SPABOX: Safeguarding Privacy in a Middlebox using Decision Tree Algorithm

M.Muthuselvi [1], Arumugathai.M [2], Kalaivani.V [3], Monisha.M [4]
[1] Assistant professor, Dept of CSE, University College of Engineering, Nagercoil, Tamil Nadu, India.
[2, 3, 4] Student, Dept of CSE, University College of Engineering, Nagercoil, Tamil Nadu, India.

Abstract – Network Security is to prevent unauthorized access, misuse, modification or denial of a computer network which consists of policies and practices and network accessible resources. Middle Boxes are used to secure and preserve privacy in communication between the server and client and also detect attacks and malwares that often deployed by the network operators. While transferring message packets, it is added into a network.DPI (Deep Packet Inspection) is a technique used along with the middle boxes for managing the network traffic which is a form of packet filtering.DPI identifies the location, classifies, reroutes or blocks of packets with specific data or code payloads. It uses simple keyword-Based matching technique for improving the security in the entire networks. However, it issues problem of network intruders.SPA Box enables privacy for preserving regular expression evaluation and machine learning by using SVM (Support Vector Machine) which is used for malware detection .In proposed system, the work is intended to use decision tree algorithm over SVM for improving performance of SPA Box.

Index Terms – SPA Box TX (Sender), SPA Box RX (Receiver), Rule Generator (RG), DPI (Deep Packet Inspection), SVM (Support Vector Machine), Privacy preserving, Middle Box.

1. INTRODUCTION

A.Network Operation

HTTPS uses both Transport Layer Security (TLS) and Secure Sockets Layer (SSL) which is a popular Internet protocol; it's to encrypt the communication between clients and servers to ensure data integrity and privacy. The network operator deployed the Deep Packet Inspection functionalities for the detection of attacks that can be provided by many Middle Boxes.It can be detected by searching specific key-words or signatures in non-encrypted traffic. Middle Boxes also detect Malware which uses many techniques such as obfuscation and polymorphic or metamorphic strategies. Industry and academia adding more advanced machine learning and data mining analysis in DPI.

In existing system, the performance of keyword or signature matching technique were downgraded. This system uses advanced machine learning analysis for malware detection in the network traffics.Song, Li and Cao [5] suggested that to store the data on the data storage servers such as mail servers and file servers in which the datas are encrypted.since,the security of such paper are reduced and also some risks occurred in privacy.However,it has one advantages too; 1)almost no space 2)communication overhead. The main challenge is to implement that in practical usage today.W.Diffie and M.Hellman [1] proposed that to provide the tools for solving cryptographic problems of long standing.It also reduce the security and privacy.so,we improved that security in the network by using SVM(Support Vector Machine).Drawback of this paper is that,it cannot learn anything more about plaintext.[8]Crepeau,Nao and Kilian tried to described most efficient protocol for 1-out-of-n OT to date.When active attacks occur,it cannot achieve security level in the network.But,it has one advantage to improved UC-security against active and adaptive corruptions.

[6]A.Yao suggested to generate and exchange secrets over the client and the server.Introducing a new tool for controlling the knowledge in cryptographic protocol design.It lost data confidentiality of this design.This is the main drawback in this paper.But,it has some advantages too;Encryption only needs two operations1)stream cipher2)block cipher.TaherElGama improve the security based on public key cryptosystem and signature scheme by using the algorithm of Discrete Logarithms.It also implemented the key exchange called Diffie-Hellman key exchange.so,it achieve a public key cryptosystem.

The drawbacks of this paper suggested that it has low complexity and also it can compute the difficult discrete logarithms over finite fields which is the pro of this paper[2].[4]MA Salehi tried to search the regular expression over encrypted data in the cloud technology.This paper has less accuracy in RESeED when compare with other.It can be efficient and secure representation of the data.Cash introduced searchable encryption for leakage-abuse attacks.

It achieve high efficiency with provable security.But,it cannot explored realistic active attacks[7].[3]Kedar and Girija suggested that intrusion detection by using robust and fast pattern matching algorithm for full regex syntax which builts only for regular expressions.Under the backtracking algorithm,it can improved by substantial manner.Kedar and Girija, both are introduced performance vulnerabilities.

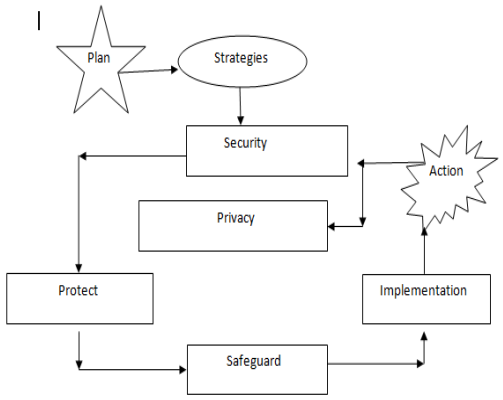

Fig 1:Illustration of safeguard privacy

B.Review of the paper

In recent years, data security, data integrity and transaction security are more priority for the enterprise. There are many security threats occurred in many forms. These threats comes from either external threats or internal threats. External threats comes under malicious attacks where as internal threats are comes from data theft. Each threat must be evaluate carefully in which the management of risk can take more effective. Safeguard is used to protect client data and utilizes system controls. These controls maintain high levels of security and minimize risk. This safeguard continuously test its system to protect against security threats. If any threats are identified, the safeguard's teams quickly implements security upgrades while other controls are restore in some place and recover any data that may be lost or any disruption.Fig 1:illustrated that the safeguard privacy in the network system

C. Scope of the paper

Evaluation of safeguard placed in many subject requirement areas 1) Record Keeping 2) Restricting Access 3) Secure Storage 4)Incident Reporting 5) Employee Awareness 6) IT Security 7) Disposal. *Record Keeping:*In safeguard privacy,We keep the record of user data in safe manner.Restricting Access:The user can access them data should not be a any restriction.It can be in user friendly.Secure Storage:The client data information are stored in secured manner that a way to store importance.Incident Reporting:If any incident be happened at any performed operation,then it can be report some message to that client side for additional evaluation.Employee Awareness:Employee should aware about inferred data from the client side.It can be lose from any source of malware detection.IT Security:Security is the mani cercern in our safeguard privacy.It can performed between client and server side.Disposal:If any token cannot be encrypted,it can be drop by MB.The RG decide that ttoken be droped at the client side(S).

2. THEORITICAL APPROACH

A. System Architecture

The system architecture consists of SPABOX, Middle box , RG (Rule Generator). The SPABOX access sender as well as receiver. RG usually provides set of attack rules. Each rule contain set of keywords and other information. Which has offset values of each keyword. This RG define and attack such as Symantec and McAfee. The trained model is provided by these RG for classifying encrypted traffic which decide whether the traffic contain the malicious attacks. The middlebox(MB) are usually deployed by AT&T which is a network operator . The traffic identified the security threats in two ways. 1) The rule set is provided by the RG which is compared with encrypted traffic by MB and observes matching between traffic and attack rules used by S, that's rules are presented in the rule set. If it is match, then there is no regular expression in the attack rule. Since MB can choose to drop the packet and informed to administrator or issue some warning message(what regular MB would do over unencrypted traffic).The packet will be forwarded to R when regular expression required to be further evaluated . 2) Trained model is provided by RG which classifies with encrypted traffic that can be done by MB and results are used by R to decide whether the traffic contain malware are not.When S and R wants to establish HTTPS connection, the following steps take place (a) First we establish a connection setup then we first run SPABOX handshake to exchange SSL section key that's indicated by K_{ssl}

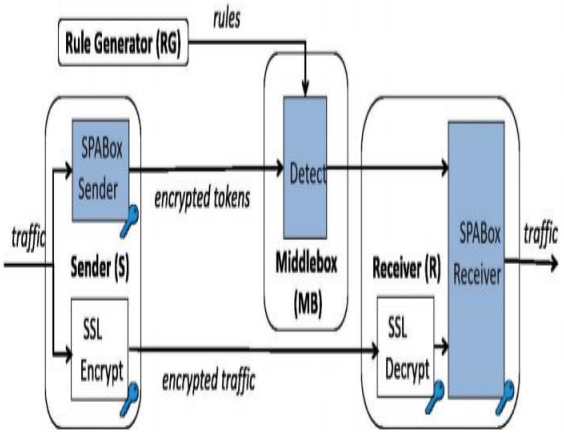

Fig 2:Illustration of system architecture

During this process,S and R needs seven parameters such as (SPAPara),g,n,r,s,N^2 ,salt and a hash function H(.).These parameters are used for encryption, decryption and detection in

the SPABox protocol.MB transparency property is preserved when SPABox handshake(in which S and R are involved) bootstraps off the existing SSL handshake.(b) At S, two S connections of SSL and SPA are setup. On the SSL connection, S encrypt the traffic with SSL. On the SPA connection S makes copy of traffic which is tokenized and then encrypts by using Discrete Logarithm Problem (DLP) which is detail described in section III. (c) MB perform two task. Once task receives encrypted tokens which is comes from S over SPA connection. (d) At R, it decrypt and authenticate the traffic by using regular SSL. After that, R tokenized and encrypt the plain text. The resulted plain text compare with traffic which comes from SPA connection. The encryption is much faster than decryption by using the protocol of SSL when comparing the decrypted packet and detokenized traffic. In other, comparing the cipher text over SSL connection the R can determine whether S follows SPABOX protocol correctly or not which including both keyword matching and malware detection. If any discrepancy is occurred, R may thing SSL attacker and further more it drops the connection immediately. Otherwise R may process information that can be forwarded by MB.

B. System Implementation

In this, we described a detailed implementation of SPABox, clients and MB. On the rule generated side -LIBSVM 3.21 library is used for traning SVM model. On the, client side we implement SPABOX on the top of OpenSSL-1.0.2d library. We also modify handshake process in this library. Another library called GMP 6.0.0 which is used to convert each token.We choose 5 bytes long in these token that can be converted I to large integer value (mpz_t format). After that we can hash and encrypt each token by using corresponding large integer. We choose the parameter as g,$salt_0$, d, n ,s and N^2 to be 80, 20, 20, 10, 160 and 4096 bits respectively in this we choose SHA-1 hash function. In security analysis basics, the parameter r is set to be between 20 and 40.

When S established the connection it create two sockets. One is used for SPABOX handshake which is sending normal HTTPS traffic and another one is used for encrypted token transmission. The additional implementation of garbled circuit and OT is performed at S side which implementation is similar to R. If R matches the traffic with regular expression successfully, then its stop the connection.

On the middle box side ,we implement MB with DPDK in click modular router. And then we build both type of hash table called THT and HHT. The THT start with 65,536 slots and when its reaches more than 50% full, it must be resized. In this hash table half of all the threats uses keyword matching technique for matching encrypted tokens. But in HHT, one threat is used to search if keyword tokens are matched. If it is matched, there is no regular expressions evaluation is needed. Since MB block the connection and notify R. If the regular expression run on the receiver side it is used to transfer input key string. The remaining threats are used for malware detection which is implemented based on the library called GMP 6.0.0 library. Otherwise, the result is send to other traffic of R.

C. Models and Protocols

To protect the privacy of user traffic from MB which performs DPI and also detect malicious traffic. Besides, MB can detect attacks over encrypted traffic with uses rule sets and model called ML model which are provided by RG.

1) Machine Learning (ML) model

In this proposed system we uses the model of ML which is used to improve their operation on behavior. It has three categories i) Supervised learning- it contain data sets which includes decide output. Such a function can calculate the error for prediction. ii) Unsupervised learning- In this type of data set does not include decide output. Therefore there is no way to perform this function. So the data set is segmented into class in which each class contain portion of dataset. iii) Reinforcement learning- This is used to learn action for given set of states. An error is not provided in it.

2) Security Model

In this model we assume that i) RG is honest ii) Mb is honest-but- curious. That is, it will implement the rules and follow some set of protocol honestly. The security model design SPABOX which can achieve goals that can be summarized as follows i) To guarantee the confidentiality of unmatched traffic ii) No private information can be inferred by the MB.

The first goal only allow the data that are exactly identical to known suspicious keyword which are presented in MB. The second goal unable to apply the techniques of data analysis about the traffic such as leakage-abuse attacks.

3) SSL and TLS protocol

This protocol handles keyword matching regular expression and malware detection via machine learning. Based on the security model this protocol is used to provide privacy of user data over HTTPS connection. Besides, SPABOX is used to maintain the properties that can be provided by these existing protocol. The property are as follows i) Private connection- All traffic are encrypted with secrete key. So this traffic has both secure and reliable. ii) Identity authentication – This can be achieve by using public key cryptography. This can be authenticated by each endpoints of identity. iii) Reliable connection- the message transmitted by one party of the section that can be able to verify by other partied in this section. It prevent undetected loss during the transmission. These protocols is used for providing a privacy to client related data. During transmission, it also provides models such as security models and machine learning model.

3. ALGORITHM DISCUSSION

a) Discrete Logarithm Problem

In this algorithm two main problems are happen which is based upon public key cryptography. These problems are 1)Integer factorization which is mainly used in RSA algorithm.The another one problem is DLP this can be occur in Diffie-hellman key exchange algorithm.When the numbers are sufficiently large and no efficient the integer factorization problem is known.Many cryptographic protocols are based on these related problems.Not all numbers are hard to factor the problem is only that semi primes when the product of two prime numbers.For instance, if these two numbers are large i.e)more than 2000 bits long.There is no efficient algorithm is known for computing discrete logarithms.It is computing($\log_b a$) to raise b to larger and larger powers k until a is found.This algorithm requires running time in a group G.It is run faster than naïve algorithm.Let G be a cyclic finite group and g€G be a generator of G.The DLP in g is the following: Given an element h€G, find the smallest positive integer x such that h=[x]g (additive group)/h=gx(multiplicative group).We will denote such an x with DLP of h.

b)Decision Tree Algorithm

This type of algorithm is the flow chart like structure where each internal node acts and attributes each branch represents outcome and each leaf node acts as a class lable.There is a root node presented at the top most layer in a tree. It is used to represent decision making. Decision tree algorithm may able to handle both numerical and categorical data. So, the proposed system used this algorithm effectively. It can be implemented by using software packages such as IBM, SPSS Mmodeler, Rabid Miner, SAS Enterprise Miner, Matlab, etc.

C.Keyword Matching

We can performed this technique at S,R and the MB.It support two more functions.An attack rule may contain position information as well as multiple keywords. On the sender side, we tokenize the trafiic that can be sent over a SPA connection using fixed-length sliding window.The length of the keyword are equal to or greater than 5 be searched, if the keyword tokenize in the rule set.Then the MB compare all the tokens that can be received at the receiver side (R) over the connection of SPA with these keyword tokens.In implementation,we use 5 bytes window length.Because of this reasons, i) message length is too large ii)longer sliding window slower in speed.After that, these token are converted into an integer of mpz_t format,that can be performed before being encrypted.After tokenization,S can encrypt each token based on DLP.Assume that the plaintext be tokenized as t,where n is the prime number,g is an element in multiplicative group Z_n^* and the element of this is that salt in the additive group Z_n ,t using the hash function as H(.). After tokenization,the n is set to be 160 bits,it can be converted into 20 bytes after encrypted that.So,we computed as follows:

$$C=g^{salt_k.H(t)} \mod n$$

Where $salt_k = salt_0 + kd \mod Z_n$ .To make our protocol secure, S is required a random number r of tokens and set this as sequence of preceding tokenized traffic over a SPA connection.Keyword Matching technique used for matching the keyword that can be tokenized.After that,it can be encrypted.That encrypted token is detected by MiddleBox(MB).

4. RESULTS AND DISCUSSION

A.Data Sets

Our keyword rule set contains 11,202 keywords and 21,035 distinct keyword tokens.

In malware detection, we uses two different data sets. They are malware data set and benign data set. The malware contains 17258 malicious program including different malware families that can be represented by different types of malware. The benign contains 1000 legitimate executablesin which most of the system files are gathered from the machines running on our campus. Now, we use 70% of the file be traing set ,remaining 30% be used as the testing set.

B. Performance

In this, we descried about performance of the entire SPABox, keyword matching used in blindbox and compare them with SPABox.

1)On the Client Side

In our client side,we use two desktops that can be equipped with Intel Core i7 processors and 16GB memory.These machines are multicore,by we use only use one thread per client.The CPU supports AES-NI instructions. So we compare our solution with the blindbox.It takes 9 x more time.The main reason is that it support for AES encryption.Blindbox takes amount of time as SPABox with similar one.When compare with Paillier encryption method saves almost 20 x time.This process takes more computation overhead on the receiver side(R).Another overhead is that when evaluating regular expression comes from OT.During performance requirement,we use OT implementation.This protocol bootstraps off Diffe-Hellman Key-Exchange protocol.Hence,it is more efficient.At 5mbps, the encryption cost is limited as the CPU continuously generate data.but the throughtput increases as 23mbps.

2)On the MiddleBox Side

MB protocol is implemented on the server with two 2.0GHz Xeon E5335 cores and 16 GB RAM.The CPU doesnot support AES instruction.SPABox can save 29.5% time when compare with blindbox.It only takes 239µs at the MB.Incurred overhead is too small.To keeping this speed,to use the hash function instead of using searching tree in our algorithm. Each keyword

token is hashed and then it can be stored in the hash table.Such that MB can quickly find out.If the hash table increases its memory,then we consider acceptable trade-off.Therefore,the memory overhead is negligible by THT.

3)Network Overhead

Assume that, we have 3K keywords and the MB is 20mbps.SPABox can finish setup with 40µs on the client side while bulidbox required interaction between client and MB .The garbled circuits for each keywords as 599KB that can be in buildbox.To compute this,to take more than 90 min and take huge computation on the client side.On the contrary,MB needs 7.5 s to evaluate them when 228byte information needs from the client.So, SPABox has better support for connections and mobile users.

5. CONCLUSION

In this paper, our goal is to secure the client side data by using Decision Tree algorithm.we have presented SPABox ,it supports both keyword based and data analysis based DPI functionalities over a encrypted traffics.we also improved privacy of the user data at the middle side. SPABox does not required any interaction between MB and a client .So, we use this.It also enables privacy-preserving regular expression and machine learning by using DLP for the detection of malware. The performance of SPABox has limited overhead. In our future work, to improve the performance evaluation inorder to support DPI.

REFERENCES

[1] W.Diffie, “New directions in cryptography,” IEEE Trans. Inform. pp. 587-594,1984.
[2] TaherElGama, ” A Public Key Cryptosystem and a Signature Scheme based on Discrete Logarithms” Feb 7, 2017.
[3] Kedar Namjoshi and Girija Narlikar “Robust and Fast Pattern Matching for Intrusion Detection”. Conference Paper · April 2010.
[4] MA Salehi “Regular Expression Search over Encrypted Data in the Cloud” 2014 IEEE 7th International Conference on Cloud Computing (CLOUD) .
[5] Song , Li and Cao , ”Practical Techniques for Searches on Encrypted Data” Conference Paper ,February 2000.
[6] A. Yao, ” How to generate and exchange secrets" Dec 21, 2017.
[7] cash, ”Leakage-Abuse Attacks Against Searchable Encryption” Oct 12, 2015.
[8] Crépeau ,Nao and Kilian ,“The simplest protocol for oblivious transfer”, Aug 23, 2015.