dataset[3], there are 249 instances where the stroke probability equals 1. In contrast, 4861 instances have a stroke probability of 0. In this case, a sophisticated anonymity model can leak SA privacy because there will be many classes/clusters where the SA value is zero (no stroke), particularly when k is small. To address this problem, a data-level solution is more appropriate because advanced algorithmic solutions like ℓ-diversity or t-closenss cannot offer much assistance.

To curate the SD, we employ the open-source implementation of the conditional tabular generative adversarial network (CTGAN)[4]. The main reason to use CTGAN is its ability to equally explore all values in the SD generation process from RD via condition which leads to SD of high fidelity and diversity. Besides, CTGAN has the least computing complexity than other alternatives (e.g., TVAE, Gaussian Copula), and is the ideal choice for tabular modality. Later, we add a few synthetic records with an under-represented (UR) SA value to improve RD quality in terms of the three characteristics cited above. The workflow of the mechanism used in amalgamating SD with RD is in Fig. 2. In this workflow, the RD is analyzed first w.r.t. SA values, and SA values are classified as major or minor/UR. Later, the imbalance ratio (IR) is computed between major and UR SA values by using Eq. 1:

$$IR = \frac{|Major\_SA\_value|}{|Minor\_SA\_value|} \tag{1}$$

[3] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[4] https://github.com/sdv-dev/CTGAN

This article has been accepted for publication in IT Professional. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MITP.2025.3634142

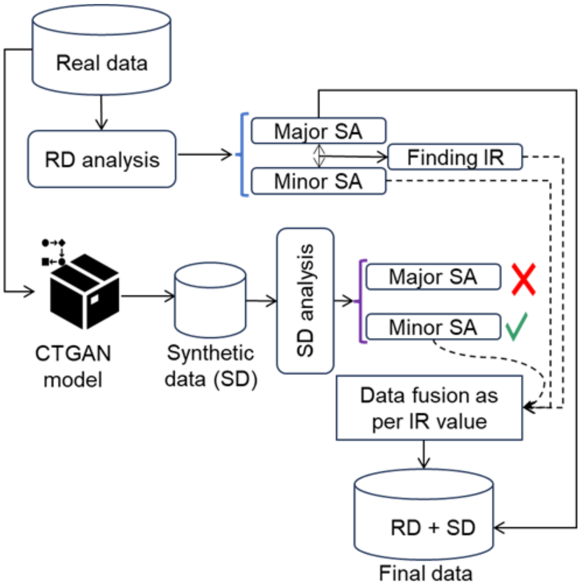


FIGURE 2. Workflow for amalgamating SD with RD.

where the numerator denotes the number of samples for the major SA value, and the de-numerator denotes the number of samples for the minor SA value.

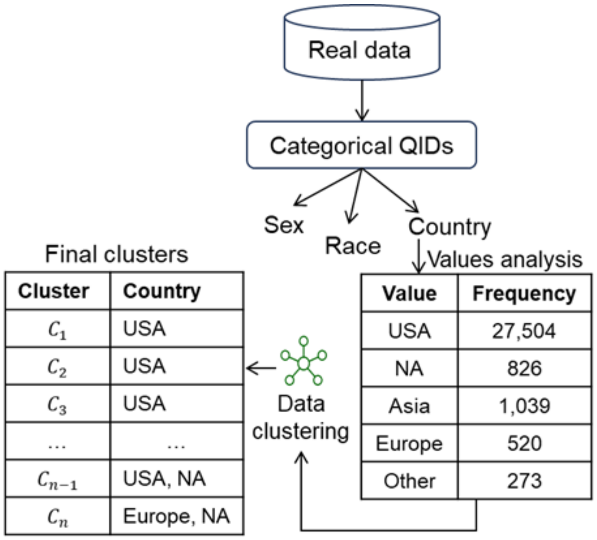
A higher *IR* indicates more imbalance between SA values. In the next step, SD is curated with the help of the CTGAN. The input to the CTGAN is the RD, condition vector, labels of the CQIDs, number of records to be curated, and other model-related parameters (learning rate, decay, # of epochs, etc.). The output is SD with a structure identical to the RD.

After curating the SD, a similar analysis is performed, and major and minor SAs are separated from each other. Since a major SA in the SD is no longer needed for data balancing, further computation is skipped. The minor SA part of the SD is further processed before fusing it with the minor SA part of the RD. It is worth noting that some quality-related issues (e.g., inconsistent ranges in NQIDs, and unexpected values in some CQIDs) can occur in SD curated with the CTGAN, and therefore, some pre-processing is applied before fusion. Later, SD and RD are fused for minor SA values while taking into account the *IR*. In the last step, instances having a major SA and an augmented minor SA are fused to curate the final data. The dataset is then sent to the *k*-A for processing.

4-Optimizing *k*-A via reduced generalization

In this component, we apply the *k*-A model on the final dataset curated from the previous step along with optimization in the generalization process. To group similar records to yield higher utility, we use the *k*-means clustering algorithm. The algorithm to partition dataset T into C clusters where the size of each cluster is at least k is given in **Appendix C**. The output of Algorithm 1 is the set of clusters, which pass through an optimized generalization process to curate anonymized data, denoted T'.

Optimized generalization: Before presenting the optimized generalization process, we discuss an example that can lead to optimized generalization. Fig. 3 shows the overview of CQID composition in the UCI Adult dataset. In this example, we highlight the value composition in the Country column (a QID); however, a similar scenario occurs for Sex and Race where one value spans many clusters.



Cluster	Country
C_1	USA
C_2	USA
C_3	USA
…	…
C_{n-1}	USA, NA
C_n	Europe, NA

Value	Frequency
USA	27,504
NA	826
Asia	1,039
Europe	520
Other	273

FIGURE 3. Analysis of Country QID value composition (NA = North America).

Referring to Fig. 3, we can see that well over 90% of the records are from the United States of America, which is regarded as population-level information. We call it population-level information because this value alone can span many clusters, offering similar uncertainty when generalized to America. As discussed in previous research [14], population-level information does not endanger privacy, and therefore can skip heavy generalization. In this work, we adopt that notion for CQIDs to optimize generalization in the *k*-A model. In the SD*k*A, NQIDs are replaced with cluster centers, and CQIDs skip generalization (if they belong to the dominant one, and all records in the cluster share the same value) or level-1 anonymity (the direct ancestor of real values in the generalization tree), leading to better utility.

It is worth noting that neither level-1 anonymity nor removing CQIDs from generalization in some clusters poses any risk based on the following six rationales: (i) all directly identifying information is removed from the data at the cleaning stage, per SDC criteria, (ii) diversity in SA values is enhanced due to SD additions that can limit an attacker's posterior probability, (iii) NQIDs and some CQIDs with no dominant values are always generalized, (iv) indistinguishability between RD and $T^{'}$ is high, which makes re-identification harder, (v) every user has $k-1$ other users with identical QIDs in each class, and (vi) record placements underwent heavier changes, making signaling difficult. Based on the above rationales, it is fair to say that introduced optimization does not lead to any privacy risk. Lastly, SD-enhanced and level-1/level-0 anonymity significantly boosts data utility for downstream tasks.

5- Anonymized data evaluation and sharing

In the last component, $T^{'}$ is evaluated with the help of diverse privacy and utility metrics and is shared with third parties for secondary use. Data seekers/consumers utilize this data for building classifiers or other kinds of analysis to improve real-world services.

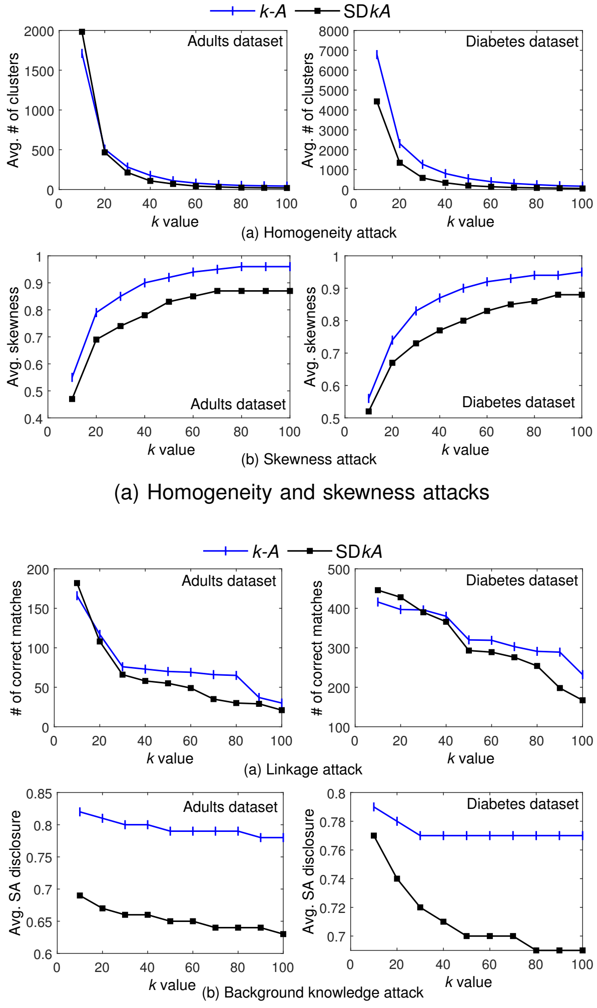
Performance Evaluation

To prove the feasibility of the proposed idea, we carried out a reasonable number of experiments in terms of privacy protection and data utility enhancement. Next, we provide the details of the datasets, implementation setup, and numerical results and comparisons.

Datasets: To evaluate the efficacy of SD*k*A, we conducted experiments on two real-life datasets: Adult[5] and Diabetes 130-US Hospitals[6], which are publicly available at the UCI ML Repository[7] and are considered benchmarks in the anonymization literature. Due to space limits, further details of the datasets and implementation setup are in **Appendix D** and **E**.

1- Privacy results comparisons

To benchmark SDkA with standard *k*-A in terms of privacy protection, we tested efficacy on four different attacks: homogeneity, skewness, background knowledge, and linking. The details of each attack can be learned from [5]. We created 10 anonymous versions of each dataset and carefully computed the privacy risk for all four attacks. For example, under the homogeneity attack, we found the number of clusters having one SA value. In the skewness attack, we analyzed the cluster that fulfills the diversity criteria but where the distribution skew for one value is high, leading to privacy disclosure. For the background knowledge and linking attacks, we exposed some information from the QIDs in the real data and computed the SA disclosure risks from successful matches. Fig. 4 contrasts the performance on all four attacks between *k*-A and the SD*k*A. In Fig. 4, we can see that SD*k*A outperformed the *k*-A and yielded better privacy for most *k* values.



(a) Homogeneity and skewness attacks

(b) Linking and background knowledge attacks

FIGURE 4. Privacy protection results: SD*k*A versus *k*-A.

From the results in Fig. 4, we conclude that the proposed SD*k*A rectifies the *k*-A issues and yields better privacy protection. In the adult dataset, the SD*k*A, on average, shows 8.8%, 11.98%, 17.70, and 21.86% improvement over the *k*-A model for four different privacy attacks(homogeneity, skewness, linking, and background knowledge). In the diabetes dataset, the SD*k*A, on average, shows 9.21%, 38.68%, 7.05%, and 11.82% improvement over the *k*-A model for four dif-

[5] https://archive.ics.uci.edu/dataset/2/adult

[6] https://archive.ics.uci.edu/dataset/296/diabetesþ130-usþhospitalsþforþyearsþ1999-2008

[7] https://archive.ics.uci.edu/

ferent privacy attacks(homogeneity, skewness, linking, and background knowledge). These results underscore better privacy protection when attacks are launched to infer either SA or identities. Lastly, we verified that SDkA can provide a defense for more attacks than previously anticipated in [5], see **Appendix F** for detailed comparative analysis. To enhance the credibility of our findings/results presented in Fig. 4, we calculate the standard deviation, which is summarized in Table 1.

TABLE 1. Standard deviation comparisons: SD*k*A versus *k*-A.

Dataset	Analysis	SD*k*A	*k*-A
Adults	Total	164.17	327.83
	Average	41.04	81.96
Diabetes	Total	1377.77	1990.17
	Average	344.44	497.54

From the results presented in Table 1, SD*k*A's performance remains stable across various privacy attacks, ensuring the consistency of the results. Based on the results, it is fair to say that SD*k*A can provide a defense for diverse attack types and is more robust than the *k*-A. It can simultaneously meet the privacy guarantees of more privacy models. These improvements have been brought on due to higher indistinguishability induced via SD blending to RD.

2- Utility results comparisons

To benchmark SD*k*A against standard *k*-A and two other SOTA methods in terms of utility enhancement, we trained a random forest (RF) classifier and computed accuracy (*A*) via Eq. 2:

$$A = {}^{T_P+T_N}/_{|T|} \qquad (2)$$

where T_P and T_N denote true positive and true negative, respectively. $|T|$ is the size of the dataset.

To gauge the effectiveness of SD*k*A in terms of utility, the RF classifier was trained on anonymized data by varying *k*. Fig. 5 presents the average accuracy after varying *k* 10 times. From the results, we can see that *A* decreases with *k* due to the increase in generalizations. However, accuracy from SD*k*A was higher than most of the baseline methods. The reason for the higher *A* is data quality enhancement via SD and optimized generalization. Although ImSKA [10] is a data-level solution and is competitive, *A* was still low because imbalance issues were not resolved. The SD*k*A, on average, shows 9.67% and 11.80% improvements in terms of *A* than ImSKA on adults and diabetes datasets, respectively. The improvements compared to two other baselines (i.e., *k*-A and KACPC) are even higher than the ImSKA across datasets. These results underscore the abilities of the proposed SD*k*A to yield higher *A* w.r.t. analytics. *A*'s results that were obtained from whole *T* are in **Appendix G**.

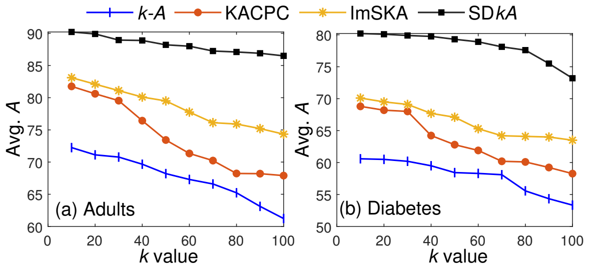


FIGURE 5. Average *A*: SD*k*A versus existing schemes.

The standard deviation of *A*'s results given in Fig. 5 for adults and diabetes datasets are: (*k*-A, 3.41; 2.47), (KACPC, 5.15; 3.76), (ImSKA, 2.92; 2.41), (SD*k*A, 1.21; 2.17). The lower values of SD*k*A ensure consistent performance in terms of *A* across datasets.

To further verify the efficacy of the SD*k*A in terms of utility, we calculated information loss (*ILoss*) and compared the results with *k*-A model. We used formalization from [15]. The formalization used in calculating *ILoss* from a cluster C_i is given as follows.

$$ILoss(C_i) = |C_i| \times T|C_i| \qquad (3)$$

where C_i is any anonymized cluster, $|C_i|$ denotes the number of records in a cluster, and $T|C_i|$ is the per-record distortion.

The $T|C_i|$ separately computes distortion from NQIDs and CQIDs induced by the generalization operation and adds them together [15]. We computed the total *ILoss* by adding each cluster *ILoss*, and the results are in Fig. 6. From the results, it can be seen that *ILoss* increases with *k* due to an increase in generalization. The SD*k*A controls the generalization issues by fulfilling diversity criteria in most clusters via SD addition, leading to lower *ILoss* in both datasets.

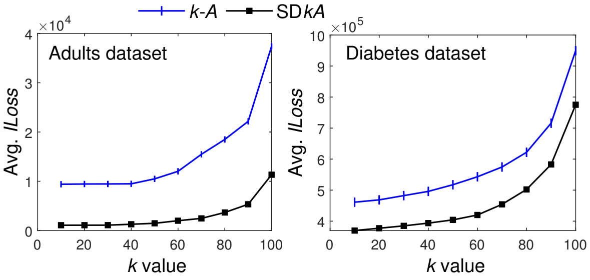


FIGURE 6. Avg. reduction in *ILoss*: SD*k*A versus *k*-A.

The results in Figures 4 to 6 underscore the potency of SD*k*A in yielding higher utility in downstream tasks while safeguarding privacy. Lastly, SDkA performs lower anonymity/generalization than k-A model. See **Appendix H** for generalization analysis. See **Appendix I** for the current scope and applicability of SDkA to heterogeneous data types.

Insights and Novel Findings

Through this work, we derived the five key insights that might help the privacy and database community to develop improved privacy models or rectify previous ones. See **Appendix J** for details of the five key insights. Apart from the insights and novel findings, the use of SD can contribute to preserving essential characteristics of RD, which can contribute to the secondary use of data in diverse applications or use cases. Lastly, recent developments have shown that SD can outperform both anonymization and semantic methods [2], therefore, its use in DS is imperative.

Conclusion and Future Work

In this paper, we first discussed the phenomenon of poor privacy guarantees and low utility when utilizing the k-A model in DS scenarios, and we analyzed the underlying causes. Subsequently, we proposed a novel anonymity model named SDkA, which comprises two sub-modules incorporating three data pre-processing techniques (SDC-based, commonly encountered, and advanced) followed by optimization of the generalization process. SDkA effectively mitigates poor privacy guarantees and low-utility issues while utilizing some high-quality synthetic records, making it more suitable for SD scenarios involving skewed datasets. Experimental results demonstrate that, compared to SOTA models, SDkA achieves higher data utility and exhibits greater defense across a variety of privacy attacks. In the future, we plan to use diverse generative models (e.g., TVAE, Gaussian Copula) to assess the impact of data repair on different anonymization models.

REFERENCES

1. L. Sweeney, “k-anonymity: A model for protecting privacy,” *International journal of uncertainty, fuzziness and knowledge-based systems*, vol. 10, no. 05, pp. 557–570, 2002.
2. Q. Razi, S. Datta, V. Hassija, G. Chalapathi, and B. Sikdar, “Privacy utility tradeoff between pets: Differential privacy and synthetic data,” *IEEE Transactions on Computational Social Systems*, 2024.
3. G. Polese and M. M. Lucia, “‘augmenting anonymized data with ai: Exploring the feasibility and limitations of large language models in data enrichment,” in *Proc. 3rd Italian Conf. Big Data and Data Sci*, 2024.
4. J. Domingo-Ferrer, K. Muralidhar, and S. Martínez, “Synthetic data generation via the permutation paradigm with optional k-anonymity,” *IEEE Transactions on Dependable and Secure Computing*, 2025.
5. J. Sáinz-Pardo Díaz and Á. López García, “A python library to check the level of anonymity of a dataset,” *Scientific Data*, vol. 9, no. 1, p. 785, 2022.
6. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” *Acm transactions on knowledge discovery from data (tkdd)*, vol. 1, no. 1, pp. 3–es, 2007.
7. N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in *2007 IEEE 23rd international conference on data engineering*. IEEE, 2006, pp. 106–115.
8. S. Srijayanthi and T. Sethukarasi, “Design of privacy preserving model based on clustering involved anonymization along with feature selection,” *Computers & Security*, vol. 126, p. 103027, 2023.
9. H. Wang, J. He, and N. Zhu, “Improving data utilization of k-anonymity through clustering optimization.” *Trans. Data Priv.*, vol. 15, no. 3, pp. 177–192, 2022.
10. B. B. Mehta and U. P. Rao, “Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 4, pp. 1423–1430, 2022.
11. Y. Wei, H. Y. Benson, J. K. Agor, and M. Capan, “Multi-objective optimization-based anonymization of structured data for machine learning,” *arXiv preprint arXiv:2501.01002*, 2025.
12. A. Majeed and S. O. Hwang, “A data-centric l-diversity model for securely publishing personal data with enhanced utility,” *IEEE Transactions on Big Data*, 2025.
13. T. Mimoto, S. Kiyomoto, K. Tanaka, and A. Miyaji, “(p, n)-identifiability: Anonymity under practical adversaries,” in *2017 IEEE Trustcom/BigDataSE/ICESS*. IEEE, 2017, pp. 996–1003.
14. M. Strobel and R. Shokri, “Data privacy and trustworthy machine learning,” *IEEE Security & Privacy*, vol. 20, no. 5, pp. 44–49, 2022.
15. J.-W. Byun, A. Kamra, E. Bertino, and N. Li, “Efficient k-anonymization using clustering techniques,” in *International conference on database systems for advanced applications*. Springer, 2007, pp. 188–200.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (RS-2024-00340882). Seong Oun Hwang & Abdul Majeed are

co-corresponding authors of this article.

Authors Biography

Abdul Majeed is currently working as an Assistant Professor at the Department of Computer Engineering, Gachon University, Korea. His research interests include privacy-preserving data publishing, information privacy, federated learning, and machine learning. Contact him at ab09@gachon.ac.kr.

Seong Oun Hwang is currently working as a Professor at the Department of Computer Engineering, Gachon University, Korea. He is a senior member of IEEE. His research interests include cryptography, cybersecurity, and artificial intelligence. Contact him at sohwang@gachon.ac.kr.