Minimum Replication of User Data Integrating Anti-Collusion Scheme in Cloud Groups

C.Pabitha[1], K. Jayapreetha[2], P. Bharathi[3], J. Jayanthi[4]

[1]Assistant professor, [2,3,4]UG Scholars
Computer Science and Engineering

***Abstract*—In cloud computing, users can share data among group members with the characters of less maintenance and little management cost. Sharing data must have security guarantees, if they are out sourced. Sharing data while providing privacy preserving is still a challenging problem, when change of the membership. It might cause to the collusion attack for an unsecured cloud. For existing technique, security of key distribution is based on the secure communication channel, however, to have such channel is a strong assumption and is difficult for practice. We propose a secure data sharing scheme for dynamic users. Key distribution done without any secure communication channels and the user can get the individual key from group manager. Data deduplication is one of the techniques which used to solve the repetition of data. The deduplication techniques are generally used in the cloud server for reducing the space of the server. CloudMe is proposed for cloud storage. All files of data owners are encrypted using AES algorithm and stored in real cloud**

***Keywords*— Cloud Computing, Privacy Preserving, Anti-Collusion, Deduplication, Key distribution.**

1. INTRODUCTION

Cloud Computing is an innovative technology that is revolutionizing the way we do computing. The key concept of cloud computing is that you don't buy the hardware, or even the software, you need anymore, rather you rent some computational power, storage, databases, and any other resource you need by a provider according to a pay-as-you-go model, making your investment smaller and oriented to operations rather than to assets acquisition. In the simplest terms, cloud computing means storing and accessing data and programs over the Internet instead of your computer's hard drive. The cloud is just a metaphor for the Internet. It goes back to the days of flowcharts and presentations that would represent the gigantic server-farm infrastructure of the Internet as nothing but a puffy, white cumulus cloud, accepting connections and doling out information as it floats. Hybrid services like Box, Dropbox, and SugarSync all say they work in the cloud because they store a synced version of your files online, but they also sync those files with local storage. Synchronization is a cornerstone of the cloud computing experience, even if you do access the file locally.

Data deduplication is used in our scheme which is one of the techniques which used to solve the repetition of data. The deduplication techniques are generally used in the cloud server for reducing the space of the server. To prevent the unauthorized use of data accessing and create duplicate data on cloud the encryption technique to encrypt the data before stored on cloud server.

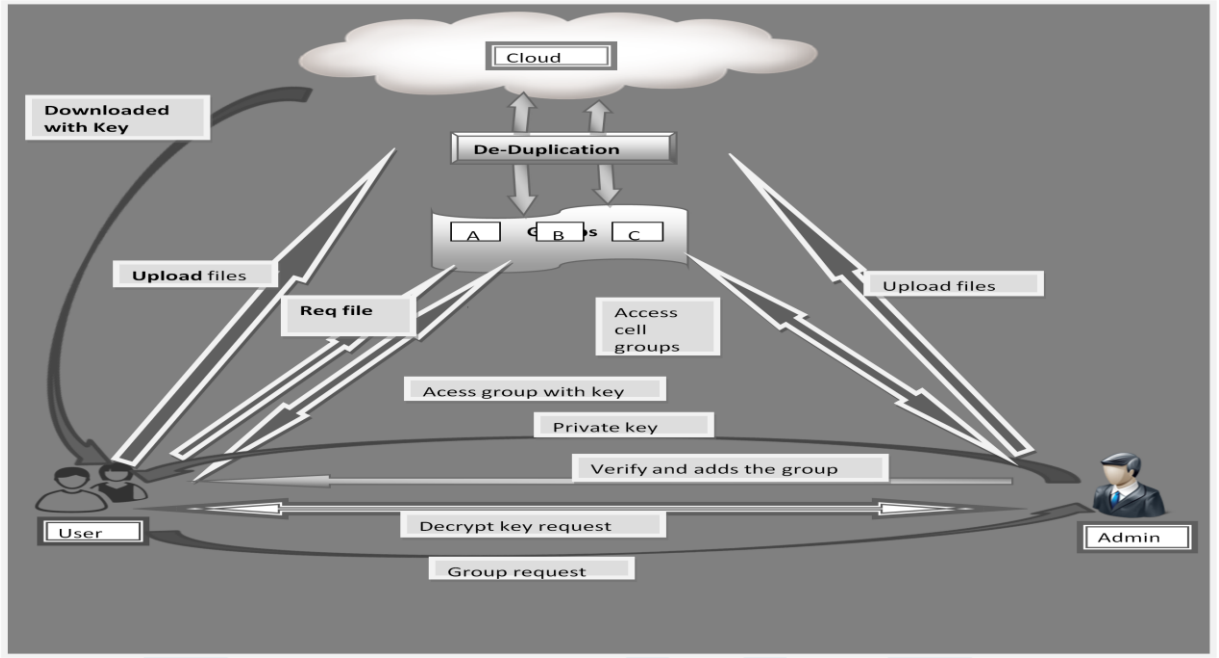
` In our scheme we propose anti–collusion scheme in cloud groups and integrated data deduplication technique. The following are the major contribution of our scheme:

1) In our scheme we provide, secure protocol for anti-collusion attack. It does not allow revoked user to access files in the cloud. There is no need for recomputing or updating private keys of other users.
2) By integrating algorithms / techniques we can implement deduplication concepts and reduce the storage cost in a cloud for the data owners.
3) Faster recovery and processing of data. This would significantly decrease the processing time of load balancer.
4) Effective and Efficient usage of cloud Storage Space. As data deduplication technique is used in this scheme the cloud storage is used effectively as there will not be any file of same content. Deduplication technique checks the content of each file while user uploads new files into the cloud.
5) Our scheme includes privacy preserving environment, only legal user can access through groups, any irrelevant entity cannot recognize the exchanged data and communication state even it intercepts the exchanged messages via an open channel, any adversary cannot correlate two communication sessions to derive the prior interrogations according to the currently captured messages.

The remainder of the paper proceeds as follows. In Section 3, we describe the system model. Our design goals are discussed in Section 4. Our proposed scheme is presented in detail in Section 5. Finally, the conclusion is made in Section 6.

2. LITERATURE SURVEY

Lu et al. [1] proposed a secure provenance scheme by leveraging group signatures and cipher text-policy attribute based encryption techniques [2]. There are two keys provided to user where one is used to decrypt the data that is encrypted by the attribute-based encryption and the group signature key is used for privacy preserving and traceability. The revocation is not



supported in this scheme. A secure multi-owner data sharing scheme [3], named Mona. It is claimed that the revoked users will not be able to access the sharing data again once they are revoked.

However, the scheme will easily suffer from the collusion attack by the revoked user and the cloud [4].The revoked user can use his private key to decrypt the encrypted data file and get the secret data after his revocation by conspiring with the cloud. In the phase of file access, first of all, the revoked user sends his request to the cloud, then the cloud responds the corresponding encrypted data file and revocation list to the revoked user without verifications. Next, the revoked user can compute the decryption key with the help of the attack algorithm. Finally, this attack can lead to the revoked users getting the sharing data and disclosing other secrets of legitimate members [5].

Zhou et al. [6] presented a secure access control scheme on encrypted data in cloud storage by invoking role-based encryption technique.

3. SYSTEM MODEL

Fig. 1 represents the System Model of our proposed system. In Section 2.1 the proposed system model is described in detail.

3.1. System Description

For user authentication Image based password system to decrypt and encrypted the file based authentication. When the Admin uploads the file in the cloud, the admin will split the image into 4 parts. The admin will hold 2 parts and the user of that respective group can view the other 2 parts. The images are spilt randomly using pseudo random generator technique. When the user tries to download a file, the user can send the requisition to the respective admin along with the user side available 2 parts. The admin will verify both the parts and if the authentication is passed, the file will be sent to the user in an encrypted way.

In our proposed project, we propose a secure architecture for handling file access in a dynamic cloud group.
The user belonging to an particular group is analysed and identified. After that a private key is sent to the user by the group manager in a encrypted format using RC4 encryption algorithm. The group manager performs the below tasks whenan new user joins the group or a user has left the particular group,

- *Update the whole user name list.*
- *Generate a secure key and encrypt the key without activation and send to the updated user list.*
- *Update the rights in the cloud server.*

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy.

Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

4. DESIGN GOALS

We describe the main design goals of the proposed scheme including Authority User Verification and Privacy-Preserving, Key distribution and access control, Detect Deduplication, Collusion Attack, Secure Data Sharing, Cloud Storage as follows:

A. Authority User Verification and Privacy-Preserving

At first Initial stage all users must create own username and password. After the Registration the user can login to their own space. This application verify the username and password which is either matched or not with the user registration form which is already created by the user while user registration process. If the valid user did not remember the username or password correctly the user can generate own password by using this application.

In the Privacy preservation environments, a reasonable security protocol would be developed to achieve the following requirements.

- *Authentication:* A legal user can access its own data fields, only the authorized partial or entire data fields can be identified by the legal user, and any forged or tampered data fields cannot deceive the legal user.
- *Data anonymity:* any irrelevant entity cannot recognize the exchanged data and communication state even it intercepts the exchanged messages via an open channel.
- *User privacy:* any irrelevant entity cannot know or guess a user's access desire, which represents a user's interest in another user's authorized data fields. If and only if the both users have mutual interests in each other's authorized data fields, the cloud server will inform the two users to realize the access permission sharing.
- *Forward security:* any adversary cannot correlate two communication sessions to derive the prior interrogations according to the currently captured messages.

B. Key distribution and access control

Group manager takes charge of system parameters generation, user registration, and user revocation. In the practical applications, the group manager usually is the leader of the group. Therefore, we assume that the group manager is fully trusted by the other parties.

Group members (users) are a set of registered users that will store their own data into the cloud and share them with others. In the scheme, the group membership is dynamically changed, due to the new user registration and user revocation.Once the user is revoked, the group manager creates the new encryption key for the specific group and transmits in an encrypted format using RC4 algorithm. Second the group manger updates the whole data list in the cloud server. Third the group manages updates the user list and activates the key for access.

C. Detect Deduplication.

Data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data in storage. The technique is used to improve storage utilization and can also be applied to network data transfers to reduce the number of bytes that must be sent. Instead of keeping multiple data copies with the same content, deduplication eliminates redundant data by keeping only one physical copy and referring other redundant data to that copy. Deduplication can take place at either the file level or the block level. For file level deduplication, it eliminates duplicate copies of the same file. Deduplication can also take place at the block level, which eliminates duplicate blocks of data that occur in non-identical files.

D. Collusion Attack

The user leaving a group is termed as revoked users. The revoked users can not be able to get the original data files once they are revoked even if they conspire with the untrusted cloud. Thus our proposed system detects the revoked users and protects the data confidentiality and privacy.

1) Secure Data Sharing

Secure data sharing is performed using private keys generated and transmitted using secure communication channels. In our scheme, the users can securely obtain their private keys from group manager Certificate Authorities and secure communication channels using RC4 algorithm.

2) Cloud Storage

The group user can upload the files in real cloud server named dropbox. Duplication of files are checked and the files is been uploaded in the cloud server. To get a file, the user needs to send a request to the cloud server. The cloud server will also

check the user's identity before issuing the corresponding file to the user. During file access the user key has to matched by the group manager and the requested file can be downloaded by the group users.

5. PROPOSED SYSTEM

5.1. DATA DEDUPLICATION

Data deduplication is a specialized data compression technique which makes all the data owners, who upload the same data, share a single copy of duplicate data and eliminates the duplicate copies in the storage.

The two approaches of data deduplication are described as follows:

- *Target-Based Data Deduplication:* The target-based approach only focuses on avoiding storing duplicate data. Those duplicate data are still uploaded repeatedly. Therefore, it cannot improve the volume of transmissions.

- *Source-Based Data Deduplication:* In this approach, users have to upload the identification of their data and query the cloud storage server whether the data are stored in the cloud storage before really uploading them. If the data have not been stored, users need to upload the whole data, and the cloud storage server completely stores them. Otherwise, users need to upload only the metadata, and the cloud storage server simply creates a pointer, which points to the first stored copy. Therefore, the source-based approach can improve both the utilization of the storage and the bandwidth. Nevertheless, it changes the familiar process of cloud storage services. When users want to upload their data, they must query the cloud storage server for the existence of the data first.

A. Genetic Programming

Cloud is unavoidable for storing and ubiquitous retrieval of data. Data deduplication is a data compression and duplicate detection for eliminating duplicate copies of data to make storage utilization efficient. In order to avoid false positive and false negative data retrieval from records, Genetic Programming Approach (GP) is being used for deduplication in databases [7]. Genetic Programming approach is a systematic domain and independent programming model for deduplication [8][9].A sequence matching algorithm and Levenshtein's algorithm are used to Text Comparison. Information as a Service is a poor storage performance where the organization need to pay more for each and every GB[10]. Hashing identifies data with identifier. Fig. 4.

This is implemented by assigning an identifier (metadata) for each chunk of data or a file, calculated by using cryptographic hash functions[11][12].

Fig. 3. Represents Genetic Programing.

Steps for GP approach

1. Randomly generate an initial population
2. Repeat until the best solution or a stop criterion is reached.
   - 2.1. Evaluate of each individual by means of the fitness function.
   - 2.2. Select a subgroup of individuals onto which genetic operators are to be applied
   - 2.3. Apply the genetic operators
   - 2.4. Replace the current population by the new population
3. End

Fig. 5. Represents Nature Inspired Deduplication framework.

The goal of NIDF is to provide a reliable system for identifying duplicates in the Cloud Server and making it a prerequisite to add files to the Cloud Storage.

Deduplication can be done at the source or destination. NIDF involves destination deduplication since there is less overhead. In case of Source deduplication, all the files in the Cloud Storage have to be sent through a network to the Cloud User (Source) which causes an unnecessary overhead and bandwidth wastage.

As the result of this approach the performance accessed is measured based on the parameters precision,Recall and F-measure are used for comparsion along with the parameter time taken to identify the duplicates[8]. F-measure is the harmonic average of precision and Recall.

This has limitations as all Cloud Service Providers and Cloud Users rests with this method as it encounters the basic constraints in storage space posed by the presence of duplicates.

This work can further be extended to a variety of file formats including images and video file formats by utilizing appropriate signal processing techniques.

5.2. PRIVACY-PRESERVING

We discuss privacy preserving technique which concentrates more to prevent Insider-Collusion attack which is fasters growing type of breaching attack. Insider attack is one of the top three central data violations [13]. According to the ‘‘2015 Verizon Data Breach Investigations Report,’’ [14] attacks from ‘‘insider misuse’’ have risen significantly, from 8% in 2013 to 20.6% in 2015.

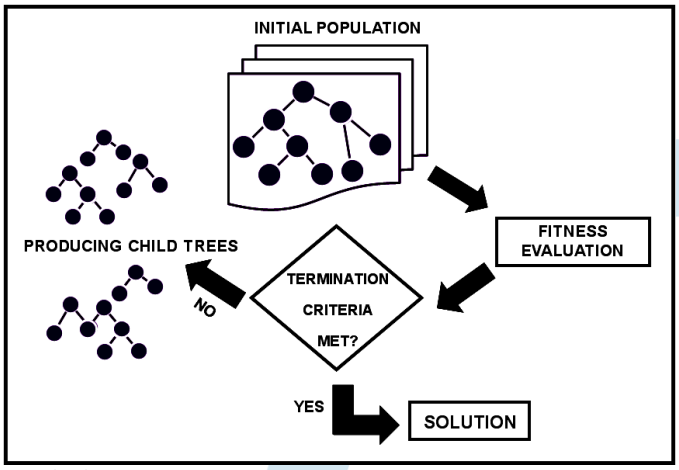


Fig. 3. Genetic Programming

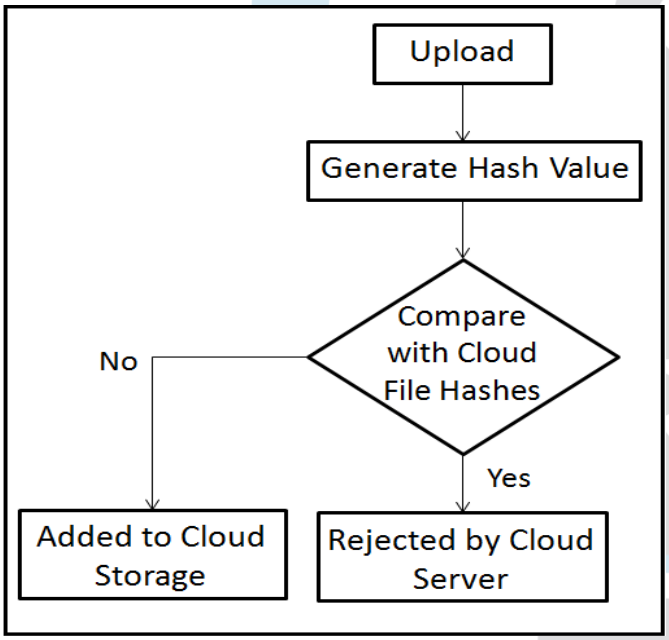


Fig. 4. Deduplication by Hashing

Insider attacks may be of lack of technical barrier. Prime area of research in privacy is Support Vector Machine(SVM). Important examples of insider threat schemes are mentioned by Claycomb and Nicoll[15] and include the following two cases:

1) Administrators of a rogue cloud service provider.
2) Employees in victim’s organizations who exploit cloud weakness for unauthorized data access.

B. Encrypted Data Deduplication

The proposed encrypted data deduplication mechanism is described in [16]. When a cipher structure is uploaded, the cloud storage server checks whether the check block (i.e., the first hash value) is in the hash list. If the check block is not in the hash list, the content of the uploaded cipher structure has not been uploaded. Then, the cloud storage server stores the whole cipher structure and adds the check block into the hash list. Otherwise, the cloud storage server converts the uploaded enabling block:

(1) First, the cloud storage server calculates the modular multiplicative inverse of the uploaded converting block and multiplies it by the first stored converting block under modulus p as the conversion factor.
(2) Then, it encrypts the conversion factor by the multiplicative asymmetric homomorphic encryption using the public key of the data owner, who uploads the cipher structure in this session.

(3) Finally, the cloud storage server performs the operation Θ on the encrypted conversion factor and the uploaded enabling block to produce the converted enabling block. After converting the uploaded enabling block, the cloud storage server only stores the converted enabling block.

5.2.1. PRIVACY-PRESERVING METHODS

The Privacy-Preseving Methods dicussed below is based on Peter Shaojui Wang, Feipei Lai, Hsu-Chun Hsiao, And Ja-Ling Wu,"Insider Collusion Attack On Privacy-Preserving Kernel-Based Data Mining Systems"[13].

A. REDUCING THE NUMBER OF THE INSIDERS

Catching Insiders based on Temporal Events
*Step 1:*Extradimensional data will likely to be insider action which should be marked as 'possible insiders data' by K-Means.
*Step 2:*Monitoring system detected malicious activities finding several data marked as 'possible inside data'. Then raise an alarm and notify the organization administrators.
Step 3: check whether those 'possible insider's data' access are true or simply erroneous judgement. This check is performed by administrator by comparing system logs of unauthorized access by which number of possible insiders can be estimated.

B. EXPANDING THE DIMENSION OF THE DATA

The next method is to counter the attack is by making data set larger than insider.

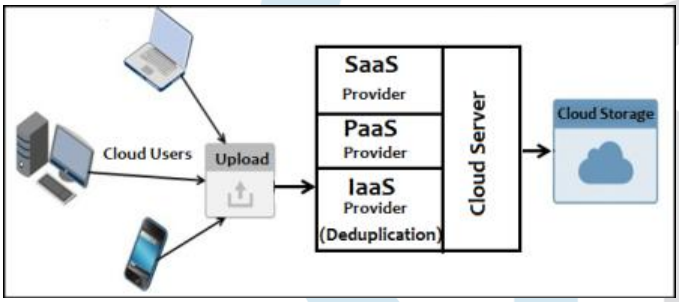


Fig. 5. Nature Inspired Deduplication framework

6. CONCLUSION

In this paper we have discussed secure anti-collusion data sharing scheme for dynamic groups in the cloud. In our scheme, the users can securely obtain their private keys from group manager Certificate Authorities and secure communication channels. Our scheme achieve secure user revocation, the revoked user cannot access any files in the cloud and also there is no need of updating or recomputing private keys of other users. Also, we have developed a effective deduplication check for users while uploading the data in the cloud environment.

REFERENCES

[1] R. Lu, X. Lin, X. Liang, and X. Shen, *"Secure provenance: Theessential of bread and butter of data forensics in cloud computing," in Proc. ACM Symp. Inf., Comput. Commun. Security,2010, pp. 282–292.*
[2] B. Waters, *"Ciphertext-policy attribute-based encryption: Anexpressive, efficient, and provably secure realization," in Proc. Int. Conf. Practice Theory Public Key Cryptography Conf. Public Key Cryptography,2008, pp. 53–70.*
[3] X. Liu, Y. Zhang, B. Wang, and J. Yang, *"Mona: Secure multiowner data sharing for dynamic groups in the cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1182–1191, Jun. 2013.*
[4] Z. Zhu, Z. Jiang, and R. Jiang, *"The attack on mona: Secure multiowner data sharing for dynamic groups in the cloud," in Proc. Int. Conf. Inf. Sci. Cloud Comput., Dec. 7, 2013, pp. 185–189.*
[5] Zhongma Zhu and Rui Jiang, "*A Secure Anti-Collusion Data Sharing Schemefor Dynamic Groups in the Cloud", IEEE Transactions On Parallel And Distributed Systems, Vol. 27, No. 1, January 2016*
[6] L. Zhou, V. Varadharajan, and M. Hitchens, *"Achieving secure role-based access control on encrypted data in cloud storage," IEEE Trans. Inf. Forensics Security, vol. 8, no. 12, pp. 1947–1960, Dec. 2013.*
[7] D. T. Meyer and W. J. Bolosky, *"A study of practical deduplication,"Proceedings of the 9th USENIX conference on File and storage technologies, ser. FAST'11. Berkeley, CA, USA: USENIX Association, pp. 1–6, 2011.*
[8] Moise´s G. de Carvalho, Alberto H.F. Laender, MarcosAndre´ Goncalves, and Altigran S. da Silva, *"A Genetic Programming Approach to Record Deduplication," IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No.3,pp.399 - 412, March 2012.*

[9] P.Shanmugavadivu,N.Baskar, *"An Improving Genetic Programming Approach Based Deduplication Using KFINDMR", International Journal of Computer Trends and Technology, Vol.3, Issue5, pp.694-701, 2012.*
[10] W. Hu, T. Yang, J.N. Matthews, *"The good, the bad and the ugly of consumer cloud storage," SIGOPS Oper. Syst. Rev. 44 pp.110–115,2010.*
[11] Amrita Upadhyay,Pratibha R Balihalli, ShashibhushanIvaturi andShrisha Rao, *"Deduplication and Compression Techniques in Cloud Design", Proceedings of IEEE International Systems Conference(SysCon), pp. 16, 2012.*
[12] Qinlu He, Zhanhuai Li, Xiao Zhang, *"Data Deduplication Techniques," Proceedings of IEEE International Conference on Future Information Technology and Management Engineering,Changzhou, China, FITME, pp.430-433, 2010.*
[13] Peter Shaojui Wang, Feipei Lai, (Senior Member, Ieee), Hsu-Chun Hsiao, And Ja-Ling Wu, (Fellow, Ieee), *"Insider Collusion Attack On Privacy-Preserving Kernel-Based Data Mining Systems", IEEE Access, Received April 18, 2016, Accepted April 25, 2016, Date of Publication April 29, 2016, Date of Current Version May 23, 2016,Volume 4, 2016.*
[14] 2015 *Verizon Data Breach Investigations Report*, Verizon, Bedminster, NJ, USA, 2015.
[15] W. R. Claycomb and A. Nicoll, *"Insider threats to cloud computing: Directions for new research challenges," in Proc. IEEE 36th Annu. Comput. Softw. Appl. Conf. (COMPSAC), Jul. 2012, pp. 387_394.*
[16] Chun-I Fan, Shi-Yuan Huang, Wen-Che Hsu, *"Encrypted Data Deduplication in Cloud Storage", 2015 10th Asia Joint Conference on Information Security*.