Preserving data privacy in data market and obtain truthful data

Miss. Aishwarya Pratap Jadhav
Department of Computer Engineering
Amrutvahini College of Engineering

Prof. M. A. Wakchaure
Department of Computer Engineering,
Amrutvahini College of Engineering

Abstract :

Truth is most often used to mean being in accord with reality. It is very important that data which are collected from different user is real data/truthful data. There are several fields were present from which the data is get emerged for the purpose of society's need. Ensuring data truthfulness and protecting the privacies of data contributors are both important to the long term healthy development of data markets. On one hand, the ultimate goal of the service provider in a data market is to maximize their profit. Yet, to reduce operation cost, a strategic service provider may provide data services based on the whole raw data set, or even return a fake result without processing the data from designated data sources. In here, TPDM is been introduced, that with efficiency integrates honesties and privacy preservation in knowledge market. TPDM is structured internally in associate degree homomorphism encryption-decryption mechanism, using partially homomorphic encryption and identity-based signature for providing the security. Also at the same the mechanism time facilitates the batch verification, data processing, and outcome verification, while maintaining identity preservation and data confidentiality.

Keywords: Truthful and untruthful data, TPDM, Data confidentiality, Homomorphic encryption

Introduction:

Now a day, large numbers of users were present on social media platform. Everyday so many users were contribute their data, share the data , download the data. The huge number of data is get collected from different users. There are many open records systems were present. Due to which many users exchanged their data on the internet. For example, Facebook and twitter's API platforms were present, that collects personal social media data of many users. So it is very much important that data which is collected and downloaded for specific purpose by the data users ,it must be a true or real data. If the users get untruthful data or fake or false data, it many be chances that those particular user can get mislead by reading this fake data or fake reviews. Also it can make people lose confidence about quality of information on web. That's why it is very important that data should be a real one. Also security of that is also maintained. This is also one of important task that providing privacy preservation mechanism for that data. It is important to develop useful tool that can help web users differentiate truthful and untruthful information.Data market is an online store where people can buy data. Data marketplaces typically offer various types of data for different markets and from different sources. Common types of data sold include business intelligence, advertising, demographics, personal information, research and market data. Data types can be mixed and structured in a variety of ways. In this paper TPDM were introduced, Truthfulness and Privacy Preservation in Data Markets. The Fullyhomomorphic encryption technique were used for providing the security to the data by applying specific operations i.e, addition and multiplication operation on ciphertext data.

Literature survey:

T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su , They construct three major classification protocols that satisfy the privacy constraint: hyper plane decision, Naive Bayes, and decision trees. Nowaday machine learning concepts were used in many fields, such as medical or genomics predictions, spam detection, face recognition, and financial predictions. There is privacy concern so that's why these classification protocols were introduced. These protocols to be combined with the AdaBoost . And protocols are efficient, also it taking milliseconds to a very few seconds to perform a classification when running on the real medical data sets. [2]
Magdalena Balazinska, Bill Howe, and Dan Suciu, They discussed about It outline some of the key challenges that such markets face and also discussed the associated research issues that our community can help solve. Also they

told the implications of the emerging cloud-based data markets on the database research community. Our community has a great opportunity in making a significant impact on these data markets, while solving exciting data management research challenges. [4]

Dan Boneh, Matthew Franklin proposed a paper on fully functional identity-based encryption scheme (IBE). The system is based on the bilinear maps between groups. In this paper identity-based encryption scheme is introduced. The security of the system is a natural analogue of the computational Diffie-Hellman assumption. The limitation of this system is Revocation for private key is not present. [5]

Seung Hyun Seo, Mohamed Nabeel, Xiaoyu Ding, proposed a paper on An Efficient Certificateless Encryption for Secure Data Sharing in Public System storage clouds. The Safely share responsive data and information in public system storage clouds. Move forward towards the effectiveness. Additionally has downside that Network Connections Dependency furthermore Cost is more this calculation utilized is public key encryption algorithms.[6]

Proposed methodology:

In the proposed system, first efficient secure scheme is used for data markets, which simultaneously guarantees data truthfulness and privacy preservation. Also The TPDM is structured internally in a way of Encode -then-Sign, using fully homomorphic encryption technique and identity based signature. The service provider must be Collect the true data and process that data. Whenever user purchase product than he/she can send a review to the system than system first checks whether the contributors have authorized a person or not. In this paper a frame work has be proposed which uses a classification technique. The classifier used is support Vector Machine (SVM) which is very much efficient in detecting the fake accounts and separate real profiles from the fake profiles. The datasets which are used by the data user is get check before it has been using. SVM algorithm is used for that. Genuine and fake datasets were detected. And then privacy preservation is done for that datasets.

System Implementation:

Implementation of achieving data accuracy and privacy protection in data markets is break down into four models :

A. Data contributor

B. Service provider

C. Data consumer.

This are the four main entities.

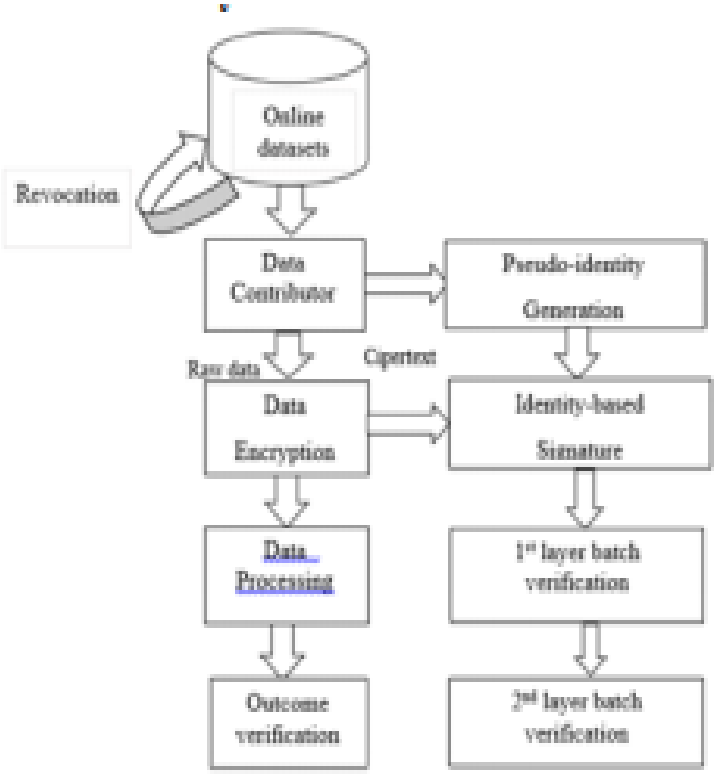


Fig.1. System architecture

The online dataset of OYO hotels were used for the project.

A. Data Contributor:

The user undergoes registration method the Registration centre can give the pseudo identity and provides them to the user. We have a tendency to assume that the registration centre sets up the system parameters at the start of information commerce. The verification conducted by each the service provider. For privacy mechanism the data contributor generate key request and that key request is send to registration center. The public and private key pair is generated for data contributor.

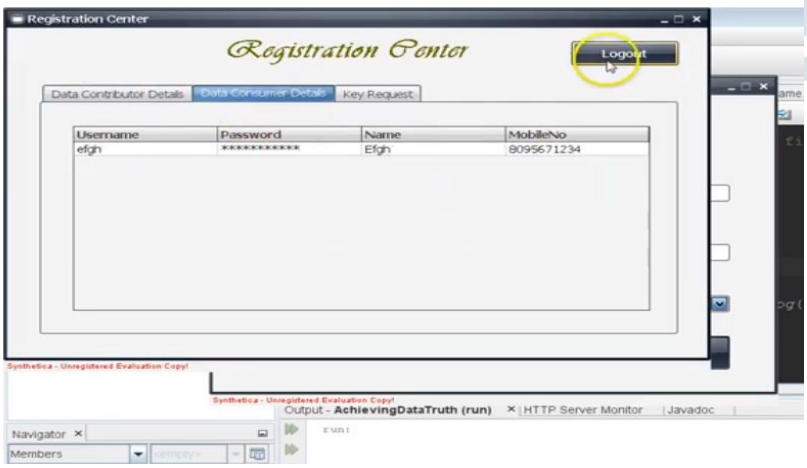


Fig. 2. Data contributor login

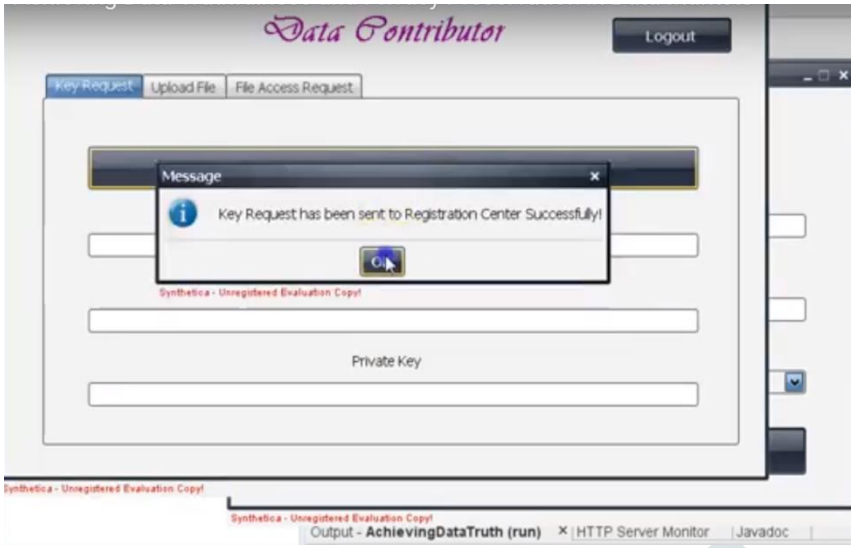


Fig. 3. Private-public key pair generated

B. **Service provider:** The service provider will ready to choose the service that in would like from the service provided by the collector. The data contributor upload the file. The datasets were used is Oyo hotels online datasets. The service provider is selected by data contributor. Then that online text file is get encrypted and signature is get generated. Also at service provider side, the first layer batch verification is done.

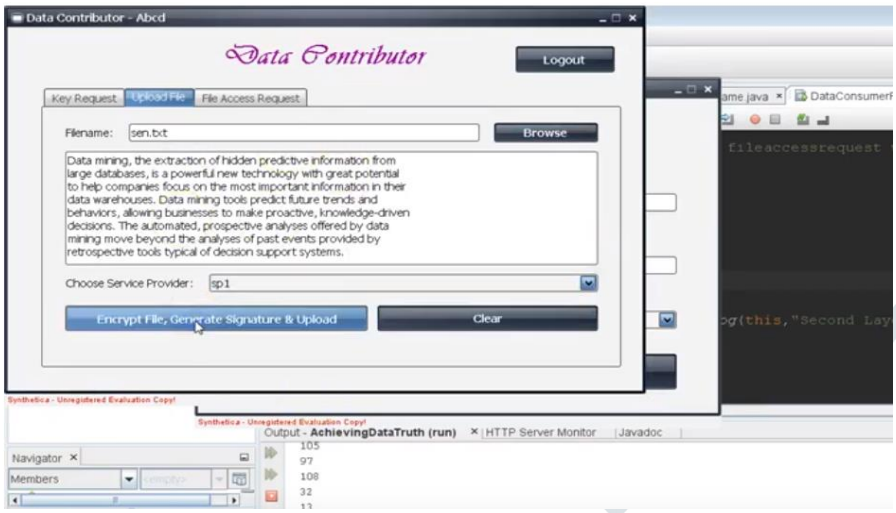


Fig 4. Data encrypted and signature is generated

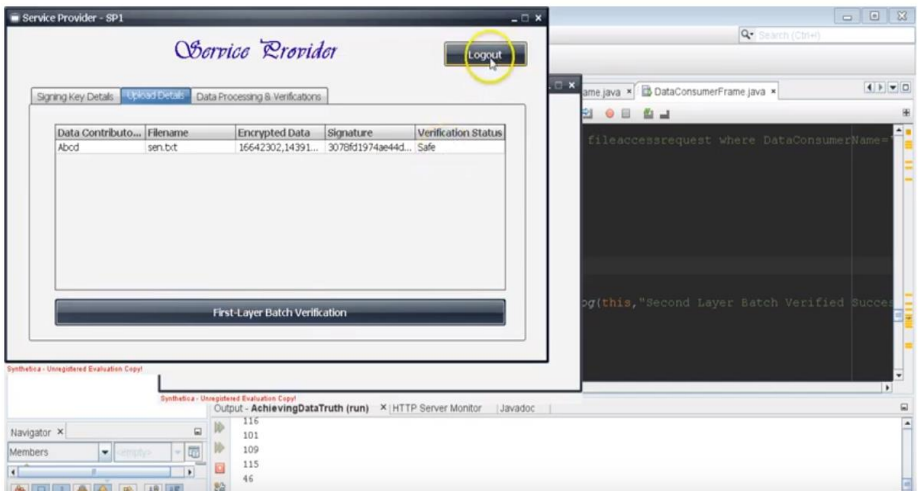


Fig. 5. First layer batch verification

C. Data consumer: Data consume will access the files, which are given by the data contributors at service provider. The status will indicates whether file is accepted or not. When data consumer give access request to data contributor at that time the details about that particular file is receive by the data consumer. At data consumer side, all details like (File name, data contributor name, signing- private key service providers name) are present. The file get downloaded. Data processing and verification is done. Second batch verification is done and data is get decrypted successfully.

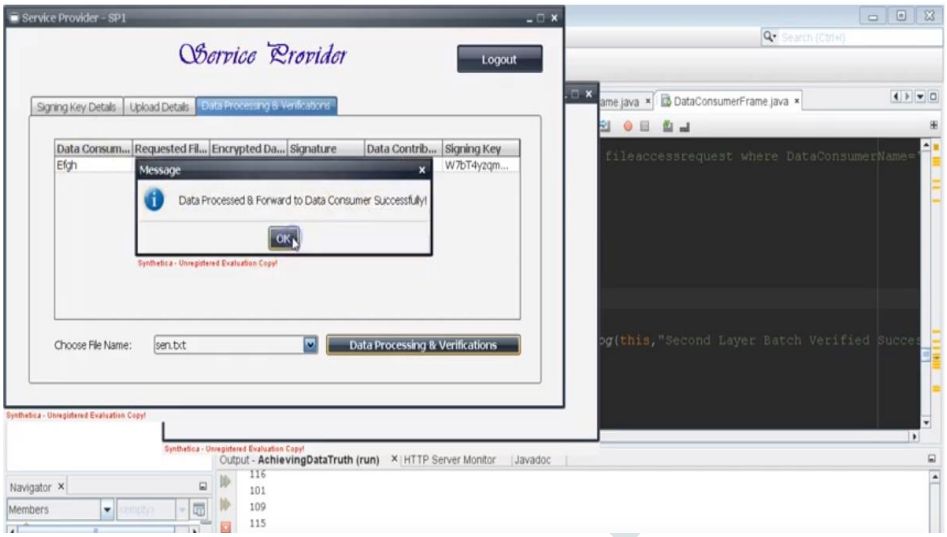


Fig. 6. Data processing

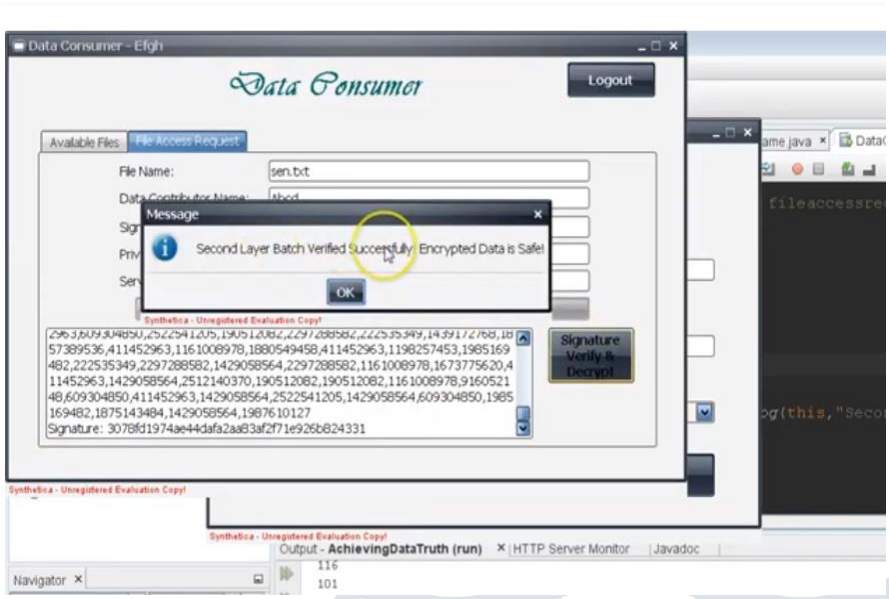


Fig.7. second layer batch verification and file decrypted successfully

Algorithm:

I. RSA Algorithm

INPUT: Required modulus bit length, k
OUTPUT: An RSA key pair ((N,e),d)
1. Select a value of e from 3,5,17,257,655373,
2. repeat
3. p ← genprime(k/2)
4. until (pmode)≠1
5. repeat
6. q ← genprime(k - k/2)
7. until (qmode)≠1
8. N ← pq
9. φ (n) ← φ (p)* φ (q)↔(p-1)(q-1) // 'φ' Euler's totient function.
10. e ← 1<e< φ (n)
11. d ← ≡ e (mod φ(N))
12. return (N,e,d)

A. Encryption:
Sender does the following:

1. Obtains the public key (n,e).
2. Represents the plaintext message as a positive integer m with 1<m<n
3. Computes the ciphertext c=m^emodn.
4. Sends the ciphertextc .

B. Decryption :

1. Person A recovers m from c by exploitation his or herprivate key exponent, d, by the computation m cd (mod n).
2. Consider m, Person A will recover the first original message M by reversing the padding scheme.

This procedure works since c me (mod n), cd (me)d (mod n), cd mde (mod n).
By the symmetry property of mods we have that mdemde (mod n).
Since de = 1 + k(n), we can write
mde m1 + k(n) (mod n), mde m(mk)(n) (mod n), mde m (mod n).

B. l- Depth tracing Algorithm [1]
Initialization: S = {δ1,δ2,…..δn}, head = 1, tail = n, limit = l,
whitelist = Ø, blacklist = Ø, resubmitlist = Ø
1.Function l-DEPTH-TRACING(S, head, tail,limit)
2. if │whitelist│ + │blacklist│ = n or limit = 0 then
3. return
4. else if CHECK-VALID(S, head, tail) = true then
5. ADD-TO-WHITELIST(head, tail)
6.else if head = tail then //Single signature verification
7.ADD-TO-BLACKLIST(head, tail)
8. else // Batch signatures verification from δhead to δtail
9. mid = │ head+tail │
10. l-DEPTH-TRACING(S,head, mid, limit -1)
11. l-DEPTH-TRACING(S,mid + 1, tail, limit - 1)

Result and analysis

The SVM training technique were used and result were generated. Depending on fake or real review of hotels online datasets the accuracy and error is get calculated. If the error is 0 then it will consider as fake reviews , if 1 then it will consider as real review.

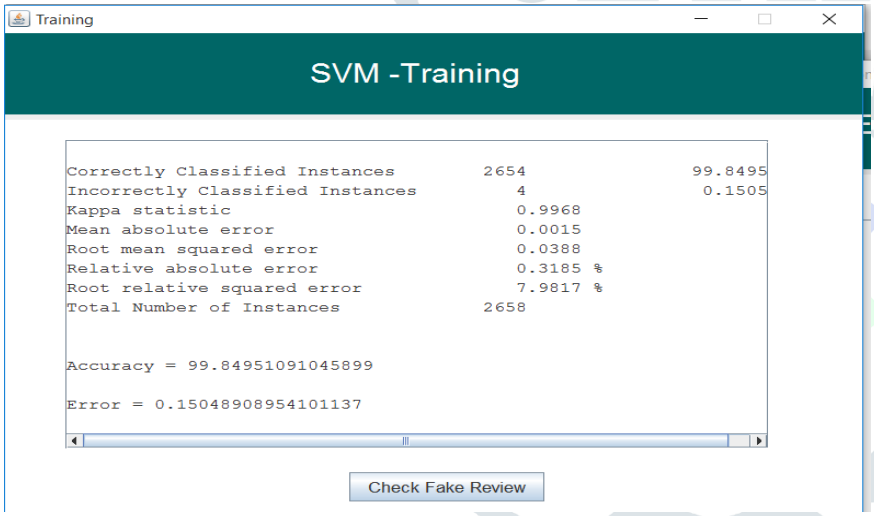
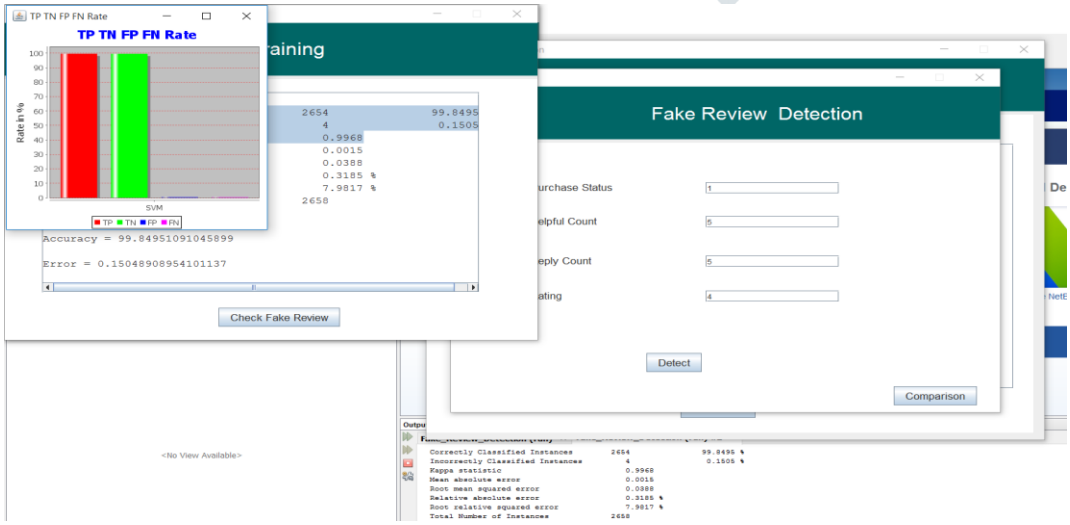


Fig. 8. Check fake review



Where, TP = True positive, TN = true negative, FP= False positive, FN= False negative

Conclusion

This paper, the information contributors must honestly submit their own data, however cannot impersonate others. Besides, the service supplier is enforced to honestly collect that data. The fake and real review will detected by using SVM algorithm at very initial phase. Only real reviews datasets were used by the data users. Also for privacy protection, the TPDM mechanism were used. For privacy protection of datasets the batch verification is done. And homomorphic encryption is used.

References

[1] ChaoyueNiu, ZhenzheZheng, Fan Wu, XiaofengGao and Guihai Chen " Achieving Data Truthfulness and Privacy Preservation in Data Markets "“ IEEE Transactions on Knowledge and Data Engineering (2018 Early Access).Study on the ISCX Dataset." Data Intelligence and Security (ICDIS), 2018 1st International Conference on.IEEE, 2018.

[2] T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su, AccountTrade: accountable protocols for big data trading against dishonest consumers, in INFOCOM, 2017.

[3] J. Camenisch, S. Hohenberger, and M. Ø. Pedersen, “Batch verification of short signatures,” Journal of Cryptology, vol. 25, no. 4, pp. 723–747, 2012.

[4] M. Balazinska, B. Howe, and D. Suciu, Senior Members, IEEE “Data markets in the cloud: An opportunity for the database community,” Vol. 4, no. 12, pp. 1482–1485, 2011.

[5] Dan Boneh, Matthew Franklin, Fellow,“Identity-based encryption from the weil pairing,” in CRYPTO, 2001.

[6] Seung-Hyun Seo, Member, IEEE, Mohamed Nabeel, Member, IEEE, Xiaoyu Ding, Student Member, IEEE, and Elisa Bertino, Fellow,An

Efficient Certificateless Encryption for Secure Data Sharing in Public system storage clouds,Vol.25, No.9, PP.2107.

[7] Ricardo Mendes, Student member, And Joaqo P. Vilela, "Privacy-Preserving Data Mining: Methods, Metrics, and Applications", Vol. 5, 2017

[8]Wakchaure MA, Sane SS. An Algorithm for Discrimination Prevention in Data Mining: Implementation Statistics and Analysis. In2018 International Conference On Advances in Communication and Computing Technology (ICACCT) 2018 Feb 8 (pp. 403-409). IEEE.

[9]Wakchaure MM, Sane SS. Performance Measurement of Various Threshold Values for Discrimination Removal and Data Quality Percentage by Different Discrimination Measures. International Journal of Applied Engineering Research. 2018;13(18):13961-8

[10] K. Ren, W. Lou, K. Kim, and R. Deng, “A novel privacy preserving authentication and access control scheme for pervasive computing environments,” IEEE Transactions on Vehicular Technology, vol. 55, no. 4, pp. 1373–1384, 2006.