

Shielding privacy: a technique of extenuating composition attacks in various independent data publication

Md. Omar Faruq¹, Md. Abul Ala Walid², Mrinal Kanti Baowaly³, Maloy Kumar Devnath³, Md. Sabbir Ejaz⁴,
Pronob Kumar Barman⁵, A H M Sarowar Sattar⁶

¹Department of Computer Science and Engineering, Bangladesh Army University of Engineering and Technology, Natore, Bangladesh

²Department of Data Science, Bangabandhu Sheikh Mujibur Rahman Digital University, Gazipur, Bangladesh

³Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh

⁴Department of Information and Communication Engineering, Noakhali Science and Technology University, Noakhali, Bangladesh

⁵Department of Statistics, University of Kentucky, Lexington, United States

⁶Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh

Article Info

Article history:

Received May 19, 2023

Revised Dec 1, 2023

Accepted Feb 24, 2024

Keywords:

Composition attacks

k-anonymity

Knowledge domain

l-diversity

Unauthorized access

ABSTRACT

Protecting personal information from unauthorized access is a critical concern for individuals. However, the accumulation of confidential information by various organizations, such as banks and hospitals, for regular communication creates a potential vulnerability. If an individual visits two hospitals and both facilities independently release the individual's gathered data, a malicious adversary could potentially deduce confidential information through a composition attack. Therefore, developing methods that protect individuals from composition attacks is crucial. According to the size of the dataset and the percentage of overlapping persons, our study examines the effectiveness of composition attacks. We propose a knowledge domain-based design to mitigate successful composition attacks, which has shown promising results in reducing such attacks and compared to existing studies based on the k-anonymity and *l*-diversity models. Our approach leverages a knowledge domain to reduce the likelihood of data breaches, demonstrating the effectiveness of our method in protecting individuals' privacy and preventing unauthorized access to sensitive information. Finally, the effects of data utility on the diverse data set have been measured.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Md. Abul Ala Walid

Department of Data Science, Bangabandhu Sheikh Mujibur Rahman Digital University

Gazipur-1750, Bangladesh

Email: abulalawalid@gmail.com

1. INTRODUCTION

In recent years, the preservation of privacy in the publication of data has grown to be an issue with amazing potential. Organizations are recipients of personal data and thereafter engage in the dissemination of such data in accordance with governmental and legal regulations. A distinct organization, such as a bank, hospital, etc., gathers and disperses personal information in a way that conceals the identity of the individual (e.g., information about a person's health or finances, for example, is disguised).

We assume that these organizations are in multiple independent data publishing environments. In that situation, a person may go to various hospitals, and their private information may be disclosed by invoking a composition attack. Although several distributed knowledge sets pose little secrecy upon personnel, studies show that a person's secrecy is at risk of raising disparate organizations to have some common tuples, and they release the common data sets individually [1].

There exists every comprehensive number concerning personal data during the arrangement preceding mean while each knowledge-based policy toward several fields [2]. These data should remain accessible through those organizations so that people support enlarging their information for decisiveness planning. Existing security measures include [3], [4] but they do not take into account the situation where a patient visits many institutions for the same disease. By this moment, an opponent may employ a composition attack [5], [6] touching distributed data sets to acknowledge the privacy that is maintained by traditional security principles.

A person's private data remained preserved within each purpose from the k-anonymization technique that produced data seems unknown when organizations are dependent on each other. Moreover, such interaction and information sharing as followed by (e.g., [7]) approximately estimating again explained exchanges (e.g., [8], [9]) are communicated to realize and discuss possible breaches. Specifically, the before-mentioned coordination does not support also may indeed imply constrained through ordinance [10], [11]. Moreover, a person goes to multiple hospitals, and the hospitals remain independent from each other, as said in Table 1 illustrates the original data sets for Organization-A (Table 1(a)) and Organization-B (Table 1(b)).

Table 1. Preliminary data sets of: (a) Organization-A and (b) Organization-B

(a)					(b)				
Name	Gender	Age	Zip code	Diseases	Name	Gender	Age	Zip code	Diseases
Samal	M	33	7075	G	Sezan	M	25	6092	H
Rita	F	21	5085	B	Mintu	M	24	6095	C
Titu	M	22	6055	D	Adam	M	21	6090	A
Roni	M	20	6095	F	Raju	M	19	6095	G
Suntu	M	23	6095	E	Nirala	F	23	6085	I
Lema	F	23	5040	C	Nipa	F	25	6040	H
Runa	F	22	5050	E	Zoba	F	28	7050	D
Saba	F	25	5070	F	Mimi	F	29	7070	F
Rana	M	28	7012	G	Ripon	M	34	7012	E
Tutul	M	29	7020	H	Safiq	M	33	7020	D
Abir	M	31	7050	D	Rifat	M	30	7050	E
Adam	M	21	6090	A	Ratul	M	32	7075	F

Every table has two different locations as, Organization-A and Organization-B for performing composition attacks. Table 2 represents the anonymized version data sets for that Organization-A (Table 2(a)) and Organization-B (Table 2(b)) respectively. Concerning illustration for these two tables, undertake that one sufferer possesses the following secret knowledge, (sex=male, age=21, zip code=6090), acknowledged through an adversary. The adversary additionally acknowledges that these victims' recordings remain inside two data collections. Table 3 records data portions of pair published information collections including single victim's records. We interchangeably refer to many independent data publications in an uncoordinated setting throughout this paper. Data publishers use a variety of data publishing approaches, including multiple-view data release [12], [13] and serial data publication [14]–[16]. The publisher is still concerned about the anonymization procedure employed on public data sets utilised by other organisations. Here, gender, age, zip code, and sensitive attribute are each represented by the letters GN, AG, ZIC, and SA, respectively. At the time, the information holder and the information publisher possess never connected through several different also previously specific methods [8], [17]–[19] imply not employed to preserve secrecy concerning non-coordinated data publishing procedures.

Table 2. Anonymous data of: (a) Organization-A and (b) Organization-B

(a)				(b)			
Age	Gender	zip code	Diseases	Age	Gender	zip code	Diseases
16-25	F	5***	B	26-40	*	7***	H
16-25	F	5***	C	26-40	*	7***	G
16-25	F	5***	E	26-40	*	7***	E
16-25	F	5***	F	26-40	*	7***	D
16-25	M	60**	A	26-40	*	7***	E
16-25	M	60**	E	26-40	*	7***	F
16-25	M	60**	D	10-25	*	60**	H
16-25	M	60**	F	10-25	*	60**	C
26-35	M	7***	G	10-25	*	60**	A
26-35	M	7***	H	10-25	*	60**	G
26-35	M	7***	D	10-25	*	60**	I
26-35	M	7***	G	10-25	*	60**	H

Table 3. An illustration of a compositional attack

	GN	AG	ZIC	SA
Organization-A	F	16-30	6070	A
	F	16-30	6070	E
	F	16-30	6070	D
	F	16-30	6070	N
Organization-B	F	10-30	6070	H
	F	10-30	6070	C
	F	10-30	6070	A
	F	10-30	6070	M

Initially, we introduce a proposal to reduce the composition attack. We organize a collection of records of the attribute age of earlier distributed knowledge collections. Next, we discover each confidential attribute (diagnosis) toward all equivalence groups of the published data collection. Therefore, implementing the post-processing technique at k -anonymization reduces composition attacks. Finally, we analyze our all the equivalent groups of the dataset including one confidential domain so we can determine that our method can protect the secret information regarding personnel that remains preserved with each data publisher compare to the k -anonymity and l -diversity techniques. The following describes the contribution of the study:

- The strength of the various composition attacks on the different amounts of datasets with overlapping pools has been observed.
- A knowledge domain based on the statistical data has been constructed and a proposed model based on the knowledge domain has been introduced.
- The analysis has been performed that reduces the composition attacks.
- Comparing the results with the existing methods of k -anonymity and l -diversity.
- The effects of data utility on the diverse data set have been measured.

2. RELATED WORK

An increasingly popular area of research is privacy-preserving data publishing with various approaches being proposed to mitigate the risk of privacy breaches. In the literature, various approaches have been proposed to mitigate composition attacks [4]. For instance, the probabilistic method is applied in [3] to reduce the likelihood of such attacks. As an alternative, the composite process is used in [20] to defend against composition assaults. It makes use of generalization, perturbation, and suppression techniques.

Single data publication techniques, such as k -anonymity, l -diversity [21], (α, k) anonymity [22], and t -closeness [23], have also been proposed as solutions to the problem of composition attacks. Those methods are susceptible to a composition attack [6]. The IoT based approach proposed to maintain security in cloud environment [24], an association rule for hiding the sensitive knowledge as proposed the integer linear programming [25] and the paper [26] presents that how to protect the smartphone users. In order to identify a person's tuple in public information collections associated to their quasi-identifying attributes, the works [5], [27] use a complex linking procedure. Necessarily, these systems within particular example publication environments do not apply through multiple independent publications [6].

The k -anonymization technique is used by various privacy protection models as a pre-processing step. The paper by Sameesha [28] introduced a two-phase top-down specialization system concerning anonymizing individual data. The paper by Zhang *et al.* [29] proposed a methodology for refining the scalability of that anonymization system across these clouds. The paper by Khan and Malluhi [30] discussed the security problems in the cloud situation and the cause for these faith matters. Nobody confides the method with the minimum mechanism in their fingers and no photograph of the way in what way data is kept.

Through the intersection of the two equivalence classes from Tables 2(a) and (b), only a single sensitive value is specified that is constant from the individual's private information. Patient "A"'s private information is "A" and the opponent knows that the patient is male, 21 years old, and resides in the zip code 6090; therefore, the adversary can ascribe and infer the patient's information using the quasi-identifier. A patient named Adam has the disease "A" according to Tables 1(a) and (b), which is a true link and contains the original information. In our case, the study utilizes a pre-existing dataset to construct a private information domain and maximize its coverage based on the principle that aims to enhance the protection of individuals' sensitive information across various independent publications.

3. STRENGTH OF COMPOSITION ATTACKS

In this section, we perform this determination by employing various unknown data collections to confirm our evaluation. The k-anonymization techniques [5] are used to demonstrate how potent the composition assault is. Table 4 demonstrates attributes and their domain size of the dataset.

Table 4. Attributes and their domain

Attribute	Domain size
Age	16-94
Gender	2
Education	18
Birthplace	40
Race	6
Salary	50

So, the strength of the composition attack relies on the result of the common record about every match equivalence group, and each amount of separated confidential values counts the specific k-anonymity part of a person in that equivalence class. When counting the total match M_{total} then is split with that complete quantity from data set D_{total} . Now, we count the percentage of total match signified as m that estimate the strength of composition attack as (1):

$$\text{Percentage of strength, } m \frac{M_{total}}{D_{total}} \times 100 \% \quad (1)$$

All the data sets have the same attribute domain. For our discussion, we further the two pairs of disparate of two autonomous hospital datasets are the same. We use the hypothesized data set D_0^* to simulate D_2^* . As seen in Figure 1, overlapping tuples of 1000, 2000, 3000, 4000, and 5000 are shared by sets of data of 10,000 bytes each. Anyone looking for a method to lessen composition attacks in uncoordinated contexts may benefit from knowing how strong the composition attack is [31].

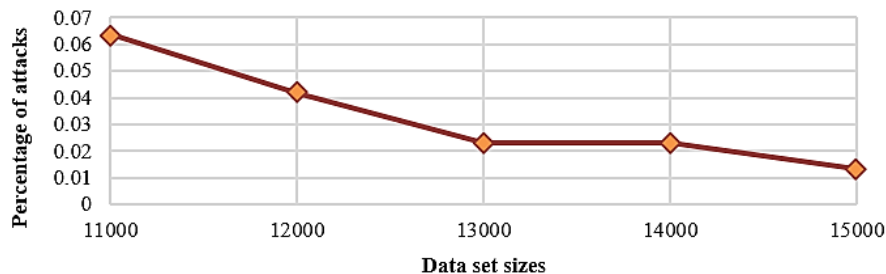


Figure 1. The computed strength rate against composition attack on k anonymized data sets for various domain sizes 11000, 12000, 13000, 14000, and 15000 with the overlapping records 1000, 2000, 3000, 4000, and 5000 respectively

4. METHOD TO REDUCE COMPOSITION ATTACKS

To illustrate our model, we first define some privacy terms and definitions that are used throughout this paper. Consider such that, any dataset introduced $D=\{t_1, t_2, \dots, t_n\}$ multiple set of concerning tuples by schema $\{ID, QI_1, \dots, QI_d, SVL\}$, wherever each t_i expresses tuple or record within D associates on individual i , assigned attribute ID comprises specific identifiers of persons QI_1, \dots, QI_d moreover comprises quasi-identifier characteristics, therefore being age, gender, and zip code, which container possibly know persons while connected between outside data. To explain those characters, assign QI to denote those set $\{QI_1, \dots, QI_d\}$. Columns within the QI container either be determined across a certain specialty (such as gender) or approximately any numeric range (such as age). Attribute SVL is the confidential value about a person, such as a disease.

4.1. Equivalence class or group

Consider an anonymized data set for the set of records from that data set that corresponds to the equivalence class, incorporating the most prevalent quasi-identifiers (like QI) with the associated values. In

an illustration, an equivalence class exists at entries 1 through 6 in Table 2(b) with attributes {age, sex, zip code}. Let D_1, D_2, \dots, D_n remain this n autonomous k -anonymized datasets comprising a minimal equivalence class of size k .

Let us think about the two autonomous anonymous data set D_1^* and D_2^* . Here follow those notes $EG_i(D_{1j}^*)$ to express each equivalence group of a person i , $QI(EG_i(D_{1j}^*))$ denotes the quasi-identifiers, and we use the symbol $SVL(EG_i(D_{2j}^*))$ to correspond to the multi-set of sensitive value in $EG_i(D_{2j}^*)$ of a published data set D_{2j}^* . Each identity group capacity k denotes an individual's anonymity as a victim. That indicates that a person should remain classified with $(k-1)$ additional people within this published data set. The level of protection may transform [31], [32] if there available various published data sets for the same person.

4.2. Knowledge of an adversary

Let's assume that a victim v is a person in the collection of data D_1 through ID equal to v and S value of s for the purposes of our discussion. The opponent and adversary used vice-versa throughout this paper. The opponent acknowledges that QI uses v , that v exists within D_1 , also these next details, and seeks to gather the person's sensitive values. Instantly, v stands additionally inside different data set D_2 . When the opponent becomes the entrance to D_1 and D_2 , the published data sets of D_1^* and D_2^* sequentially. That opponent can achieve an intersection on applying that corresponding equivalence collection to acknowledge the s of the person.

4.3. Knowledge of a data publisher

A data publisher may have not comprehensive information about the different data set that has access to the coincidental records including its distributed information collection. Nevertheless, any publisher predicts that these consequences of QI including SA of different datasets match this same source of population and that both data sets employ the corresponding k -anonymization technique. Though, QI values and confidential values of datasets D_1 including D_2 observe that corresponding arrangement. We do not study the before-mentioned limited population under here work. These data sets are anonymized correspondingly as needed by-law against a field [20].

4.4. Knowledge domain

Assume that the $D_1, D_2, \dots,$ and D_n published data set wherever the identical equivalence associations are $EG_{1D_1}, \dots, EG_{1D_1}, EG_{1D_2}, \dots, EG_{1D_2},$ and $EG_{1D_n}, \dots, EG_{1D_n}$ including the confidential attributes $SVL_{EG_{1D_1}}, \dots, SVL_{EG_{1D_1}}, SVL_{EG_{1D_2}}, \dots, SVL_{EG_{1D_2}}$ and $SVL_{EG_{1D_n}}, \dots, SVL_{EG_{1D_n}}$ sequentially. Every amount of confidential values is acknowledged to the data publisher for a particular age, gender, also country. These three quasi-identifiers, which call for some sensitive values, are situations that are often equivalent to the age, gender, and country (or location) groups. A particular knowledge domain is then indicated in (2) and the related knowledge domain described in Table 5.

$$K(\epsilon) = EG_{1D_1} \cap EG_{1D_2} = [SVL_1, SVL_2, \dots, SVL_n] \tag{2}$$

Table 5. Knowledge domain

Gender	Age	Country	Diseases
M	15-30	India	C
	15-30	India	B
	15-30	China	G
	15-30	China	D
F	15-30	India	E
	15-30	India	A
	15-30	China	F
	15-30	China	E

Wherever these acknowledged knowledge requirements denote $K(\epsilon) \neq \text{NULL}$. Here the quantity of situations the same disease exists identified upon this publisher including QI as quasi-identifier attributes one which values remain common as unique equivalence class has which kinds of conditions are the same disease followed by published data sets of Table 5.

In the illustration of Table 5, assume EG_{1D_1} plus EG_{1D_2} hold pair of equivalence groups wherever Age (EG_{1D_1})=(16-25), Age (EG_{1D_2})=(16-25), $SVL(EG_{1D_1})=(A,D,E,F)$ including $SVL(EG_{1D_2})=(B,C,E,F)$ sequentially. Hereabouts, we ought to associate information as some male patients $K(\epsilon)=\{B, C\}$ for nation India also associated information $K(\epsilon)=\{D, G\}$ for nation China. Hither, we ought to acknowledge information concerning these female patients $K(\epsilon)=\{A, E\}$ to the nation India and recognized information $K(\epsilon)=\{E, F\}$ for the nation China.

4.5. Threshold value

Every smallest amount of common confidential values by connected knowledge domain concerns determined through that publisher necessarily contain an equivalence group. That is expressed through Θ . Concerning an instance, if this threshold $\Theta < 2$ when the opponent is capable to complete a strong composition attack. Through Table 3, only one confidential value is common and the victim falls to a composition attack.

4.6. Privacy breach

Assume that $D1$ holds one data set $D1^*$ denotes its anonymized translation and i implies an individual inside that dataset. Accordingly, a secrecy breach \mathcal{B} happens when $|EG_1(D1)|$ regards the number of generally separated confidential values including that area of discovered information. Since this privacy breach transpires while $|EG_1(D1)| < \Theta_{E1}$ during our published data set, the content of our privacy policy $|EG_1(D1)| < \Theta_{EG1}$ when the two published data sets before-mentioned as Tables 2(a) and (b) a person Adam's confidential preference smaller than Θ . Later, the composition attack transpires.

Algorithm to arrive at our proposed approach

Input: Take a collection of k -anonymized data with y quasi-identifying features in $D1$.

Output: Final processed data.

```

1: Find the secret values SVL( $EG_n$ ) for each equivalence group as well as the equivalence
   groups  $EG_n$ .  $EG_n$  should stand for the number of equivalence classes.
2: While there is a class of equivalence that does not satisfy the statement  $SVL(EG_n) \geq \Theta$  do
3:   For every equivalence group, figure out the common SVL( $EG_n$ ).
4:   if  $SVL(EG_n) < \Theta$  then
5:     Select the common private value from the pool of private values as chosen  $\Theta$ 
6:   end if
7: end while
8: Final processed data.
```

Initialization (step 1) first, the input data sets are pre-processed by k -anonymity and l -diversity methods, and then we contain the entire amount of equivalence groups. We collect the equivalence group preference (label) EG_n and also the distinct sensitive values concerning every equivalence group $SVL(EG_n)$. Checking the criteria (steps 2–4) following, we determine the group of sensitive values that seem coincidentally inside an equivalence group of that hypothesized data valued $D1$, where $SVL(EG_n)$. See this S can comprise any combination of different sensitive values of S . As an instance if $SVL(EG_n) \geq \Theta$ and, then their container holds potential groups, wherever the threshold, Θ satisfies that privacy principles that each group has a minimum of two common distinct confidential values. If the equivalence group missing to complete the condition (steps 5-7) while an equivalence group abandons to provide the privacy model and the requirement at step 4 (i.e., the equivalence group is possibly controlled to a composition attack), select the common private value from the available sensitive values as chosen Θ . Output table (step 8) certainly, the before-mentioned steps produce the outputs generalized concerning dataset $D1$.

Particular sensitive values. The total of each value for the general confidential attribute concerning the identical equivalence group necessity remains larger than or similar concerning the threshold value i.e., $\Theta_{Ei} \geq \Theta$ holds our picked confidential values that directions preserve a person (the sufferer) and produce anonymity to the opponent. If the confidential values of the E_{D1} and E_{D2} two equivalence groups for any data set $D1$ and $D2$ are $SVL(EG_{D1})=(A, D, C, E, F, H)$ and $SVL(EG_{D2})=(2,6,2,8)=(A, B, C, H)$ and the selected confidential values as (A, C, H) for the two equivalence class which is larger than the threshold value.

From (2), we use to acquire the knowledge domain as our base knowledge. Table 3 shows that the adversary performs a successful composition attack for a victim through the data published by satisfying the privacy model. Although the protection model applied to Table 2 the sensitive information gathered conducted by the composition attacks. The use of the knowledge domain in this approach is an innovative task. The result of Figures 2 and 3 shows that the proposed model is better as the comparison between our method and the previous k -anonymity and l -diversity models [5], [21].

5. ANALYSIS OF RESULTS AND DISCUSSION

Figures 2 and 3 present that our proposed method is better compared to the other existing method based on k -anonymity and l -diversity. Our experimental results demonstrate that our approach is more effective in mitigating composition attacks on datasets that are continuously published by data publishers, compared to the well-known k -anonymity and l -diversity [5], [21].

Hither, this actual Census Bureau of the United States datasets (found at <http://ipums.org>) is utilized, where we make use of a data source with about 500,000 records. Age, sex, education, race, marital status,

and place of birth are used as quasi-identifying characteristics. We use the occupation and salary properties as the confidential attribute. Table 4 displays the field dimensions of all characteristics. Now, we employed two distinct data sets, namely D1 and D2, from two autonomous organisations’ data sets, from dissimilar publishers as two disparate locales. The overlapping pools were split into two groups and arranged in Figure 2 so that each group of data sets contained overlapping tuples of the same size (1010, 1020, 1050, 1080, and 1100) that were each 10, 20, 50, 80, and 100.

Making five copies of each data set in each class, we randomly added 101000, 102000, 103000, 104000, and 105000 records from the suitable pool to each copy, resulting in five sets of data with sizes of 1000, 2000, 3000, 4000, and 5000, respectively, as illustrated [33] in Figure 3.

The two established models k-anonymity and *l*-diversity make up the earlier models. Finally, the arch shows how the composition attack is reduced when applied to several data sets of various sizes. We now understand that composition assaults are always mitigated by this method. The usefulness of the output data is measured using the utility metric. To calculate the data utility, we employ (3) [34]:

$$Data\ utility = \frac{|Total\ original\ tuple - Total\ processed\ tuple|}{Total\ original\ tuple} \tag{3}$$

To assess the impact on data utility, we utilize in (3) and present the results in Figure 4. Our proposed method has a minor effect on data utility, with a loss rate of approximately 2% in the general case [33]. Even though data publisher employs the security techniques k-anonymity and *l*-diversity opponent create composition attacks to gather the confidential data of a person. On the way to prove the mitigation of composition attack in this analysis, we apply these disparate data sets of one collection of various sizes within our approach. This result shows the comparison with the k-anonymity and *l*-diversity paradigm as the earlier study.

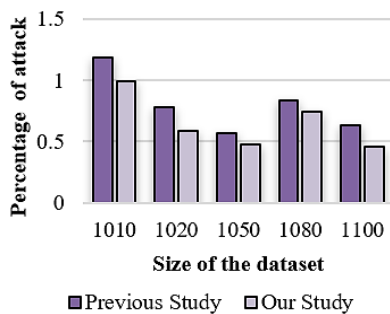


Figure 2. The comparison between the k-anonymity and *l*-diversity models [5], [21] and our approach on the D1 set where the distance between the groups is 10

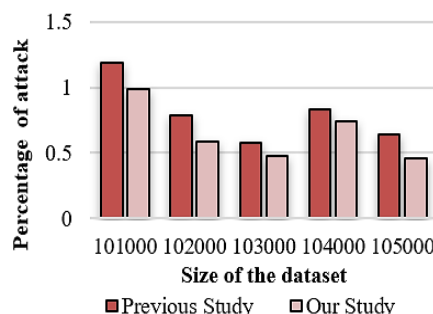


Figure 3. The comparison between the k-anonymity and *l*-diversity models [5], [21] and our approach on the D2 set where the distance between the groups is 1000

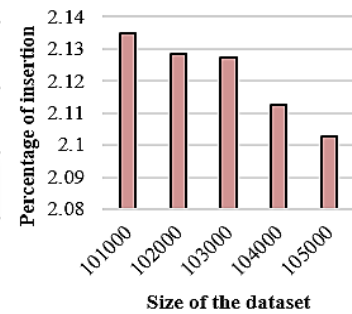


Figure 4. The rate of utility loss

6. CONCLUSION

Our study has demonstrated the efficacy of our proposed method in reducing composition attacks on privacy in multiple independent data publication. However, while our approach has shown promising results in mitigating such attacks by leveraging knowledge domains, there are still opportunities for further improvement. Our design operates under the assumption that all features of the tuple are independent, but we acknowledge the challenge of maintaining the connection between quasi-identifiers and sensitive attributes. For example, in a healthcare context, a female patient may be more likely to be diagnosed with breast cancer than a male patient. Accounting for such complex relationships between national factors is a difficult task, but we suggest that more precise analytical modeling of national associations based on distinctive attributes such as country, gender, and age range can help to reduce data erosion and improve data utility. Our proposed approach has been compared with two well-established models, k-anonymity and *l*-diversity, and has demonstrated a promising reduction in composition attacks. Future research could explore additional means to enhance the performance of our method and further improve the privacy and security of multiple independent data publications.





REFERENCES

- [1] A. H. M. S. Sattar and S. Helal, "Privacy Risk Against Composition Attack," *International Journal of Innovative Research in Computer Science & Technology*, vol. 6, no. 2, pp. 18–23, Mar. 2018, doi: 10.21276/ijrcst.2018.6.2.3.
- [2] F. Provost and T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013, doi: 10.1089/big.2013.1508.
- [3] A. H. M. S. Sattar, J. Li, J. Liu, R. Heatherly, and B. Malin, "A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments," *Knowledge-Based Systems*, vol. 67, pp. 361–372, Sep. 2014, doi: 10.1016/j.knosys.2014.04.019.
- [4] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: a survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, Jun. 2010, doi: 10.1145/1749603.1749605.
- [5] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002, doi: 10.1142/S0218488502001648.
- [6] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith, "Composition attacks and auxiliary information in data privacy," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2008, pp. 265–273, doi: 10.1145/1401890.1401926.
- [7] B. Malin, "k-Unlinkability: A privacy protection model for distributed data," *Data & Knowledge Engineering*, vol. 64, no. 1, pp. 294–311, Jan. 2008, doi: 10.1016/j.datak.2007.06.016.
- [8] W. Jiang and C. Clifton, "A secure distributed framework for achieving k-anonymity," *The VLDB Journal*, vol. 15, no. 4, pp. 316–333, Nov. 2006, doi: 10.1007/s00778-006-0008-z.
- [9] B. Malin, "Secure construction of k-unlinkable patient records from distributed providers," *Artificial Intelligence in Medicine*, vol. 48, no. 1, pp. 29–41, Jan. 2010, doi: 10.1016/j.artmed.2009.09.002.
- [10] R. D. Cebul, J. B. Rebitzer, L. J. Taylor, and M. E. Votruba, "Organizational Fragmentation and Care Quality in the U.S. Healthcare System," *Journal of Economic Perspectives*, vol. 22, no. 4, pp. 93–113, Oct. 2008, doi: 10.1257/jep.22.4.93.
- [11] B. Malin and L. Sweeney, "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems," *Journal of Biomedical Informatics*, vol. 37, no. 3, pp. 179–192, Jun. 2004, doi: 10.1016/j.jbi.2004.04.005.
- [12] C. Yao, X. S. Wang, and S. Jajodia, "Checking for k-anonymity violation by views," *VLDB 2005 - Proceedings of 31st International Conference on Very Large Data Bases*, vol. 3, pp. 910–921, 2005.
- [13] B. Yang, H. Nakagawa, I. Sato, and J. Sakuma, "Collusion-resistant privacy-preserving data mining," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, Jul. 2010, pp. 483–492, doi: 10.1145/1835804.1835867.
- [14] K. Wang and B. C. M. Fung, "Anonymizing sequential releases," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2006, pp. 414–423, doi: 10.1145/1150402.1150449.
- [15] R. C.-W. Wong, A. W.-C. Fu, J. Liu, K. Wang, and Y. Xu, "Global privacy guarantee in serial data publishing," in *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 2010, pp. 956–959, doi: 10.1109/ICDE.2010.5447859.
- [16] X. Xiao and Y. Tao, "M-invariance," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, Jun. 2007, pp. 689–700, doi: 10.1145/1247480.1247556.
- [17] P. Jurczyk and L. Xiong, "Privacy-preserving data publishing for horizontally partitioned databases," in *Proceedings of the 17th ACM conference on Information and knowledge management*, Oct. 2008, pp. 1321–1322, doi: 10.1145/1458082.1458257.
- [18] N. Mohammed, B. C. M. Fung, K. Wang, and P. C. K. Hung, "Privacy-preserving data mashup," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, Mar. 2009, pp. 228–239, doi: 10.1145/1516360.1516388.
- [19] S. Goryczka, L. Xiong, and B. C. M. Fung, "M-Privacy for Collaborative Data Publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2520–2533, 2014, doi: 10.1109/TKDE.2013.18.
- [20] J. Li, M. M. Baig, A. H. M. S. Sattar, X. Ding, J. Liu, and M. W. Vincent, "A hybrid approach to prevent composition attacks for independent data releases," *Information Sciences*, vol. 367–368, pp. 324–336, 2016, doi: 10.1016/j.ins.2016.05.009.
- [21] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkatasubramanian, "L-diversity: privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3, Mar. 2007, doi: 10.1145/1217299.1217302.
- [22] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α , k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2006, pp. 754–759, doi: 10.1145/1150402.1150499.
- [23] 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*, Apr. 2007, pp. 106–115, doi: 10.1109/ICDE.2007.367856.
- [24] S. M. Rukmony and S. Gnanamony, "Rough set method-cloud internet of things: a two-degree verification scheme for security in cloud-internet of things," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, p. 2233, Apr. 2023, doi: 10.11591/ijece.v13i2.pp2233-2239.
- [25] B. Suma and G. Shobha, "Association rule hiding using integer linear programming," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, p. 3451, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3451-3458.
- [26] A. Zharova, "The protect mobile user data in Russia," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 3, pp. 3184–3192, Jun. 2020, doi: 10.11591/ijece.v10i3.pp3184-3192.
- [27] W. E. Winkler, "Advanced Methods for Record Linkage," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1994, no. 1645, pp. 467–472.
- [28] V. Sameesha, "A Scalable Two Phase Top Down Specialization Approach For Data Anonymization Using Mapreduce On Cloud," *International Journal of Computer Trends and Technology*, vol. 45, no. 1, pp. 50–53, 2017, doi: 10.14445/22312803/ijctt-v45p110.
- [29] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud," *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 1008–1020, 2014, doi: 10.1016/j.jcss.2014.02.007.
- [30] K. M. Khan and Q. Malluhi, "How can cloud providers earn their customers' trust when a third party is processing sensitive data in a remote machine located in various countries," *Emerging technologies can help address the challenges of trust in cloud computing*, 2010.
- [31] M. O. Faruq and A. H. M. S. Sattar, "Strength of Composition Attacks in Multiple Independent Data Publication," *2020 IEEE Region 10 Symposium, TENSYP 2020*, pp. 1022–1025, 2020, doi: 10.1109/TENSYP50017.2020.9230903.





- [32] S. Kim, M. K. Sung, and Y. D. Chung, "A framework to preserve the privacy of electronic health data streams," *Journal of Biomedical Informatics*, vol. 50, pp. 95–106, Aug. 2014, doi: 10.1016/j.jbi.2014.03.015.
- [33] M. O. Faruq and A. H. M. S. Sattar, "An Approach to Mitigate Composition Attacks on Privacy in Non-coordinated Environments," in *2020 2nd International Conference on Advanced Information and Communication Technology, ICAICT 2020*, 2020, pp. 123–128, doi: 10.1109/ICAICT51780.2020.9333538.
- [34] A. S. M. Hasan, Q. Jiang, and C. Li, "An Effective Grouping Method for Privacy-Preserving Bike Sharing Data Publishing," *Future Internet*, vol. 9, no. 4, p. 65, Oct. 2017, doi: 10.3390/fi9040065.

BIOGRAPHIES OF AUTHORS



Md. Omar Faruq     was born at Naogaon. He received the B.Sc. at 2016 and M. Sc. at 2020 in computer science and engineering department from Rajshahi University of Engineering and Technology (RUET), Rajshahi, Bangladesh. Now, he is working as an Assistant Professor from August 2016 at Bangladesh Army University of Engineering and Technology. He can be contacted at email: omarfarukcse10@gmail.com.







Md. Abul Ala Walid     received his B.Sc. Engineering degree from Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh, in CSE, and completed M.Sc. Engineering from KUET in the department of CSE. His areas of interest in research include computer vision, biomedical image processing, clinical data analysis, advanced machine learning, and deep learning. He can be contacted at email: abulalawalid@gmail.com.






Dr. Mrinal Kanti Baowaly     is a faculty member in the Department of CSE at Bangabandhu Sheikh Mujibur Rahman Science and Technology University in Bangladesh. He received his B.Sc. in CSE from Khulna University; and Master in Information Technology from University of Dhaka, Bangladesh. He has a Ph.D. in Social Networks and Human-Centered Computing from National Chengchi University in Taiwan and has conducted his research at Academia Sinica. His research interests include machine learning, deep learning, data science, natural language processing, social network analysis, and health informatics. He can be contacted at email: baowaly@gmail.com.






Maloy Kumar Devnath     is a doctoral student in Information Systems at the University of Maryland, Baltimore County. His research interest is mainly in Applied Machine learning, Graph-Based ML, and Sensor Computing. Prior to his doctoral studies, Maloy served as an Assistant Professor at Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj. He has completed his B.Sc. Engineering degree in Computer Science and Engineering from Bangladesh University of Engineering and Technology. He can be contacted at email: maloy.cse.buet@gmail.com.




Md. Sabbir Ejaz    is the faculty member at Noakhali Science and Technology University in Department of Information and Communication Engineering. Showing unwavering enthusiasm for research, his interests span across digital image processing, deep learning, machine learning, and artificial intelligence. He has made consistent contributions as an author or co-author to multiple impactful research publications. He can be contacted at email: sabbirejaz.ice@nstu.edu.bd.



Pronob Kumar Barman    is a doctoral candidate specializing in Information Systems at the University of Maryland, Baltimore County. His research is centered around applied machine learning and artificial intelligence. Before embarking on his doctoral journey, Pronob served as Deputy Director at Bangladesh Bank. He holds a Bachelor of Science (BS) degree in Statistics from the University of Dhaka and earned his Master of Science (MS) in Statistics from both the University of Kentucky and the University of Dhaka. He can be contacted at email: pbarman1@umbc.edu.



A H M Sarowar Sattar    was born in Rajshahi. He received the B.Sc. degrees in computer science & engineering from the Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh and M. Sc. degree from KTH Royal Institute of Technology, Germany and the Ph.D. degree in data privacy from University of South Australia. Working at Rajshahi University of Engineering and Technology as a former professor at computer science and engineering department. He is an expert on data privacy, data engineering, and cyber security. He can be contacted at email: sarwar@gmail.com.