

An Enhanced Differential Homomorphic Model using N-Prime Scheme for Privacy Preservation

*¹Hafiz O. Sanni, ¹Rafiu M. Isiaka, ¹Akinbowale N. Babatunde, and ²Muhammed K. Jimoh

¹Department of Computer Science, Kwara State University, Malete, Nigeria

²Department of Computer Science, University of Ilorin, Ilorin, Nigeria

{sannihafiz|jmkclarularu}@gmail.com | {abdulrafiu.isiaka|akinbowale.babatunde}@kwasu.edu.ng

Received: 02-FEB-2023; Reviewed: 15-MAR-2023; Accepted: 20-MAR-2023

<http://doi.org/10.46792/fuoyejt.v8i1.999>

ORIGINAL RESEARCH

Abstract- In preserving individual privacy in data publishing, several efforts have been made by scholars globally to develop an individual privacy-preserving model and hybridized models which harness the strength of the individual model to increase privacy preservation in data Publishing (PPDP). The Differential homomorphic model (DHM) was among the hybridized models developed that combine differential and homomorphic models. Though is one of the state-of-the-art hybridization methods for privacy preservation because of the Differential model and Homomorphic model strengths of the two hybridized models which are the ability to prevent composition problems and database attacks respectively. However, applying this model is challenging because of the high computational complexities due to the modular exponentiation problem available in the Pailler encryption scheme used in DHM. In this research, an N-PRIME homomorphic encryption scheme was proposed to replace the Pailler encryption scheme in the differential homomorphic model (DHM). The designed model was 51% faster than the existing model (Differential Homomorphic Model) in terms of computation time and 48.5% faster when generating the graphical data set, though the designed model consumed 4% more storage space than the existing model.

Keywords- Homomorphic, Differential, Privacy Preservation, Database Security, N-prime, Hybridized Model, Database Attack.

1 INTRODUCTION

Information and communication technology (ICT) creates a vast amount of data characterized by velocity, volume and veracity. Data are collected by government agencies and private organizations directly or indirectly from the users. Most times such data are published for public consumption for statistics, mining among other uses. Published data usually contain personal information which may be sensitive. Leakage of such sensitive information violates individual privacy.

A privacy policy is a statement or a legal document that discloses some or all of the ways a party gathers, uses, discloses, and manages a customer or client's data (Connor, 2014). It is a legal requirement to protect a customer or client's privacy. Personal information can be anything that can be used to identify an individual, not limited to the person's name, address, date of birth, marital status, contact information, ID issue, and expiry date, financial records, credit information, medical history, where one travels, and intentions to acquire goods and services. The primary goal of privacy preservation is to hide sensitive data before it gets published (Chatterjee & Sengupta, 2015). For example, a hospital may release patient records to enable researchers to study the characteristics of various diseases. The raw data contains some sensitive information about individuals treated over time in the hospital.

In other to preserve the privacy of an individual, Data scientists have developed many individuals and hybridized techniques to prevent privacy right (Connor, 2014). The two major challenges in privacy preservation are composition and database attacks. This work considers the computation strength ability of the Nprime Model which was introduced to reduce the computational complexity of the differential homomorphic model which is the existing model as shown in the result section of this paper. It also gives better data security by generating long cipher text than the existing model.

The beneficiaries of this research are institutions that possessed individual records in their system and want to publish the record online for purpose of research to derive new knowledge and still want to maintain individual privacy to a high extent. For instance, medical centres will be able to publish the patient record online for the purpose of research without violating their patient privacy right even the National identity management system can also publish records of citizens online for the purpose of research without violating individual rights and the quality of data is being maintained. All other institutions that possessed individual records within their system can also publish such records online for purpose of research without violating individual privacy right.

The major security challenges in privacy preserving are composition problem and database attack.

1.1 COMPOSITION PROBLEM

This is when the suppressed data was detected as a result of having background knowledge of the data set. For example, If the dataset of a particular domain (Kings Groups of school) dataset as follows (name, email, age, skin colour, disease) was published and private data was

*Corresponding Author

Section B- ELECTRICAL/COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:

Sanni H.O., Isiaka R.M., Babatunde A.N., and Jimoh M.K. (2023). An Enhanced Differential Homomorphic Model using N-Prime Scheme for Privacy Preservation, FUOYE Journal of Engineering and Technology (FUOYEJET), 8(1), 23-27. <http://doi.org/10.46792/fuoyejt.v8i1.999>

suppressed. Even though the private data has already been suppressed name and email if someone runs a query like “people with white skin colour disease” and it returns one record like (age=17, Skin Colour=white, disease=Gonorrhoea). If there is someone who knows that in the king Group of schools, there is only one person with white skin colour, he can deduce the white skin colour person he knows in Kings Group of School has Gonorrhoea. This means the individual privacy of that person has been violated.

1.2 DATABASE ATTACK

This is when an unauthorized user force gains access to the database and views all the suppressed data. Data may be suppressed for researchers from the database but what if the unauthorized user forcefully gains privileges to the database backend that means the original dataset can be viewed and individual privacy will be compromised.

The related works are shown in Section 2 and also, the application area of the designed model is shown in the same Section 2. The methodology and developed model are shown in Sections 3 and the results of the performance evaluation is shown in Section 4 and Section 5 provides the conclusion, summary and future research.

2 RELATED WORKS

The importance of privacy preservation in information security necessitates the efforts made by scholars globally on individual and hybridized techniques. All individual techniques have limited strength. Among the efforts made by scholars globally on individual techniques include; anonymization, suppression, generalization and slicing. Others are fuzzy logic, differential privacy and homomorphic encryption (Ahmad et al., 2014). Some Hybridized techniques include combining Fuzzy membership function and random rotation, combining suppression and perturbation. Other Hybridized approach is combining differential privacy and homomorphic encryption

Anonymization is the first technique introduced to preserve the privacy of individuals in data publishing (Majeed & Lee, 2021). It is an approach for preserving the privacy of published data by hiding the sensitive data and identity of record owners by suppressing or removing the identifying values. Ahmad et al. (2014) stated that removing the key identifier is not the best way to prevent privacy preservation (Mansour et al., 2021). National data was published in the year 2005 without a primary identifier to the public (Mansour et al., 2021), that is, all identifying attributes have been replaced with asterisk but this data was combined with other public data to identify the suppressed values using a quasi-identifier, in other to improve on this.

Another approach was introduced by Fletcher and Islam (2014) called generalization. It replaces attribute values with semantically unvarying but less particular values. Due to this replacement, many records have the same quasi values. Generalization replaces exact values with a more general description to hide the details of attributes, making the quasi values less identifying. If the value is

numeric, it may be changed to a range of values. For example, the age attribute value of 45 can be changed to a range of 40- 60 but all those approaches however reduce the quality of data (Jiang, 2018). To deal with problems that occur in generalization and anonymization, slicing was introduced. Slicing partition, the data set is both vertically and horizontally. By organizing attributes into columns based on how they relate to one another, vertical partitioning is accomplished. Each column contains a subset of attributes that are highly correlated.

Horizontal partitioning is done by grouping tuples into buckets. Then, within each bucket, values in each column are randomly permuted (or sorted) to break the linking between different columns. The limitation of this technique is that when there are repeating sensitive attributes probability of membership disclosure is high (Jiang, 2018).

Furthermore, the fuzzy membership function was also introduced to preserve the privacy of an individual (Ahmad et al., 2014). It removes all the sensitive attributes before publication. In this approach, a fuzzy membership function is applied to the original data to hide the sensitive data which will generate the sanitized data with fuzzification methods. The problem of this approach is the composition problem, any member of the public that has prior knowledge of published data can be identifying the data owner by mapping (Xie & Wang, 2016). To reduce the risk of identifying the data owner, Differential Privacy was introduced by adding noise to the output. This privacy-preserving data publishing model (Differential privacy) was proposed by Dyda et al. (2016). This model also follows that even if several datasets are joined there will be no disclosure risk. The Limitation of this approach is that it does not consider database attacks by not removing the sensitive data from the database (Dyda et al., 2016).

As an improvement on previous individual techniques to prevent database attacks a new approach was introduced. In this approach, all the databases were encrypted in case of a database attack (Chatterjee & Sengupta, 2015). The first homomorphic encryption introduced was Paillier homomorphic encryption which prevents database attacks by encrypting all the data in the database. The advantage of the approach is that only an addition operator can be conducted over encrypted data without decrypting the data (Kumar & Srivastava, 2019). Another Homomorphic encryption scheme introduced to preserve privacy is RSA (Rivest, Shamir and Adleman) (Mansour et al., 2021). It also prevents database attacks by encrypting all the data in the database. The advantage here is that only multiplication manipulation can be conducted over encrypted data without decrypting the data. The problem of the two popular homomorphic encryption schemes used for data privacy (Paillier and RSA) are composition problems (Yichen et al., 2021).

The limitations of the individual privacy techniques necessitated the introduction of hybridized method. Hybridization approaches were introduced to harvest the individual privacy techniques. Examples include

combining fuzzy membership function and random rotation perturbation (RRP) (Vo-Van et al., 2020), the problem with this hybridized approach is distorted data that produce bad output (Salim et al., 2021). Another approach is suppression and perturbation which is a hybridization technique introduced (Mohammed et al., 2019). but it leads to a loss of information drastically due to some sensitive values being suppressed. Among the other efforts is the combination of Differential privacy with homomorphic encryption as proposed by Mohammed et al. (2019), the data in the database were encrypted using Paillier homomorphic encryption, the strength is the ability to take care of the database attack and composition problem but the limitation of this approach is high complexity and processing time as a result of modular exponentiation of Paillier homomorphic encryption.

NPrime was proposed as a homomorphic encryption scheme with high computation strength than the paillier scheme. This work considers the computation strength ability of the Nprime Model which was introduced to reduce the computational complexity of the differential homomorphic model (Yichen et al., 2018) as shown in the result section of this paper. It also gives better data security by generating long cypher text but the system trades off the storage space for better security and fast computation time.

3 ADAPTATION OF DIFFERENTIAL HOMOMORPHIC MODEL

The differential homomorphic model (Yichen et al., 2018) was adapted. It comprises five (5) layers; Data Entry layer, Encryption layer, Data Request Layer, differential layer and decryption layer as the adaptation of the model is shown in Figure 1. The Enhancement part in the model is

layer two which is the encryption layer where Nprime was used in place of the Paillier Encryption Scheme in the Differential homomorphic model (DHM). There is a noticeable drawback in the existing system. In other to improve the existing system by providing better privacy preservation and speeding up the computational speed, the Nprime encryption model was introduced. Since Prime works on Numeric numbers, all the characters needed to be converted into numeric using equivalent ASCII values. The system architecture of the adapted model is shown in Figure 1.

The first layer of the scheme is the data entry layer, this is the layer where all the dataset is loaded into the model. The second layer is the Encryption layer, in this layer, all the data in the dataset were encrypted using Nprime encryption Scheme as compared to the existing model using pallier. The third layer is the data request layer, this is the layer where the user can request data in the dataset and before data will be released to the user, noise is added to the data to avoid composition problems. Lastly, the decryption layer which is the fourth layer is the layer where the data can be decrypted back to the original data using the decryption key.

3.1 DATA ENTRY LAYER

This is the first layer where the model will read the dataset and all characters in the dataset will be converted to numeric.

Mathematical Representation of layer 1

$$\text{Data} = [\text{character}, \text{Ascii}]$$

$$\text{Data} = \text{encode}(\text{Data})$$

Interpretation:

The equation (1). Matching characters with equivalent ASCII keys.

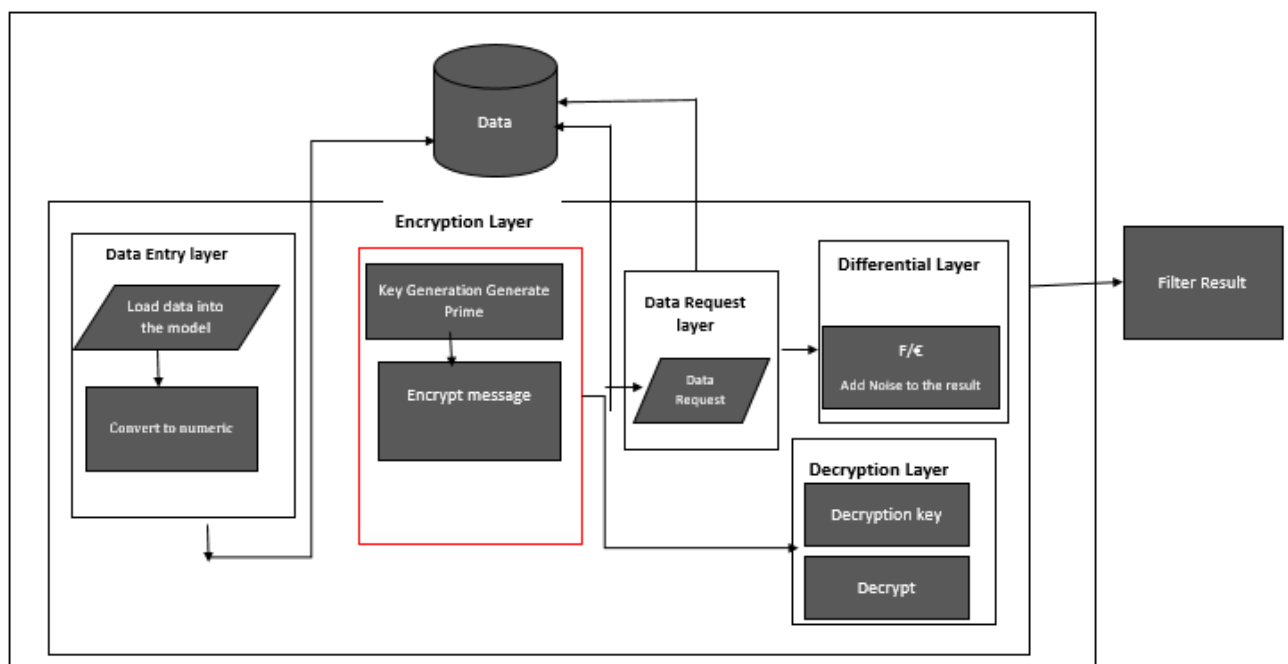


Fig. 1: Developed Model (Enhanced differential homomorphic model)

3.2 ENCRYPTION LAYER

Differential Homomorphic Model is the existing system that is using Pailler as encryption scheme. The equation for Paillier encryption is shown in equation 1.

$$C = g^m \cdot r^n \bmod n^2 \quad (1)$$

Where C is the cipher text (encrypted data), m is the message and (g, n) is the public key and r is the random number. Condition to select g and n; where (n, g) are prime numbers and larger numbers respectively.

The greatest common divisor of r and n = 1 as below:
 $\text{GCD}(r, n) = 1$

Condition to select r, where $\text{GCD}(r, n) = 1$

Equation (1) is the reason for the enhancement because of (r^n) where n is a large number, this leads to higher process time and higher computation time because of higher modular exponentiation in the pailler encryption scheme as shown in equation (1).

3.3 ENHANCEMENT OF THE ENCRYPTION LAYER

Nprime will be used to replace Paillier in the existing model so as to reduce the computation time.

Nprime encryption scheme

Algorithm for Key Generation

Step 1: Key generation

Generate $l \in \mathbb{Z}_n$

Generate l, where l is a big prime integer value

Generate n, where n is a multiplication of various prime numbers

Step 2: Encryption algorithm

$$ct = m + r \cdot i \cdot n \quad (2)$$

Where $m \in [0, L-1]$ and $n = p_1 \cdot p_2 \cdot \dots \cdot p_i$; ct is the cipher text. The encryption scheme will be faster because of no presence of modular exponentiation and it generated more encrypted text length. And m is the message, ct is the cipher text and I is the big prime value. This layer will generate an encryption text.

3.4 DATA REQUEST LAYER (THIRD LAYER)

This layer will accept a query from the researcher, and the researcher will request for types of data he/she want to fetch from the database.

Algorithm for data search

Int arr [] = {data, n}

Data count = findPair(count(int arr[]))

This layer search for the requested data from the encrypted database.

3.5 DIFFERENTIAL LAYER (FOURTH LAYER)

After the user has searched query to search for particular data or information, the count result will be distorted by adding minimal noise using differential privacy and the search result will be released.

$$DT = \text{int} \frac{(\text{dataCount})}{D} \quad (3)$$

where D is the frequency and DT is generated distorted data from the query searched which will be released to the data recipient.

This layer adds minimal noise to the data search.

3.6 DECRYPTION REQUEST LAYER (FIFTH LAYER)

Data can be decrypted back by an authorized person i.e.; the encrypted data can be converted back to the original data using generated decryption key

$$m = ct \bmod i \quad (4)$$

Where m is the message, ct is the cipher text and I is the big prime value generated.

4 RESULTS AND DISCUSSION

4.1 PERFORMANCE EVALUATION

The performance analysis is done by comparing the previous differential homomorphic model and the improved differential homomorphic model with N-Prime in terms of 2 key cryptanalysis metrics (Time and space).

4.1 ENCRYPTION & DECRYPTION TIME REPRESENTATION

Table 1 shows the Encryption time process (ms) and Decryption time process of the two Algorithms with their respective homomorphic schemes with fixed data size and varied key length. Table 1 represents the four different sizes of keys and the corresponding Decryption time taken by the Differential Homomorphic Model (DHM) and the Developed Model. By analysing Table 1 it was concluded that the time taken by the developed Model has a smaller Encryption time. It was also shown in Table 1 that the developed model was 67% faster than the present model (DHM) and the decryption process of the developed model is 51% faster than the present model.

Table 1. Encryption time and Decryption Time process

Type of Scheme Key	Size (Bit)	Develop Model (sec)	Differential Homomorphic model (sec)
Encryption Process	256	1.13	2.3
	512	2.2	4.33
	1024	2.75	5.31
	2048	3.5	5.71
Decryption Process	256	0.34	0.39
	512	0.67	0.9
	1024	0.84	1.22
	2048	0.91	2.22

Table 1 shows that the developed model has lower computational time compared to the differential homomorphic model (DHM). The encryption time has given us a linear equation which means the higher the bit size the higher the time taken. As shown in Table 1, Figure 2 shows that the developed model line is below the existing model which implies that the developed model is faster.

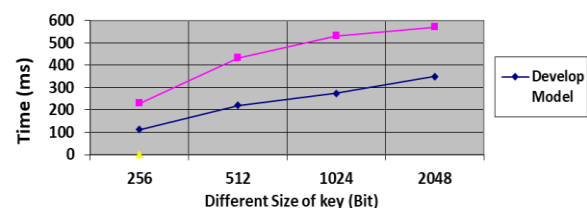


Fig. 2: Encryption Time Representation

4.2 ENCRYPTED STORAGE SPACE

The storage space of the encrypted file consumes in the database by the developed model and existing model using different text sizes (6KB, 20KB and 2.3MB). The encrypted space taken by both models is shown in Table 2. When a 6KB text file size was used it took the developed model's 128KB storage space size and the existing model's 100KB storage space. For 20KB developed, the model consumes more space than the existing model, the developed model consumes 390KB of storage space while the existing model consumes 210KB, also when a higher file size which is 2.3MB developed model consumed 46MB while the existing model consumed 39MB. It shows that the developed model is consuming more storage than the existing model.

Both models also provide a linear equation graph which implies that the higher the file size the higher the encrypted storage. The existing model consumes lesser space than the developed model, nevertheless the higher the text strength the better the security. For example, to encrypt 100 with the previous model it will generate less cipher text as shown below $188^{100.97^{187}} \bmod 34969 = 26118$ (small cipher text) and to generate cipher text of 100 using Nprime it produces high cipher text $100 + 67291 * 56771 * 67291 = 257063554799051$. The developed model trades away the storage space for better security and fast encryption time; see Table 2.

Table 2. Encryption Storage Space

Type of Algorithm & Scheme	Original Size		
	6KB	20KB	2.3MB
Developed Model	128kB	390KB	46MB
Differential Homomorphic model	100KB	210KB	39MB

Table 2 shows that the encrypted storage space increase as data size increases. The developed model Consume more storage space than the existing model

5 CONCLUSION

The primary aim of privacy preservation is to produce quality data for researchers and maintained individual privacy by cratering composition problems and database attacks. Efforts have been made by scholars globally on individual and hybridized techniques, based on the most recent hybridized model, the differential homomorphic model was adapted with N-prime for better security features and less computational speed which trade-off storage space. In the future, the research can be concentrated on how to reduce the cost of implementation of the model by reducing the storage space size without losing the security function and computation speed.

REFERENCES

Ahmad, S. M. T., Haque, S., & Faizanut Tauhid, S. M. (2014). A Details Study of Data Transformation for Privacy Preserving in Data Mining. *International Journal of Scientific & Engineering Research*, 5(1). Retrieved from <http://www.ijser.org>

Chatterjee, A., & Sengupta, I. (2015). Searching and Sorting of Fully Homomorphic Encrypted Data on Cloud. *IACR Cryptology EPrint Archive*, 1–14. <https://doi.org/10.1109/TCC.2015.2481416>

Connor, P. O. (2014). *An International Comparison of Approaches to Online Privacy Protection an International Comparison of Approaches to Online Privacy Protection Peter O' Connor*. (May). <https://doi.org/10.1007/3-211-27283-6>

Costante, Elisa; Sun, Yuanhao; Petković, Milan; Hartog, Jerry (2012). "A machine learning solution to assess privacy policy completeness". *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society - WPES '12*: 91.

Dyda, A., Purcell, M., Curtis, S., Field, E., Pillai, P., Ricardo, K., ... Lau, C. L. (2021). Differential privacy for public health data: An innovative tool to optimize information sharing while protecting data confidentiality. *Patterns*, 2(12), 100366. <https://doi.org/10.1016/j.patter.2021.100366>

Fletcher, S., & Islam, M. Z. (2014). Measuring Information Quality for Privacy Preserving Data Mining. *International Journal of Computer Theory and Engineering*, 7(1), 21–28. <https://doi.org/10.7763/ijcte.2015.v7.924>

Jiang, Z. (2018). T-Closeness Slicing: A New Privacy Preserving Approach for Transactional Data T- Closeness Slicing: A New Privacy Preserving Approach for Transactional Data Publishing. (January). <https://doi.org/10.1287/ijoc.2017.0791>

Kumar, V., and Srivastava, N. (2019). Chinese Remainder Theorem based Fully Homomorphic Encryption over Integers. *International Journal of Computer Applications* 14(2), 203–208.

Kumar, Vinod, Kumar and Rajendra. (2018). Fully Homomorphic Encryption Scheme with Probabilistic Encryption Based on Euler's Theorem and Application in Cloud Computing. 10.1007/978-981-10-6620-7_58.

Liu, Qinju, Lu, Xianhui, Luo, Fucui, Zhou, Shuai, He, Jingnan, Wang and Kunpeng. (2020). SecureBP from Homomorphic Encryption. *Security and Communication Networks*. 2020. 1-9.

Majeed, A., & Lee, S. (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. In *IEEE Access* (Vol. 9). <https://doi.org/10.1109/ACCESS.2020.3045700>

Mansour, H. O., Siraj, M. M., Ghaleb, F. A., Saeed, F., Alkhamash, E. H., & Maarof, M. A. (2021). Quasi-Identifier Recognition Algorithm for Privacy Preservation of Cloud Data Based on Risk Reidentification. *Wireless Communications and Mobile Computing*, 2021. <https://doi.org/10.1155/2021/7154705>

Mohammed, Mohammed Anwar and Fadhil Salman Abed (2019). "An improved Fully Homomorphic Encryption model based on N-Primes." *Kurdistan Journal of Applied Research*

Salim, Mikail, Kim, Inyeung, Doniyor, Umarov, Lee, Changhoon and Park. (2021). Homomorphic Encryption Based Privacy-Preservation for IoMT. *Applied Sciences*. 11. 8757.

Saranya, N., Karpagam, M., & Muruganandham, N. (2017). Hybrid approach for data publishing using Privacy preservation techniques. *ARPN Journal of Engineering and Applied Sciences*, 12(1), 155–159.

Vo-Van, Tai, Nguyen-Hai, A., Tat-Hong, M., Nguyen-Trang and Thao. (2020). A New Clustering Algorithm and Its Application in Assessing the Quality of Underground Water. *Scientific Programming*. 2020. 1-12. 10.1155/2020/6458576.

Xie Q, Wang L (2016). Privacy-Preserving Location-Based Service Scheme for Mobile Sensing Data. *Sensors* (Basel). Nov 25;16(12):1993. Doi: 10.3390/s16121993.

Yichen, J., Wang, C., Wu, Z., Du, X., & Wang, S. (n.d.). *Privacy-preserving biomedical data dissemination via a hybrid approach*. *AMIA Annu Symp Proc*. 2018 Dec 5;2018:1176-1185. PMID: 30815160; PMCID: PMC6371369.

Zhan, J., Hsieh, C. L., Wang, I. C., Hsu, T. S., Liao, C. J., and Wang, D. W. (2010). Privacy-preserving collaborative recommender systems. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 40(4), 472–476. <https://doi.org/10.1109/TSMCC.2010.2040275>