Data Leakage Prevention for Data in Transit using Artificial Intelligence and Encryption Techniques

Mohammed Ghouse
DM, Network and Cyber Security,
Bharat Electronics Limited,
Bangalore, India
mohammedghouse@bel.co.in

Manisha J. Nene
Dept. of Computer Science and Engineering,
Defense Institute of Advanced Technology,
Pune, India
mjnene@diat.ac.in

VembuSelvi C
MRS, Cyber Security
Central Research Laboratory, BEL,
Bangalore, India
vembuselvic@bel.co.in

***Abstract*—Data leakage in an organization is a very important concern that leads to the ex-filtration of data, The work in this paper addresses a novel concept for prevention of data leakage for data in transit. The text under consideration is classified to confidential or non-confidential category based on the content and context using Machine Learning technique. Subsequent action for encryption is performed on the confidential data and then transmitted from the Intranet domain to Internet domain ensuring that the data is not compromised to unauthorized users. In addition, normal transactional data which is non-confidential in nature is not prevented from transmitting and is easily accessible by a third party. Encryption is applied only to selected data and not the entire data in transit, ensuring that the hardware resources are efficiently utilized.**

An adversary can simply compose an e-mail with the organization's confidential information as the body of the mail, in such scenarios our method will classify the e-mail content and will encrypt the data so that the data is not revealed to an outsider.

Index Terms*—Text Classification, Data Leakage Prevention (DLP), Encryption, Gateway, Intranet, Internet, Decryption, Artificial Intelligence (AI), Machine Learning (ML), Trusted Platform Module (TPM).

I. Introduction

Data leakage is a condition where the confidentiality of the data is compromised. There are several means for data leakage to happen, major causes include 1) Intentional data leakage by an adversary internal to the organization, 2) Data leakage by a person external to the organization but has got temporary access rights to the victim organization's resources, 3) Un-Intentional leakage by Internal user or administrator.

It is said that data is an asset of an organization hence protecting this asset becomes a very important aspect for an organization's growth and development. Since the invention of the internet and e-mail systems, the risk of data leakage in any organization has become a serious issue and immediate action for leakage of confidential data from an organizations domain (The Intranet domain) to the Internet domain is to be addressed.

Data leakage prevention is applicable for data in the following stages:
1) Data in transit
2) Data in use and
3) Data at rest

Some of the techniques that can be used as a counter measure for data leakage detection are:
1) Scrambling: Where the sequence of the data is changed so that the meaning is preserved but the sequence is altered.
2) Perturbation: Is a very well known technique where the confidential content of the document is modified and made less sensitive, here the noise is introduced in the text so that the overall Signal to Noise ratio is reduced and hence the data leakage detection capability comes down.
3) Interleaving: Is a technique where the sequence of occurrence of the words is changed resulting in under weighing the confidentiality of the document.

Traditional approach for data leakage prevention include
1) Watermarking: Where in a unique code is embedded in each document which is confidential,
2) Finger Printing: Here the documents or the text is represented as a set of strings and a hash value is generated. Hash value's for each document is generated in a sliding window method, Now that the Database is available, each document is analyzed and its hash values are then compared against the hash values in the database, if sufficient number of matches in the hash values are found then the document is considered confidential.

In this paper a novel concept for DLP for data in transit is introduced, accomplished using techniques of Machine Learning employing the content and context based text classification and later employing the encryption technique to preserve the confidentiality.

Although the introduction of Data Diode was one of the recent contribution towards DLP for data in transit, in this case the data from the Secure Domain (Intranet Domain) to the Un-secure domain (Internet domain) is completely blocked. For scenarios where a user wants to send some data from the Intranet domain to the Internet domain, though not confidential but required for the proceedings of the project, the user is unable to send any information to even his peers which ultimately results in an adverse impact on the progress of the project and hence affects the business prospects of the organization and the reputation of the organization. In such scenario's our approach will be ideally suitable, where data

is allowed to flow from Intranet to Internet domain but after classifying the data and then either encrypting it (if its found confidential) or sending as it is (if it is found non-confidential).

II. Related Work

In this section, we discuss the previous work on clustering and classification of texts/documents, various approaches are used eg: attribute reduction method, graph representation of texts/documents is presented.

A. Clustering Method for Text Data

1) Rule-Based: In rule based approach various rules/policies are defined for the terms that appear in the text/document [5], these rules are based on the user behavior and transactions, these policies may differ from employee to employee based on his/her role and designation, this technique is presented in various studies [7] [8] and it is also implemented in several products.

The main drawback of Rule Based approach is the high rate of False-Positives(FP's) and also the time incurred in defining a new policy for scenarios where the roles of a user are frequently updating, since the rate of FP's is very high this approach is not being used nowadays. In comparison to rule based approach, Role based approach defines policies and these policies are then applied based on the role of an employee in an organization.

2) Vector-Based: Vector Space Model(VSM) is a Vector based method, this approach was first presented by Salton in 1960. In VSM model, each document is represented in terms of vectors defined by each term and its corresponding weight [1], i.e each term is categorized into a vector space [4], and for each document the angle cosine value is measured, [2] now the classification is done by calculating the similarity between the different documents under consideration [3]. Each document D is represented as $D = D((T_1, W_1), (T_2, W_2), ...(T_n, W_n))$, where n is the total number of terms in the document D, T_i is the i^{th} term, and W_i is its corresponding weight in the document.

3) Frequency Based: Term Frequency and Inverse Document Frequency(TF-IDF) is a more often used technique where the importance of the term is directly proportional to the frequency of occurrence of the term in a document and is inversely proportional to the frequency of occurrence of the term in the whole document. This technique is been widely used in several fields such as Data Mining, Text Classification etc. [10]

4) Graph-Based: In Graph based representation, the text is represented in the form of graph instead of the existing models, one main advantage is this model can not only capture the contents of the text/document but it can also correlate the terms along with its context, this technique is mostly applicable for text related tasks [9]. A similar approach where the DLP model is implemented based on not only the confidential keywords, but also the context terms is called CBDLP(Context Based Data Leakage Prevention) [11] where, the graph structure with terms and their context is adapted to indicate documents of the same class and then the confidentiality score is calculated and the document is classified as confidential or non confidential.

Datasets available in public domain can be used for training the AI Model or a customized dataset suitable for an organization can be created [17] for achieving optimum performance.

B. Data Leakage Prevention

Prevention of disclosure of data to unauthorized users is the main goal of Data Leakage Prevention and the information security team. Secret data leakage are liable to cause major financial losses and can also damage the reputation of the organization.

Various efforts have been brought out in order to cater for this serious issue, which include systems like Firewalls, Intrusion Detection System's(IDS's) and/or Intrusion Prevention Systems(IPS's). These systems can perform optimally provided the data to be protected from leakage is constant and structured, but these techniques will not work for data which are unstructured and not constant, to overcome this limitation, DLP system's were introduced. DLP system's are designed such that they have the ability to identify and protect confidential data and it also detects misuse based on rules defined in advance.

The DLP system's are being addressed by various names few of them include Data Loss Prevention, Data Leakage Prevention, Information Loss Prevention etc. [12], [13] has provided a brief survey on DLP's, they describe taxonomy of DLP solutions along with examples. [14] provides a detailed history of the DLP systems available and how they can be deployed in other Secure network technologies.

Now, these DLP system's would not work well if the data to be protected is modified contextually or semantically, thus these systems lack in terms of their ability to prevent leakage of the confidential data.

DLP systems are different from the conventional systems such as Firewalls, VPN's etc, in that the later devices are less dedicated for the task of DLP and are not proactive in operation, they may block a legitimate user from accessing the data as a means of sensitive data protection as seen in Firewalls, or they can simply encrypt all the traffic, as in a VPN, whereas Anomaly based IDS's can be triggered proactively they focus mainly on meta data such as size, source and destination etc. and not the sensitive content of the text. Content based DLP's are more common and are generally preferable when compared to context based DLP'S since they are more logical as they focus on the confidential terms of the Document under consideration [12]. A cloud computing environment employing SEDoS-7 [18] can very well be used along with the DLP systems.

Few more method's used for DLP are the fingerprinting method(Hash based) and piece wise hashing method [16] that are developed especially to detect the data that is evolved, these techniques will not be able to identify the sensitive data if the data is altered in terms of its sequence of appearance or if the data is modified with the semantics in tact, fingerprinting which in the background uses the Hashing technique such as

the MD5 or SHA are most likely to change, any small change in the document/text will result in complete change of the hash value [15]. This problem of hashing can be solved partially by creating chunks of the data and then taking hash value of the chunk of data separately [6]. However, these hash values for the smaller data are also susceptible to change for even a small change in the data and also the processing required is more therefore it is inappropriate to use.

III. Architecture

We propose a novel architecture **D-SeGATe (Data Leakage Prevention using Secure Gateway Analysis Technique)**.

Our Architecture is applicable for scenarios where there is a separation between the Secure domain(the intranet domain) and the Unsecure domain (The internet domain), The goal is to ensure that the data can easily flow from the internet domain to the intranet domain and vice-versa, the former can be achieved using a data diode approach for cases where data flow is from the internet domain to the intranet domain as seen in Figure 1. where as the latter requires a continuous monitoring environment.

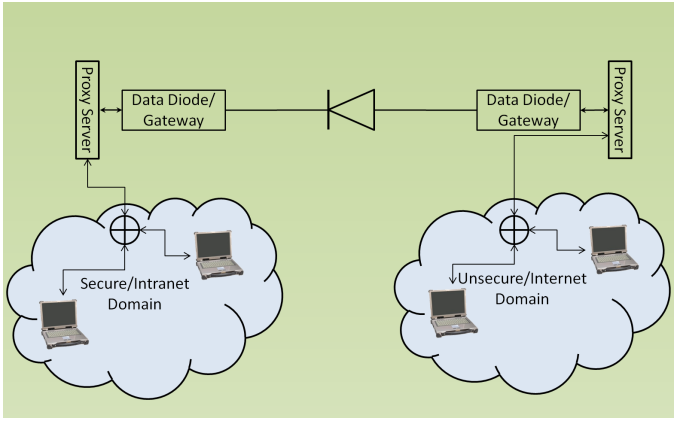

Fig. 1. Example of a Data Diode Deployment and Operation.

The data flow in the reverse direction should happen within a closed monitored environment, that is, every data that is passing from the intranet to the internet domain should be monitored carefully for its content so as to analyze if it is confidential or not. If the data is found to be confidential then the data will be encrypted and sent, an alert message will be sent to the information security administrator regarding the access of the confidential data. A sample representation of our architecture is as indicated in Figure 2. The figure depicts data flow from the secure domain(intranet domain) to the unsecure domain(internet domain) only.

However data flow from the internet to the intranet domain can take place without any data classification by means of data diode.

A. Elements of the Architecture

1) Intranet Endpoint Devices: These include all the network endpoint devices that can be used to access the data from the intranet network domains, these devices are assumed to have sufficient resources to access the network.

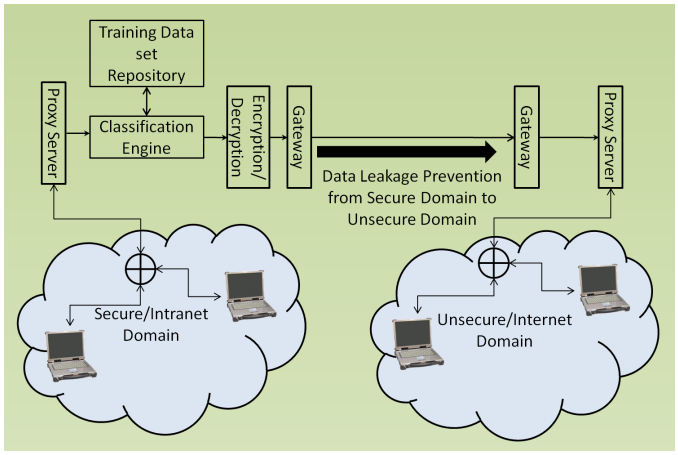

Fig. 2. DLP using AI and Crypto Technique.

2) Internet Endpoint Devices: These include all the network endpoint devices that can be used to access the data from the internet network domains, it is also assumed that these devices are provided with sufficient resources required for accessing a network.

3) Intranet Domain Gateway: This is a common device for all the intranet endpoint devices, all the data traffic flowing outside the intranet domain passes through this device,

4) Internet Domain Gateway: This is a common device for all the internet endpoint devices, all the data traffic flowing outside the internet domain passes through this device.

B. Data Leakage Prevention Approach

Following are two possibilities for data that is tried to be uploaded, Figure 3 shows flow Diagram for **D-SeGATe-Tx** (upload use case).

1) Confidential Data: With the document being confidential following procedure will take place,

The classification takes place at the gateway of the transmitter (i.e at the Secure domain/Intranet domain), all the data (including documents, texts, images) is passed through a ML based classifier loaded with all training data-set repository at the gateway, the classification module is trained using the training data-set repository and only the testing data is provided to the classification module, the classification module, a) Analyses data, b) Does all pre-processing, c) Classifies the testing data. It is to be noted that the classification module is provided with all forms of data(text, image, document) coming out of the gateway, so that even if the adversary tries to compose an e-mail with confidential content, the respective text is analyzed by the engine and classified as confidential.

Now that the data is classified as confidential, it is encrypted using a common key shared only to the intranet end devices(key sharing module is explained in the next section), The receiver receives the encrypted data over internet domain, the user by himself will not be able to read back the data in the

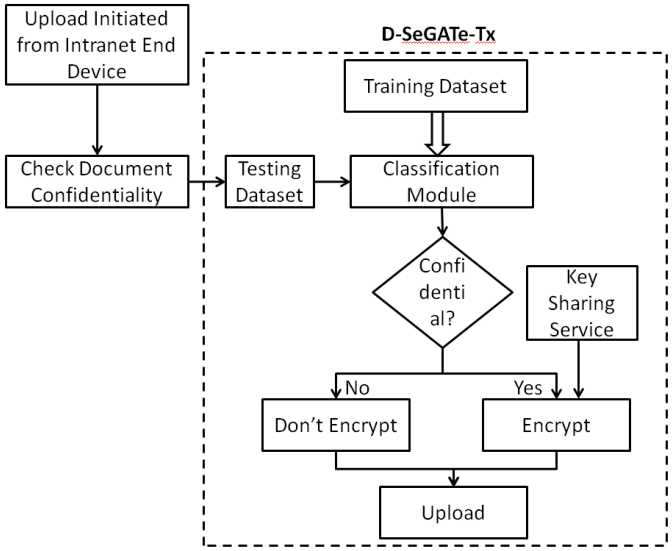

Fig. 3. Upload Implementation Flow Chart

internet domain since the data is encrypted and the key is not shared to any of the internet end devices, the user then sends data to the intranet domain, now once the data is transferred to the intranet domain the user can easily decrypt the data (since the common key is shared among all the intranet domain devices) and can read it back. (Figure 4 shows the flow of data)

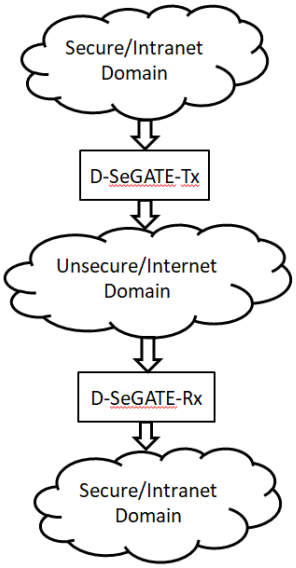

Fig. 4. Data flow in a D-SeGATe System

Above scenario is best suited for an incident where a confidential data of the organization is tried to be leaked, our architecture would encrypt the data, so that only authorized persons are allowed to read it back after decryption, Note that the confidential data is not simply blocked so that even if an authorized user wants to send some confidential data to his peer who is using the same Intranet domain but is physically isolated and the isolated intranet domains are connected by an internet backbone, the peer user is allowed to access the confidential data without any obstruction, provided the machine in which the data(document/text/image file) is accessed is an authorized intranet end device.

2) Non Confidential Data: With the document being non confidential following procedure will take place

The classification of data is carried out at the gateway of the transmitter, the document/text/image is passed through the ML based classification module, the text/document is analyzed for confidential data in terms of content and context. The document is classified as non confidential, and is allowed to pass through without encryption, now the user at the receiver end receives the data in a plain form and can easily access the data even in the internet domain(unsecure domain).

Above scenario is applicable when an employee wants to send some queries to an external vendor or respond to his queries, since this is a non confidential data and also since such queries need be responded on a real time basis (as the reputation of the organization is at stake)our approach provides an efficient solution. Our architecture will not introduce any delay in the complete process and therefore can be used in any organization dealing with confidential as well as normal transactional data.

C. Common Key Sharing

All intranet domain end devices are provided with a common key for encryption and decryption of data that is classified as confidential, the common key would be generated by a central machine called the key management server.This server will generate a key by using a TPM Module, the TPM module enables confidentiality of the key, since no body is able to access the internal memory of the TPM and hence the key generated is automatic, truly random, not pre-configured and not predictable. The key management server generates the key and it is distributed to all the devices connected in the intranet domain. The key is distributed in a way that all the Intranet devices are provided with the latest key. Key sharing would make use of asymmetric encryption for key exchange which is the public private key sharing technique for each of the end device whose public Key is stored in the central Data base(DB). It is to be noted that the key is shared only among the intranet end devices and no internet end device is aware of the decryption key.

Sharing Interval: The Information security administrator will define the interval for which the key should be changed, the interval depends on the organization and the level of confidential data it handles. Generally, an Interval of 1 month is acceptable.

IV. Implementation

A. Flow Chart

The implementation flow chart for our approach for data upload is as depicted in figure 3. The user initiates for a

document/text/image to be uploaded from the intranet domain to the internet domain, Data is stored at the gateway, the classification engine at the gateway is activated, the engine classifies the document/text/image. If the document/text/image is found confidential then the document/text/image is encrypted and sent over the network, if the document/text/image is found to be non confidential then the document/text/image is sent as it is over the network.

The implementation flow chart for data to be downloaded is depicted in figure 5.

User in the internet domain downloads the data and tries to read it, if the data is confidential then the user would never read it because it is encrypted and the key is not known, at this stage even if an adversary leaks the confidential data to the internet domain, since the data is encrypted, the data is of no use to the entire domain outside the intranet. The user then sends the data to the intranet domain and requests for data to be downloaded, the meta data of the document/text/image is analyzed, if the data is found to be encrypted then the common key is used to decrypt the data, else the data is read back as it is without decryption.

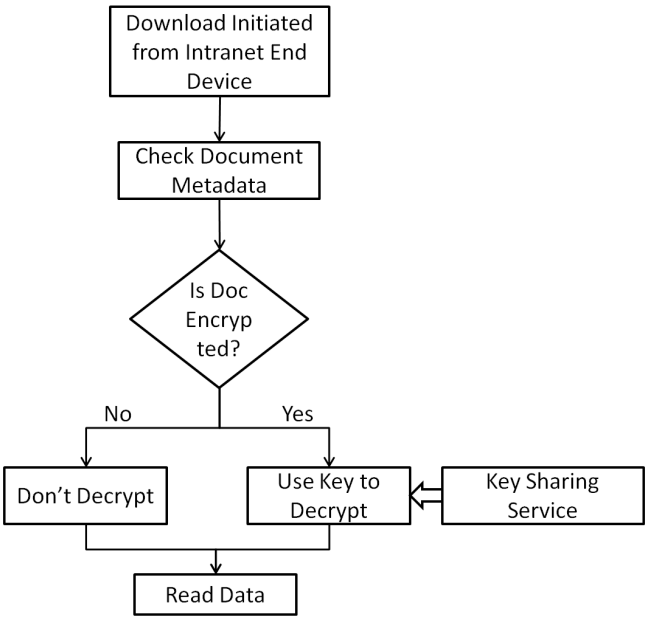

Fig. 5. Download Implementation Flow Chart

B. Algorithm

The details of D-SeGATe for the data transfer mode of operation is brought out in Algorithm 1.

As far as the decryption part of the model is concerned, data would be analyzed first for whether it is encrypted or not. If data is encrypted then the encryption version would be looked for, now this version would be useful for scenarios where data is being downloaded after a long time since it was transmitted, during which the symmetric keys are changed/updated. The version would help in recognizing the correct key that needs to be applied for the data to decrypt.

Result: Plain or Encrypted Data
initialization;
while *Document/Text/Image are Selected* **do**
  check the Confidential Content;
  if *Data is Found Confidential* **then**
    Encrypt it;
    Tag it with the Encryption Version;
  else
    Pass Data as it is;
  end
end

Algorithm 1: DLP for Data in Transit

C. Classification Model

The classification model consists of the training stage and the testing stage, the training phase includes clustering and graph building. The documents/text/images are first categorizes into different clusters, and then these clusters are represented by graphs, during the testing phase, the graphs are matched to the documents/text/images, and a confidentiality score is generated. A documents/text/images is considered as confidential if its confidential score is exceeding some predefined threshold value, otherwise the documents/text/images are considered non confidential.

V. Conclusion and Future Work

In summary, we have proposed D-SeGATe, A novel architecture for DLP for data in transit is introduced, this technique is applicable to leakage of data by means of documents/text/images. This architecture also utilizes all necessary classification functions such as, data pre-processing, feature extraction, text classification and accordingly sending the data with or without encryption. Our approach reduces the risk of confidential data leakage outside the organization without affecting the business life cycle of the project.

In addition, our method permits confidential data to flow in a confined environment so that only authorized person accesses the confidential Data. Also since the entire data coming out of the intranet domain is not encrypted, the hardware resources are not unnecessarily burdened.

Future work: In future, we will aim for the experimental results of our proposed Data leakage prevention solution for a system designed for an organizational point of view. We wish to explore Data leakage prevention for data in different states (i.e data at rest, data in use,encrypted/compressed/zipped data). We will also explore DLP solution for Hand held endpoint devices such as tab, smart phone etc.

Acknowledgment

The author wish to acknowledge Mr. Mahesh V, Director R&D, BEL; Mr. Manoj Jain, GM, BEL, Bangalore; Mr. Kiran Kumar, GM Mil-Com, BEL; Mr. Shrinivas Mugali V, AGM Network and Cyber Security, BEL; Mrs Umadevi B, DGM NW&CS, BEL and Mr. Jagan Mohan Rao B, DGM NW&CS, BEL for their continuous support, motivation and guidance.

REFERENCES

[1] K. R. Rao and P.Yip, Discrete Cosine Transform: Algorithms, Advantages, Applications, Academic Press, Boston, Mass, USA, 1990.

[2] S. M. Katz, "Estimation of probabilities from the sparse data for the language model component of a speech recognizer," IEEE Transactions on Signal Processing, vol. 35, no. 3, pp. 400-401, 1987.

[3] R. Jin, L. Si, A. G. Hauptmann, and J. Callan, "Language model for IR using collection information," in Proceedings of the 25th annual international ACM SIGIR conference, pp. 419-420, 2002.

[4] W. W. Cohen, "Learning rules that classify e-mail," in Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access,pp.18–25,1996.

[5] J.I. Helfman, C.L. Isbell "Immediate Identification of Important Information", AT and T Labs Technical Report, 1995

[6] Kantor et, al. "Methods for document-to-template matching for data-leak prevention", 2009.

[7] J. Staddon, P. Golle, et al, A content-driven access control system, in: Proceedings of the 7th Symposium on Identity and Trust on the Internet, ACM, Gaithersburg, Marylands, 2008, pp. 26–35.

[8] J.D.M. Rennie "ifile: An Application of Machine Learning toE-Mail Filtering" Proceeding of the KDD Workshop on Text Mining, 2000

[9] Gilad Katz, Y.Elovici and B.Shapira, "Cobanacontextbased model for data leakage prevention," Information Sciences, vol.262, pp. 137–158, 2014.

[10] J Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries" Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 2003.

[11] Xiang Yu, "A Data Leakage Prevention Method Based on the Reduction of Confidential and Context Terms for Smart Mobile Devices", School of Electronics and Information Engineering, Taizhou University, Taizhou 318000, China.

[12] Mogull R, "Understanding and selecting a data loss prevention solution",2010.

[13] Shabtai A, "A survey of data leakage detection and prevention Solutions", Springer: 2012.

[14] Kanagasingham P, "Data loss prevention", SANS Institute, Information Security Reading Room, 2015.

[15] Saphira et al, "Content-based data leakage detection using extended fingerprinting", Ben-Gurion University of the Negev, Israel, 2013.

[16] Kornblum J. "Identifying almost identical files using context triggered piecewise hashing" Digit Investig 2006.

[17] M S Girija Devi. "Scarce Attack Datasets and Experimental Dataset Generation". Proceedings of the 2nd International conference on Electronics, Communication and Technology (ICECA 2018).

[18] R. Gopeshwar Rao, "SEDoS-7 A Proactive Mitigation Approach Against EDoS Attacks in Cloud Computing". IEEE WiSPNET 2017 conference.