International Journal of Internet, Broadcasting and Communication Vol.13 No.2 67-75 (2021)
http://dx.doi.org/10.7236/IJIBC.2021.13.2.67

IJIBC 21-2-10

Zero-knowledge proof algorithm for Data Privacy

Youn-A Min

Professor, Applied Software Engineering, Hanyang Cyber University, Korea
yah0612@hycu.ac.kr

Abstract

As pass the three revised bills, the Personal Information Protection Act was revised to have a larger application for personal information. For an industrial development through an efficient and secure usage of personal information, there is a need to revise the existing anonymity processing method. This paper modifies the Zero Knowledge Proofs algorithm among the anonymity processing methods to modify the anonymity process calculations by taking into account the reliability of the used service company. More detail, the formula of ZKP (Zero Knowledge Proof) used by ZK-SNAKE is used to modify the personal information for pseudonymization processing. The core function of the proposed algorithm is the addition of user variables and adjustment of the difficulty level according to the reliability of the data user organization and the scope of use. Through Setup_p, the additional variable γ can be selectively applied according to the reliability of the user institution, and the degree of agreement of Witness is adjusted according to the reliability of the institution entered through Prove_p. The difficulty of the verification process is adjusted by considering the reliability of the institution entered through Verify_p. SimProve, a simulator, also refers to the scope of use and the reliability of the input authority. With this suggestion, it is possible to increase reliability and security of anonymity processing and distribution of personal information.

Keywords: *Blockchain, ZKP, Encryption, zk-SNARK*

1. Introduction

With the Industrial Revolution of 2020, the three laws of data about (Personal Information Protection Act, Credit Information Use and Protection Act, and the Information Communications Network Act) were implemented to enable pseudonymization when the processor of personal information has a purpose in research and public preservation of records [1]. Pseudonymized data refers to the unidentifiable data regarding a specific individual without any additional data by deleting or replacing part of the personal data [1].

Table 1 shows the utilization range disclosed by the policy wiki. As shown in the table, there is increasing interest regarding the unauthorized use and infringement of personal information due to the revised act stressing harder on multiple areas regarding usage of personal information without acquiescence.

Manuscript Received: February. 14, 2021 / Revised: February. 18, 2021 / Accepted: February. 24, 2021
Corresponding Author: yah0612@hycu.ac.kr
Tel: +82-2-2290-0872, Fax: +82-2-2290-0872
Professor, Applied Software Engineering, Hanyang Cyber University, Korea

Table 1. Summary of data bills [2]

Concept	Range of Usage
Personal Information: Information regarding a specific individual	Available for use within the range of consent, literal and specific
Pseudonimized Data: Information of a specific individual processed to not be accessible without additional information	Available without consent with the following motives : Statistics (Including commercial use) Research (Including industrial research) Public record preservation

For pseudonym information, the revised Article 28 of the Personal Information Protection Act excludes the application of personal information protection-related regulations such as notification of the collection source of personal information collected from other than the information subject, destruction of personal information, and notification of personal information leakage. Threats can arise [1-2,12].

This paper suggests an algorithm that includes company reliability within the calculation of Zero Knowledge Succinct Non-interactive Arguments of Knowledge, an encoding method, for a more secure management of pseudonymized personal information.

2. Related work

2.1 Data Pseudonymization Method

Methods of deleting or replacing the personal data, part or whole, is being used for pseudonymization[1]. Among the methods, the encryption methods include bidirectional, one-way, order-preserving, format-preserving, and homomorphic encryptions [1, 4].

Bidirectional encryption is a method that uses the same private key in encoding and decoding, giving a very fast calculation for the process but at the same time a risk for security threats to personal information once the key is leaked [5]. A one-way encryption includes application of hash function, which can be classified to Message Digest Code (MDC), Message Authentication Code (MAC), and Salt.

Order-preserving encryption is a method that maintains the code's order same as the original data, which gives the advantage of having the same order even after encoding. The format-preserving encryption has the same encoded value as the original information, which does not bring issues of schema modifications of storage. Homomorphic encryptions have the characteristic of enabling calculations in encrypted mode, allowing for various analysis.

The pseudonymization methods based on protocols for data transmitted through the network include Differential Privacy, Multi-party Computation, and Zero Knowledge Succinct Non-interactive Arguments of Knowledge (zk-SNARK).

Differential Privacy is information security technology used in Google and Apple that does not clearly show that Group B and Group C within a Database A does include personal information and mixes in appropriate noise regarding whether the information is included [7, 8]. Multi-Party Computation is based on Interactive Protocol that may use homomorphic encryption or a chosen calculation based on all the nodes maintaining their online state until the shared objective is complete[6, 7-8].

zk-SNARK is a technology in which a prover proves to the verifier that there is information in hold

without revealing the information itself. Recently, the method is being used frequently in data de-identification that requires a high level of anonymity with calculations that apply range proof methods that prove the range of concealed information [6, 7-8]. The methods mentioned above have security risks in terms of methodology, giving a need for a more secure pseudonymization.

2.2 zk-SNARK

Zero knowledge Succinct Non-interactive Arguments of Knowledge (zk-SNARK) is a pseudonymization method that proves the correctness of a data without revealing any of the private information [9, 10]. zk-SNARK uses a hash function that makes original inference and decryption impossible and is used as a cryptological method for data pseudonymization in blockchain [10, 11]. Figure 1 is the process of zk-SNARK in a blockchain.

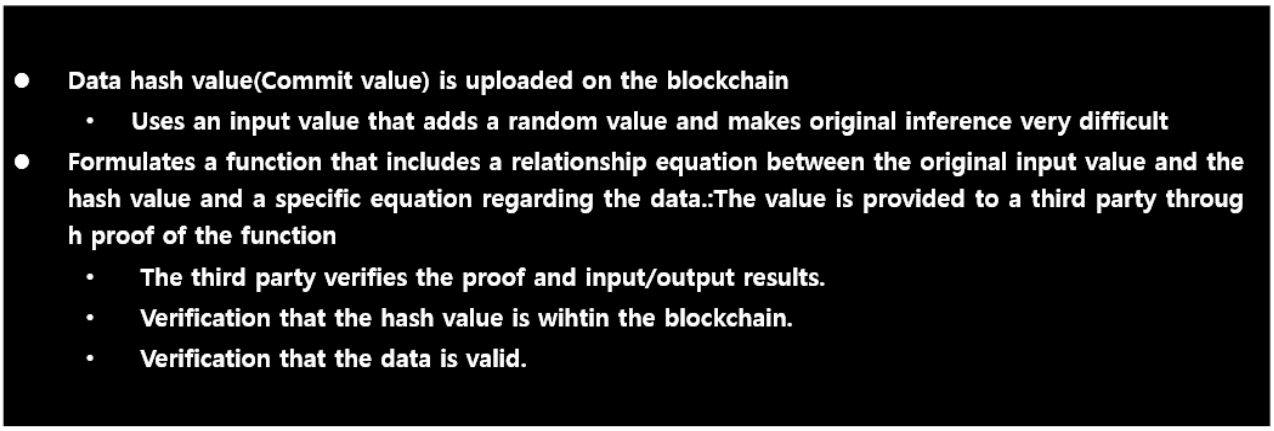

Figure 1. zk-SNARK Process

zk-SNARK is an example of Zero Knowledge Proof (ZKP) used in blockchain, a distributed data sharing environment. zk-SNARK uses ZKP to formulate a circuit that forms a certain function into a single multiplication formula and multiple addition formulae[9]. The hash value given from the ZKP conceals the original data, solving the security issues regarding personal information, and also enables very fast verification of proofs, increasing the function of the blockchain.

3. zk-SNARK-based Personal Information Pseudonimization Algorithm

3.1 Comparison of Data Encryption Methods

With the enforcement of the three laws of data, the utilization range of personal information is broadened and gives need for a verified and secure pseudonymization method.

This paper uses and modifies the ZKP formula used in ZK-SNAKE for personal data pseudonymization. The ZK-SNAKE algorithm of the formula is shown as VC (Verifiable Credentials).

The suggested algorithm enables an addition of user variable and control of difficulty level according to the credibility of used service company and usage range. Through Setup_re, the added variable γ can be selectively applied according to the company credibility. The agreement level of the witness can be controlled according to the credibility of input company through Prove_re. Verify_re takes into account the input credibility to enable control of the verification difficulty. The simulator SimProve also takes the utilization range and input credibility into account.

3.2 Modified Algorithm of zk-SNARK

Suggested environment is Non-interactive ZKP, assuming Relation T that includes the node information for formula application.

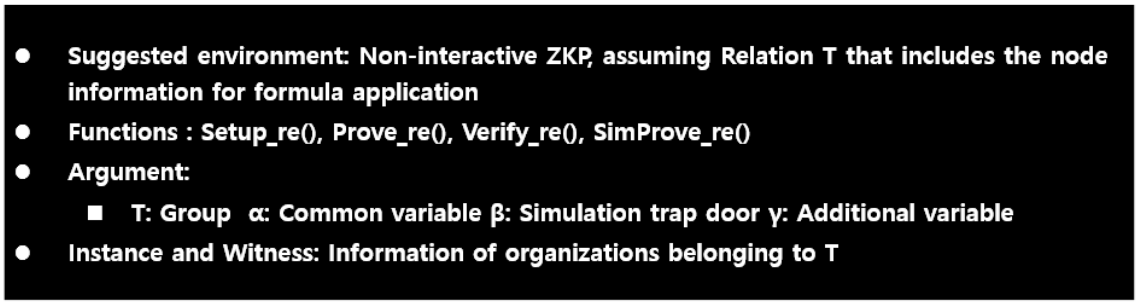

Figure 2. Functions and Factors about Propose Algorithm

The explanation of each used functions and factors are as follows

Table 2. Process for each Function of the Proposal

● *Setup_p(T)->(α,β)* : Output common variable α and simulation trap door β through Relation *T*. Selectively apply additional variable γ, which contains measurement of used company credibility, when using *Setup_p* function in order to efficiently control the speed of identification and verification.
● *Prove_p(α,Instance, Witness)->P:* Output value of *Prove_p* by inputting common variable α and the instance and witness within *T* relative to Relation *T*. The witness agreement level is controlled according to the input credibility. The difficulty of agreement can be controlled as such that in case of high credibility, the difficulty is lowered while questionable credibility increases difficulty.
● *Verify_p(α, Instance, P)->True or False :* The validity of the proof is output as *True or False* according to input of common variable α, Instance, and *P*. The verification difficulty is controlled according to the company credibility input as Instance. The process selectively applies additional variable γ to modify the equation according to the DID (Decentralized identifier) utilization range and importance of verifier credential.
● *SimProve_p(α,β,Instance)->P :* The simulator SimProve inputs common variable α, trapdoor β, and Instance to output *P*. The verification difficulty is controlled by taking the credibility Instance into account.

The process pseudo-code of Algorithm is as follows.

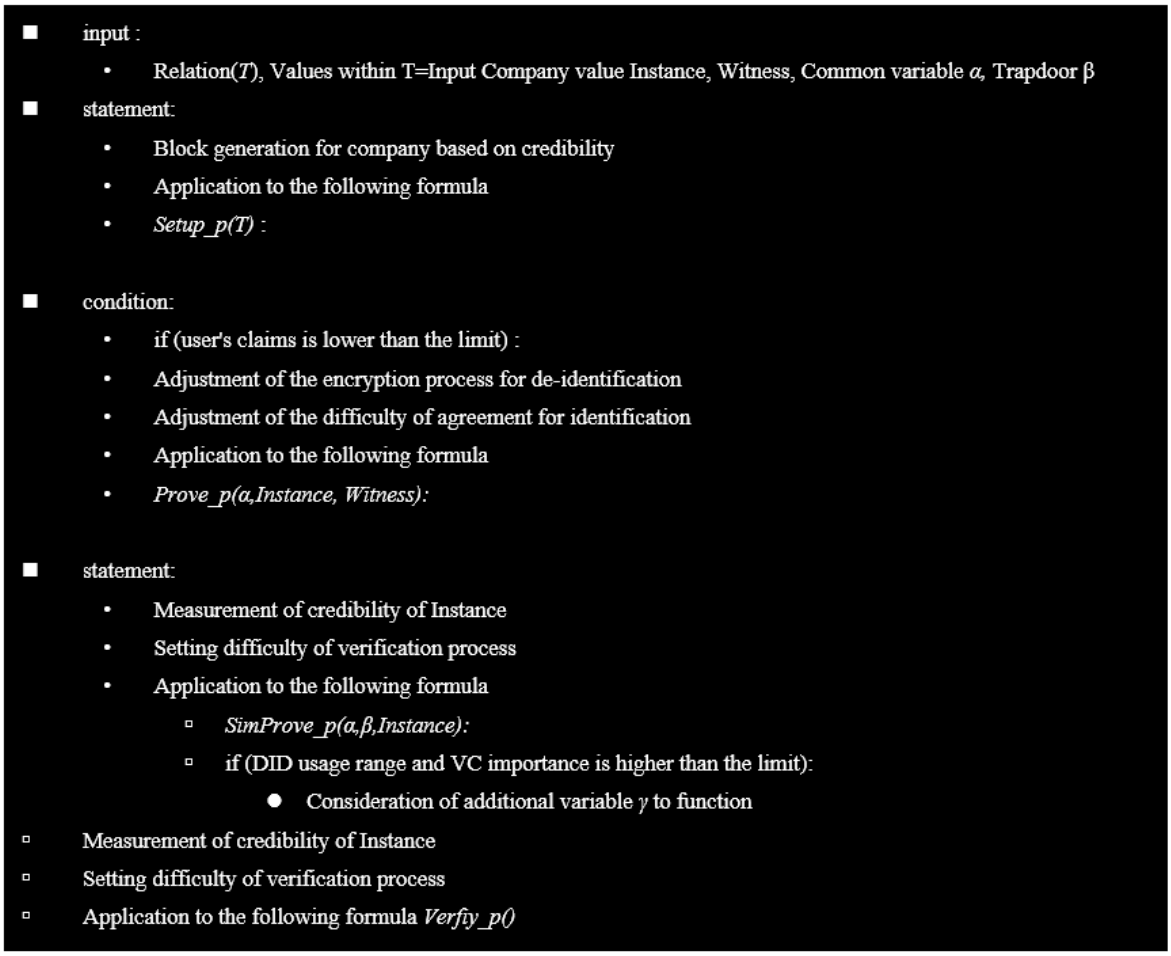

Figure 3. Pseudo-code of Propose Algorithm

3.3 Performance evaluation

In order to evaluate the performance of the algorithm proposed in this paper, the verification time of de-identification of pseudonym data was measured for 25 randomly generated datasets in a form similar to personal information such as ID (identifier), name, and address. Experimental environment and performance the evaluation environment is as follows.

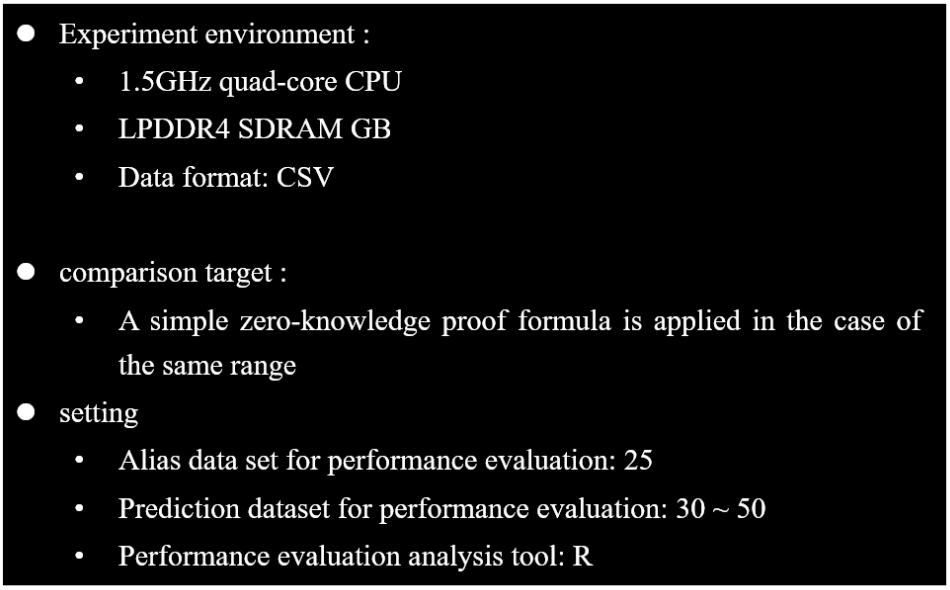

Figure 4. Evaluation Environment

Five tries were repeated in identical environments based on Figure 4, and the result measurement time data are organized as shown in Table 3.

Table 3. Verification time of Sample data

[unit : data set/ms]

Data Set	1	3	5	8	10
Suggested formula	691	751	788	820	998
Comparison formula	729	769	831	952	1071

The experiment above used a small number of data in order to eliminate error from overloading of the experiment environment, and future performance development could be inferred by organizing the values into data through R.

```
# summary of formulas and predictions about "ZKP(A)"
Call:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
x1 27.556 3.113 8.532 0.00331 **
Multiple R-squared: 0.9604,
Adjusted R-squared: 0.9472
F-statistic: 72.79 on 1 and 3 DF, p-value: 0.003383

# summary of formulas and predictions about "Propose ZKP(B)"
Call:
Coefficients:
Estimate Std. Error t value Pr(>|t|)
x2 31.106 1.551 19.89 0.000278 ***
Multiple R-squared: 0.9925,
Adjusted R-squared: 0.99
F-statistic: 395.5 on 1 and 3 DF, p-value: 0.0002778
```

Figure 5. Summary of formulas and predictions code

The difference in algorithm processing between ZKP(A) and ZKP(B) proposed in this paper is that when applying the algorithm for data pseudonymization, the reliability and frequency of use of data users are considered. According to the formulae, the formula of ZKP before modification $y \fallingdotseq 23.1x+642$ and suggested formula $y \fallingdotseq 32.15x+691$ were conjectured and the performance of the data set regarding the formulae is organized in Table 4.

Table 4. Formula about Algorithm

[unit : data set/ms]

Data set /ms	20	30	40
Suggested formula $y \fallingdotseq 27.1x+642$	1152.112	1411.912	1799.351
Comparison formula $y \fallingdotseq 31.15x+691$	1321.391	1640.001	1971.887

The performance increase can be assumed regarding the performance evaluation of the suggested formula and comparison formula.

Table 5. Performance about various data set

Data set	1	3	5	8	10	20	30	40
Performance improvement	8.2%	8.5%	8.9%	9.5%	10.2%	10.5%	11%	11.5%

In the case of the suggested formula in Table 9, there was a 10% increase in performance compared to the existing formula, and showed an increase in performance as the data set grew larger.

In order to measure the meaningfulness of the algorithm function of Table 5, we can hypothesize as the following in Figure 6.

Null hypothesis: μ=0 (the prediction is unreliable)
Alternative hypothesis: μ>0 (There is confidence in the prediction)

Figure 6. Hypothesize about Performance test

According to the performance test t values of the suggested formula and compared formula were 8.532 and 19.89, respectively, and the p values at 0.05 significance level were 0.00338 and 0.000278, respectively. Through this analysis, the null hypothesis was rejected and the suggestion was proven to have significance as a formula capable of measurement and prediction.

4. Discussion

With the enforcement of the 3 major data acts, the regulations regarding personal information were revised and the range of usage of personal information was expanded. This paper modified and suggested the zk-SNARK as a method of pseudonymization of personal information. When accessing personal information through various organizations, the credibility of the organizations and the argument factor of the utilization range were taken into consideration to be implemented into the process. The verification time was measured for algorithms before and after modification as a performance evaluation, in which the performance was shown to have increased by at least 10% on average in the significant range.

References

[1] S.O. Kim, "Balance points for safe processing and rational use of pseudonym information-Combined with constitutional evaluation of the 3rd Data act ," Korea Public Law Research. Vol. 49, No. 2, pp.371-407, Dec 2020.
DOI : 10.38176/PublicLaw.2020.12.49.2.371

[2] Y.B. Lee," A Study on the Revision Trend of Data 3 Act," Korean Society for Comparative History, Vol.27, No 89, pp.423-463, May, 2020.

[3] H.J. Chun, H.J. Yi, Y.K, Kim, Dongrae, "Data Quality Measurement on a De-identified Data Set Based on Statistical ", The Journal of the Korea Contents Association, Vol. 19, No. 5, pp.553-561, May 2019.
DOI : http:// doi.org/10.5392/JKCA.2019.19.05.553

[4] H.C. Yang, Y.J.Lee, S.G. Kim, "Effects of Application Level of Personal Information De-identification Technology on Intention to Use Big Data", Journal of the EA Society of Korea, Vol. 13, No. 3, pp. 395-

404.Sep 2016.
UCI : I410-ECN-0102-2017-560-000523124

[5] S.H. Kim and S.H. Jeon, “Big data integration using data de-identification”, Journal of the Korean Intelligent Systems Society, Vol 29 No 3, pp. 235-241, Jun 2019.
DOI : 10.5391/JKIIS.2019.29.3.235

[6] ISO/IEC 20889, Privacy enhancing data de-identification technology and classification of techniques, pp.1-50, 2018.

[7] The Ministry of Government Administration and Home Affairs, Guideline for non-identification measures for personal information-Standards for non-identification measures and support and management system guidance, pp.10-89, 2016.

[8] Zhang. Zeyu, Lu. Zhiyang, Tian. Youliang, "Data Privacy Quantification and De-identification Model Based on Information Theory," in Proc. International Conference on Networking and Network Applications (NaNA) International Conference on, pp.213-222, Oct.10-13, 2019.
DOI:10.1109/NaNA.2019.00046

[9] H.C.Yang., The effects of applying personal information de-identification technology on intention to use big data, Ph.D. Thesis. Gwang-UnUniversity, Seoul, Korea., 2016.
UCI : I410-ECN-0102-2017-560-000523124

[10] J.H. Lee et al., "Personal Information Management System with Blockchain Using zk-SNARK", The Journal of the Society for Information Security, Vol. 29, No. 2, pp. 299-308, April 2019.
DOI : 10.13089/JKIISC.2019.29.2.299

[11] Y.H.Kang, "A Study on an Enhancement Scheme of Privacy and Anonymity through Convergence of Security Mechanisms in Blockchain Environments," The Journal of the Korean Convergence Society, Vol. 9, Issue 11, pp. 75-81, Dec 2018.
DOI:10.1109/3ICT51146.2020.9312014

[12] H.K.et al., "Big Data Analysis System Based on Public Data", Journal of the Institute of Internet, Broadcasting and Communication (JIIBC), Vol.20, No.5, pp.195-205, Oct 2020.
DOI: https://doi.org/10.7236/JIIBC.2020.20.5.195