Cell-key Perturbation Data Privacy Procedure for Security Operations Center Team

Supornpol nukrongsin
College of Digital Innovation Technology
Rangsit University
Pathumthani, Thailand
supornpol.n65@rsu.ac.th

Chetneti Srisa-An
College of Digital Innovation Technology
Rangsit University
Pathumthani, Thailand
chetneti@rsu.ac.th

2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE) | 979-8-3503-0050-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/JCSSE58229.2023.10202111

***Abstract*— Data privacy laws such as GDPR in Europe and PDPA in Thailand are both laws to protect personal data. The data center task is also a data service organization that needs to do data publishing services among their stakeholders. The challenging task for the Security Operation Center (SOC) team is to analyze all security risks such as data breaches. Most cases of data breach problems are overlooked cases that occur indirectly by guessing from other prior knowledge. For example, attackers combine our dataset with other data sets to reidentify personal data. This attack is called a re-Identification attack that causes a data breach. To fix the risk, statistical noise control techniques for data anonymization are explored and implemented in this study. A Cell-key perturbation is to fix the attack without modifying an original dataset but return an answer dataset with noise addition per query instead.**

Keywords*— *data privacy, security operation center, cell-key perturbation, data anonymization

I. Introduction

The UK Statistics and Registration Services Act's responsibility is to produce official statistics and ensure good practice usage of statistics. In these acts (2007), “personal information” is information that identifies directly or indirectly a particular person. Data privacy laws such as GDPR [18] in Europe and PDPA [19] in Thailand are both laws to protect personal data. Since then, data privacy has become a hot topic in many countries.

The difference between data security and data privacy is ‘Who’ can access data in data security and ‘when and what can they access data in data privacy. For example, login authentication term in data security is about ‘Who’ can access but Authorization in data privacy is about ‘what’ and ‘when’ of data that the user is accessible. Therefore, data privacy is a task for the SOC team rather than an individual system administrator.

According to The National Institute of Standards and Technology Framework (NIST), it is almost inevitable for a firm to have a data breach. This paper intends to propose a new framework for mitigating a data breach effect. Thailand PDPA laws argue the data controller to notify the data breach incident within 72 hours after detecting the event.

The basic approach to preventing data breaches is to hide Personal Identifier Information (PII) from being exposed by others without permission. For example, all PIIs are separately stored on different tables. Unfortunately, the datasets that exclude PII combined with other knowledge can reveal unintentionally the personal data. Because data breach causes serious problem by-laws, this research proposed a new framework to protect personal data and comply with privacy laws. For example, attackers combine our dataset with other data sets to reidentify personal data. This attack is called a re-Identification risk [9] that causes a data breach.

Data center becomes a data controller or data processor by GDPR/PDPA laws because it holds a lot of personal data. PDPA is a new data privacy law in Thailand. The riskiest data controller is a data breach. The simple way of protecting data is cyber security, but there is no guarantee for data privacy. The incident response model was developed based on the procedures and policies that were implemented within the institution. Once an incident happens, the SOC has to team determine the adjustments needed for the controls that will be implemented in different processes whether for technical or nontechnical aspects.

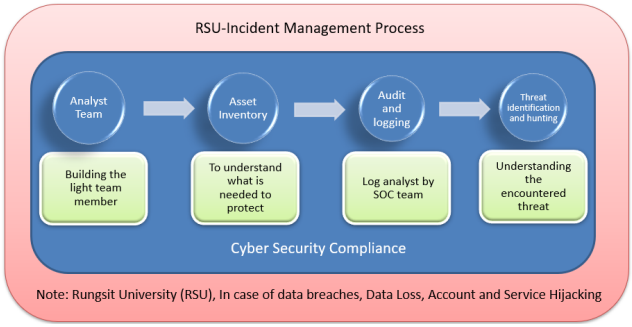


Fig. 1. RSU Incident Management Process

Figure 1 shows the RSU incident management process that handles some serious incidents such as data breaches. By Thailand law, the data controller has to report the data owner within 72 hours. The first step is to set up an analyst team who can do data analytics to find the root causes of incidents. This paper focuses on data breaches by using data from an anonymizer server shown in Figure 3.

A. Cell-key Perturbation Methodology

Sweeney [9] introduced a re-identification risk. The quasi-identifier of {age, zip code, and date of birth} can re-identify many personal data [16]. There are two main concepts to prevent a reidentification risks such as K-anonymity [9] and Cell-key perturbation [20].

A cell-key perturbation (CKP) method was developed by the Australian Bureau of Statistics (ABS) in 2021. The origin purpose is developed for the census.

A perturbation method that is one of the new data privacy concepts adds noise to frequency tables. The noise calculates from every cell value. It is more practical and faster to produce a privacy dataset and is also recommended by the UK

government. Rather than using the data swapping technique, the Cell Key Perturbation Methodology is much faster.

The CKP method adds a small amount of noise to some cells in a table, meaning that users cannot be sure whether differences between tables represent a real person, or are caused by the perturbation. The method is consistent and repeatable, so the same cells are always perturbed in the same way [4].

Cell Key Perturbation Methodology is to add statical noise to a dataset that can be reidentified. The key concept is to make a long time for attackers to reveal their original personal data.

```
df.groupby(['age','sex','race']).size().reset_index(
    name='count').sort_values('count')
```

	age	sex	race	count
402	59	Male	Amer-Indian-Eskimo	1
503	75	Male	Asian-Pac-Islander	1
501	75	Female	Black	1
291	46	Male	Other	1
434	63	Male	Other	1
...	...	...	...	...
215	38	Male	White	532
176	34	Male	White	542
185	35	Male	White	543

Fig. 2. Frequency count on quasi-identifier

Figure 2 shows a re-identification risk on the frequency of group 1 by using the group-by command. Attackers can guess who they are if they are unique. A simple group-by-command cannot reveal all other dimensions; therefore, a Perturbation table or P-table is invented for multidimensional calculation. A perturbation table applied to a cell is a frequency table that counts a group of attributes. Such a P-table can be obtained using the R-package P-table [3].

The record keys are random values for each record. When calculating a table cell, these record keys will be combined into a cell key. The noise added is based on a cell's value and a 'cell key' calculated for each cell. Cell key values are not unique and the same noise can be added to multiple cells. It is important to note that the noise is not added randomly: if the same query is run multiple times, the query will always be perturbed in the same way.

In order to apply the cell key method, a microdata file must be supplied containing categorical data. An unperturbed frequency table is generated based on the variables defined, then a 'cell key' for each value is calculated by summing the record keys - integers randomly assigned to each row - from the contributing rows in the source microdata. A lookup is then performed on a perturbation table such as a CSV file containing a list of perturbation values specifying how much noise should be applied to a cell. The final output is a perturbed frequency table that is no longer disclosive.

B. Anonymizer server

A firewall is normally a hardware for a firm to protect an organization's assets from hackers. It acts like a gatekeeper that let only authorized connection pass in a network. This equipment is vital for every network to distinguish between the internet world and the intranet world.

This paper proposed a new architecture that uses a firewall as a division line. The Anonymizer server in Figure 2 is a server in front of the firewall. After the firewall, there is a PII segment that contains all sensitive personal information. The other non-PII segment is stored separately. Both tables are linked with the primary key hashed function and stored behind a firewall. The preparation process is discussed in section 3.

The paper is organized as follows. Some previous literature is reviewed in section 2. Section 3 shows a demonstration of the research methodology. Section 4 describes the data preparation process. Section 5 reveals the experimental result. Section 6 is Future Work. Finally, in Section 7, we describe the conclusion of the study.

II. LITERATURE REVIEW

M. Mutemwa, J. Mtsweni, and L. Zimba [1] stated that SOCs' responsibilities are to detect, investigate and report on all malicious activities that occur.

Hebert Silva et al. [2] demonstrated and reaffirmed that even de-identified data that removed all direct personal data is not enough to protect against re-identification risk. Prior knowledge combined with dataset can reidentify personal information easily [2].

T. Enderle and S. Giessing [3] developed "Open Source tools for perturbative confidentiality methods". The cell key method is one of the statistic disclosure control (SDC) methods.

Dove, Iain & Ntoumos, Christos & Spicer, Keith. [20] stated that a benefit of Cell-key perturbation over k-anonymity is that it can provide better privacy protection for individual values. This is because Cell-key perturbation adds noise to each value independently, which can make it harder for an attacker to infer the true value. In contrast, k-anonymity generalizes the data within groups, which can lead to information loss and may not provide as strong privacy protection for individual values.

B. Fraser and J. Wooton. [21] proposed a method for protecting confidential information in tabular output by making it more difficult to use differencing attacks to identify sensitive information.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith [5] proposed a new mechanism that was designed to protect privacy called Differential privacy. Differential privacy (DP) itself is not an algorithm, but a framework for data privacy.

P. Cichonski, T. Millar, T. Grance, and K. Scarfone [6] stated principles for handling events are offered in the handling for reviewing data and selecting the best course of action for each occurrence.

Janos, Feher & Nguyen, Phuoc Dai [7] described all security concerns for data center industry. The main tasks of the SOC team are to detect incidents, investigate, and respond to each incident.

Emam, Khaled et, al [8] argued that most privacy laws allow healthcare institutes disclose medical records without consent if all datasets are de-identified but they tend to fail by the re-identification attacks.

$$pr(re-identification) = pr(re-identification|attempt) * pr(attempt) \quad (1)$$

There will be fewer attacks attempted on data sets that have a very low likelihood of being re-identified by an adversary [8].

The re-identified risk was found by Sweeney [9] who demonstrated that merging normal hospital discharge records with voter list records can re-identification the personal medical records of the ex-governor of Massachusetts. This evidence shows that only the removal of directly identify fields from a dataset is not enough to protect against data breach risk.

Benitez, K., & Malin, B. [12] stated that the common three fields of date of birth, gender, and residential zip code can re-identification a personal medical record.

Narayanan and Shmatikov [11] demonstrated that attackers can easily re-identification the subscriber's personal data from an anonymized Netflix shared dataset that is combined with the IMDB dataset.

Y. Ma , Y.X. Lin, et al. [15] argued that traditional methods of data protection, such as data masking or encryption, are not sufficient for protecting aggregate business data.

Khordadpour and Peyman [10] demonstrated that the main benefit of Cell-key perturbation in this paper is that it can help to reduce the risk of data breaches or unauthorized access to sensitive data.

James Bailie and Chien-Hung [13, 14] demonstrated that Cell-key perturbation is a new kind of data privacy protection and give many advantages greater than Differential privacy (DP).

Kato Mivule demonstrated [17] in his paper that data perturbation utilizing noise addition is a suitable methodology for published data sets.

III. METHODOLOGY

According to GDPR laws, Personal Identification Information (PII) has to be stored secretly away from hackers; therefore, they have to be decomposed from the original data. In this process, the SOCs team has to decompose the original data into two segments including the PII segment and the non-PII segment. For the non-PII segment, we need to add noise to mitigate and re-Identification risk. In this experiment, the SOC team applied additive noise to the numerical attributes. The following steps are the data preparation process. Firstly, PII has to be identified and then separated into a new secret table. Secondly, Cell Key Perturbation Methodology [2] is applied to the non-PII segment to prevent re-Identification risk.

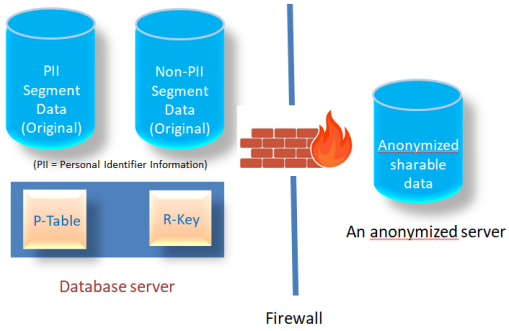


Fig. 3. Privacy data segmentation process

A. Statistical Noise control technique

The Cell-key method is one of the Statistical Noise control techniques to protect data privacy. The random record key to every data record acts as an index of the cells of a pre-defined perturbation table.

In order to return the same answer on the same query, the Cell-key method intend to assign permanently the record key to each record. These row keys are used to generate consistent values for the perturbation. Whenever the same query, the same perturbation will be applied.

Cell-key perturbation[13] is a technique used to protect sensitive data by adding random noise to the original data values. Perturbation of zeros is a specific type of cell-key perturbation where random noise is added to the cells of a dataset that contain zero values. The purpose of perturbing zero cells is to protect the privacy of individuals by preventing attackers from inferring any sensitive information from the absence of data.

By using R-key indexes, the frequency of the group was counted and summed into an integer number and then divided by modulo arithmetic to produce a cell-level key to determine the perturbation.

A cell-level key in the lookup table is the value of the perturbation that will be applied to the cell, based on the original cell value and the cell-level key. the cell-level key as a row axis, the cell values in the column is a perturbation value to add to microdata.

At this point, the method can determine the amount of noise to add to the table cell. This method ensures that each cell receives the same perturbation in a repeated process. Zero cells stand for records that do not receive any perturbation. Most of the cells are not affected by perturbation in a large dataset.

An advantage of Cell Key Perturbation is its consistent and repeatable. The cell key and the cell count are input to look up in the perturbation table and then added to the cell value. The value obtained from looking up this perturbation table is a perturbed result.

IV. DATA PREPARATION

Figure 4 shows that the original data is separated into the de-Identified segment and the Non-PII data segment. The PII segment is a table that contains all PII data such as social security number, name, last name and etc. The non-PII data segment is a de-identified dataset that removes all direct identifiers and has a primary key such as SSN on both tables for linkage.

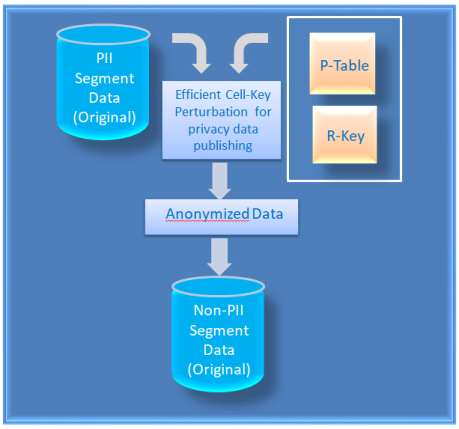


Fig. 4. Cell-key perturbation method

Figure 4 shows the Cell-key perturbation process by using R-key and P-table. The goal of the P-table package is to produce perturbation tables that can be used for applying noise to statistical tables.

V. EXPERIMENT

Python is a language to demonstrate our methodology in this study because of its flexibility and its ability to data manipulation. However, it is not suitable for commercial products and is not intended to be. This study is to demonstrate the concept of cell-key perturbation. Commercial tools are available such as an R-package and τ -argus.

Jupyter Notebook tool and python are implemented in this experiment to demonstrate a key concept. Pandas is a built-in package in the Jupyter Notebook. From de-identified data in section IV, the panda's crosstab () function in Python is used to build a frequency table in a cross-tabulation table form. It can show the frequency with which certain groups of data. A famous adult dataset is chosen because it contains a lot of quasi-identifier attributes such as {age, sex, race}. The set of three variables is a quasi-identifier because it is a minimal set. The cell key perturbation method is as following steps:

A. Step 1: Create a record key by assigning a random number between 0 and 99.

The R-key package in Figure 2 generates a row-key to use in a frequency table. The row key is added to a non-PII segment. The larger the random number, the longer computation is. Only two digits between 0 and 99 is chosen as a random number range.

```
np.random.seed(0)
df.insert(0, "row_key",np.random.randint(
    0, 100, df.shape[0]))
df.tail(10)
```

	row_key	age	sex	race
32551	87	43	Male	White
32552	67	32	Male	Amer-Indian-Eskimo
32553	46	43	Male	White
32554	6	32	Male	Asian-Pac-Islander
32555	12	53	Male	White
32556	9	22	Male	White
32557	26	27	Female	White
32558	73	40	Male	White
32559	11	58	Female	White
32560	69	22	Male	White

Fig. 5. Excerpt from experiment script in Step 1

B. Step 2: Create a frequency table using the crosstab function.

In this step, a P-table is constructed, and the crosstab function is used to count the frequency of the cell values group.

```
counts = pd.crosstab(
    [df[v] for v in list(df.columns)[1:-1]],
    df[list(df.columns)[-1]],
    dropna=False)
display(counts)
```

	race	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
age	sex					
17	Female	2	2	15	5	162
	Male	1	0	19	2	187
18	Female	2	8	18	4	236
	Male	2	3	28	2	247
19	Female	5	10	29	4	308
...	...	...	...	...	...	...
87	Male	0	0	0	0	1
88	Female	0	0	0	0	1
	Male	0	0	0	0	2
90	Female	0	0	2	0	12
	Male	0	5	2	0	22

146 rows × 5 columns

Fig. 6. Excerpt from experiment script in Step 2

C. Step 3: Calculate perturbation values on the frequency table from cell value and cell key.

```
perturbations = counts * 0

for col_idx in range(counts.shape[0]):
    for row_idx in range(counts.shape[1]):
        cell_key = cell_keys.iat[col_idx, row_idx]
        count = counts.iat[col_idx,row_idx]
        perturbation = 0
        if count > 0:
            perturbation = ptable.loc[min(count, 3), cell_key]
        perturbations.iat[col_idx, row_idx] += perturbation

display(perturbations)
```

	race	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
age	sex					
17	Female	2	2	1	1	1
	Male	0	0	2	2	2
18	Female	2	2	-3	-3	1
	Male	2	1	1	-1	-3
19	Female	-3	1	1	1	2

Fig. 7. Excerpt from experiment script in Step 3

D. Step 4: Apply all perturbation to cells

```
perturbed_counts = counts + perturbations
display(perturbed_counts)
```

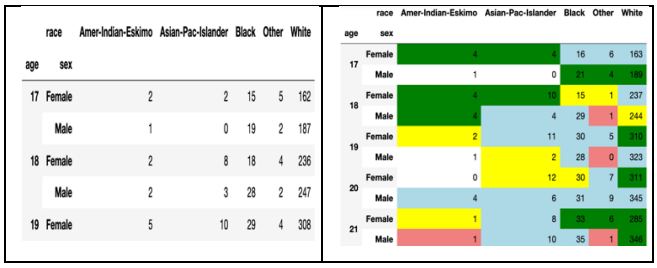
	race	Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
age	sex					
17	Female	4	4	16	6	163
	Male	1	0	21	4	189
18	Female	4	10	15	1	237
	Male	4	4	29	1	244
19	Female	2	11	30	5	310

Fig. 8. Excerpt from experiment script in Step 4

E. *Step 5: Generate new anonymized dataset.*

The final step is to construct an anonymized dataset that lies in front of a firewall. Unlike another method, this method intends to add minimal noise to the original dataset. The advantage of this method is most of the cells are not modified called “Zero cells”. The other word not all cells are modified.

This experiment shows an output from the Jupyter Notebook.



race		Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
age	sex					
17	Female	2	2	15	5	162
	Male	1	0	19	2	187
18	Female	2	8	18	4	236
	Male	2	3	28	2	247
19	Female	5	10	29	4	308

race		Amer-Indian-Eskimo	Asian-Pac-Islander	Black	Other	White
age	sex					
17	Female	4	4	16	6	163
	Male	1	0	21	4	187
18	Female	4	10	15	1	237
	Male	4	4	29	1	244
19	Female	2	11	30	5	310
	Male	1	2	28	0	323
20	Female	0	12	30	7	311
	Male	4	6	31	9	345
21	Female	1	8	33	6	285
	Male	1	10	35	1	340

Fig. 9. Unperturbed and Perturbed table output

Figures 9 (a) and 9 (b) show the output generated for all 3 variables before and after perturbation respectively. Figure 9 (a) contains the unperturbed frequency counts before values from the P-table are applied. Figure 9 (b) contains the perturbed counts; cells with noise added have been highlighted. Here is an explanation highlighted as follows in Figures 10.

Color	Summation
yellow	-3
red	-2
Light coral	-1
white	0
Light blue	+1
Green	+2
light green	+3

Fig. 10. Color highlighting table

Since cell key values are modulo with 4, the range of summation is between -3 and 3 in step 3. Most of the cells are applied all perturbation to cells randomly in step 4.

The re-Identification of the record corresponding to the cell with value 1 may take place based on any combination of age=17, male, and Eskimo shown in Figure 9 (b) ;However, attackers are not sure if the dataset is real values or perturbated values.

VI. FUTURE WORK.

The perturbation of zeros is a specific type of Cell-key perturbation used in data privacy and security to protect sensitive information.

If we perturb all rows, it damages data utility. Data privacy tries to preserve some risk groups. In our case, f=1. The effect of using the perturbation of zeros is that it adds random noise to the cells of a dataset that contain zero values. The perturbation of zeros is however out of scope in this paper. We will cope with these issues in our future work.

Cell-key perturbation is a technique that is widely used in Europe. Comparing performance with an existing algorithm is difficult because Cell-key perturbation is an alternative way to handle the differencing attacks. Therefore, we hope to make a performance comparison in future work.

VII. CONCLUSION

Cell-key perturbation data privacy procedure for the Security Operations Center Team is explained and implemented in this paper. In Figure 4, attackers cannot be so sure whether the information in the table are affected by perturbation or its original values. Some of cells are added with noise. perturbation tables (p-tables) are a goal of the Cell-key method. Jupyter Notebook tool and Python are implemented to compute perturbation tables.

The anonymizer server in Figure 2 was designed to allow users to do queries online. Consistency is so important for online service. The same user will get the same anonymized dataset all the time; therefore, the Cell-key perturbation method is a suitable answer for E-Commerce service. The server performs data preparation in section 4 and then runs Cell Key Perturbation in section 5.

REFERENCES

[1] D. Shahjee, N. Ware, "Integrated Network and Security Operation Center: A Systematic Analysis," *IEEE Access*, vol. 10, pp. 27881-27898, Mar 2022, doi: 10.1109/ACCESS.2022.3157738.

[2] H. Silva, T. Basso, R. Moraes, D. Elia, S. Fiore, "A Re-Identification Risk-Based Anonymization Framework for Data Analytics Platforms," in Conf. *2018 14th European Dependable Computing Conference (EDCC)*, Lasi, Nov 2018, pp. 101-106, doi: 10.1109/EDCC.2018.00026.

[3] T. Enderle and S. Giessing, “Implementation of a ‘p-table generator’ as separate R-package”, Deliverable D3.2 of Work Package 3 “Prototypical implementation of the cell key/seed method” within the Specific Grant Agreement “Open-Source tools for perturbative confidentiality methods”,(2018).

[4] Scotland’s Census, “Scotland’s Census 2022 Cell Key Perturbation October 2020,” National Records of Scotland, Oct 2020, Available: https://scotlandscensus.gov.uk/media/d1yn5gu3/pmp017-Cell-key-perturbation-emap-5940.pdf.

[5] Z. Hassanzadeh, R. Biddle, S. Marsen, "User Perception of Data Breaches," *IEEE Transactions on Professional Communication*, vol. 64, no. 4, pp. 374-389, Oct 2021, doi: 10.1109/TPC.2021.3110545.

[6] A. Mbombo, N. Cavus, "Smart University: A University In the Technological Age, TEM Journal, Oct 2021, pp. 13-17.

[7] Feher David Janos, Nguyen, Phuoc Dai. “ Security Concerns Towards Security Operations Centers,” 2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI), pp. 000273-000278, Aug 2018, doi: 10.1109/SACI.2018.8440963.

[8] K. Emam , E. Jonker, L. Arbuckle, B. Malin, “A Systematic Review of Re-Identification Attacks on Health Data,”, PLOS ONE 10(4): e0126772, Dec 2011, doi: 10.1371/journal.pone.0028071.

[9] L. Sweeney, “Uniqueness of simple demographics in the U.S. population,” Laboratory for International Data Privacy Working paper LIDAP-WP4, 2000.

[10] Khordadpour, Peyman, “The Security of Data Governance in the Digital World,” TechRxiv, Feb 2023, doi: 10.36227/techrxiv.22129160.v1.

[11] A. Narayanan, V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets,” 2008 IEEE Symposium on Security and Privacy (sp 2008), May 2008, doi: 10.1109/SP.2008.33.

[12] K. Benitez,B. Malin, “Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule,” Journal of the American Medical Informatics Association, vol. 17, no. 2, pp 169-177, doi: 10.1136/jamia.2009.000026.

[13] Meindl, B.: cellKey: Implementing ABS Cell-key method for adding noise to frequency and continous tables, 2020. R package version 0.19.1

[14] Bailie, J., & Chien, J. (2019). ABS Perturbation Methodology Through the Lens of Differential Privacy.

[15] Y. Ma, Y.X. Lin, J. Chipperfield, J. Newman, V. Leaver. “A new algorithm for protecting aggregate business microdata via a remote system,” in Conf. International Conference on Privacy in Statistical Databases, Dubrovnik, 2016, pp. 210-221.

[16] S. Prada, C. Gonzalez-Martinez, J. Borton, J. Fernandes-Huessy, C. Holden, E. Hair, T. Mulcahy, "Avoiding disclosure of individually identifiable health information: a literature review," SAGE Open, pp. 1-16, Dec 2010, doi: 10.1177/2158244011431279.

[17] K. Mivule, "Utilizing Noise Addition for Data Privacy, an Overview," in Conf. Information and Knowledge Engineering, IKE 2012, Las Vegas, USA, July 2012. pp. 65-71.

[18] Official Journal of the European Union, "Regulation (EU) 2016/679 of the European Paliament and of the Council," The European Union, [Online], Apr 27 2016, Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[19] Government Gazette, "Personal Data Protection Act, B.E. 2562 (2019)," Official Emblem of Royal Command, May 24 2019, Available: https://thainetizen.org/wp-content/uploads/2019/11/thailand-personal-data-protection-act-2019-en.pdf.

[20] I. Dove, C. Ntoumos, K. Spicer, "Protecting Census 2021 Origin-Destination Data Using a Combination of Cell-key Perturbation and Suppression," in Conf. International Conference on Privacy in Statistical Databases, Dubrovnik, 2018, pp. 43-55.

[21] B. Fraser, J. Wooton., "A proposed method for confidentialising tabular output to protect against differencing," Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality [Online], Nov 9-11 2005, Available: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2005/wp.35.e.pdf.