ISSN 2278-3091

Volume 9, No.3, May - June 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse03932020.pdf

https://doi.org/10.30534/ijatcse/2020/03932020

Web Intrusion Detection System through Crawler’s Event Analysis

S. Ponmaniraj[1], Dr. Tapas Kumar[2], Dr. Amit Kumar Goel[3]

[1]Research Scholar, [2,3]Professor

[1,2,3]School of Computing Science and Engineering, Galgotias University, Greater Noida

ponmaniraj@gmail.com

ABSTRACT

In this digital era, Reliability of a specific web page and its referral linking address or sites are questionable due to its security parameters. Though user maintains various protecting software and highly configured firewall systems against attack still some kind of assaults are happening at communication devices between sender and receiver. Hackers are deploying their molest code in the form of intrusions and it causes to attacks such as spoofing, Denial of Service (DoS) etc. Identifying those intrusion attacks are not limited to analyzing few parameters considerations. Rapid development of new tools and technologies, hackers are intruding targeted victim’s systems to steal/snoop their sensible data. Security breaches of the targeted system arises through intruding victim’s system by IP tracing and spoofing. Intrusions are obtaining on the following categories, Stand alone system, Networks and Web oriented intrusions. Many research activities are already taken into the place to identify causes for intrusions and preventive methods. There are more possibilities on web oriented intrusions due to web sites and their referral links are coming from number of intermediate agents. This research paper is going to discuss about Intrusion Detection System (IDS), Web Intrusion Detection System (WIDS), Proposed system architecture and implementation of WIDS.

Key words: Anomaly Detection, Bayesian Classifier, Data Scraping, Intrusion Detection System, IP Spoofing, Web IDS.

1. INTRODUCTION

Intrusion is the mechanism of attacking the targeted victim’s system and legal data. Sometimes attacks are in the passive manner, which means hacker simply tries to listen to the communication in between end users. There is no harm in this type of attacks. On the other hand, active types of attacks are harmful to the user system and data. Intrusion causes to both of those attacking types. IP spoofing is the example for passive mode of intrusions and Denial of Service (DoS) is an example for active mode intrusion attacks. DoS leads to crash the system by rising continuous request to the targeted systems. Due to continuous requests system losses its resources for respond and also this requests will make system to lose the data which are in queue [1].

The w3 consortium announced that year by year web site attacking is getting increased due to intrusions. Vulnerabilities on security parameters such as Confidentiality, Integrity and Availability (CIA) are the important issues in web intrusions. Virtual machines are comes into the concept for maintaining cloud services to provide better performances. Cloud and virtual machines leads to connect many intermediate agent systems for requests and responses. From those intermediate communications, hackers can intrude the targeted systems [2].

Intrusions attacks are giving more loads on web servers and due to heavy loads easily servers are getting shut down or crashed with mandatory applications. Since cloud is implemented on most of the computer applications and servers, effective load balancing algorithms are needed to avoid heavy loads of intrusions kind of attacks [3].

TCP/IP protocols policies are simply looking for the matching rules and most of those rules replicated in firewall implementations. In general these rules are doing the tasks as follows; 1. Session creations, 2. Log files creations, 3. Session terminations, 4. Data flows acceptance or rejections etc [4]. Since numerous fragmented data are passing through on communication mediums, Intrusions are easily happens at IP fragmentation functionalities.

TCP/IP port 139 and port 445 are the vulnerable ports to do attack by the hackers. Commonly attackers scan for the above said ports to intrude inside the system for collection of system details such as user id, computer name, LAN/WAN addresses, private IP address etc. Port numbers 139 is “NBT over IP” and Port Number 445 is “SMB over IP”. Many corporations and organizations are configuring their firewall to protect networks and systems from IP spoofing attacks based on the information received from TCP/IP ports 139 and 445 [5].

2. INTRUSION DETECTION SYSTEM – OVERVIEW

In the earlier attacking methods hackers tried with spam texts and spam emails to evade the victim’s sensitive data. After the invention of pattern analysis and matching system the text oriented attacking methods are came into control. Later hackers used images to imbed their malicious codes to attack targeted systems security functions. The malfunctioned images are called as spam images and these spam images are spread into the victim's web pages through broadcast mechanisms. Once the user clicks on these images automatically molest contents are starting to execute or it takes into some other vulnerable websites or do IP spoofing to perform eavesdrop [6, 7]. To avoid spam image attacks many research had been taking into the place to identify spam contents embedded in images. Optical character recognition (OCR) tools implemented to read spam contents from images and it is then applied to text pattern analysis to compare spam texts indulged in images [8]. Intrusion detection system is the mechanism for creating alert or acknowledging central security system or the administrator about vulnerabilities going to be happening at system and its security functions. It's kind of malicious or violation of privacy activities against victim's system and data. Intrusions happening at three levels as follows;

1. Host Intrusion Detection System (HIDS)
2. Network Intrusion Detection System (NIDS)
3. Web Intrusion Detection System (WIDS)

Host Intrusion Detection System (HIDS): This intrusion detection method concentrates on separate/stand alone host system. In which all the log files, kernel files, monitoring of system files and system connection based analysis [9]. Once the targeted system is intruding with malicious codes then automatically the attacking functions are seeking for the above mentioned file system. Network Intrusion Detection System (NIDS): In this method, attacking codes are focusing on the communication links. Hackers are implementing their molest functions through IP address of the network's connected systems. More than one system is connected together via connecting mediums such as modem, switch and router etc., by an IP addresses. Through these mediums, Hacker performs IP Spoofing to trace the communication links and then intrude targeted systems to hew the victim's legal data [10].

Web Intrusion Detection System (WIDS): Web intrusions are happening in the path of crawling progresses. After entered search phrase by the user, web bots are moving towards the server systems to bring the requested page as responses. Before moving to the server, all the requests are bypassing the intermediate agents to connect with server.

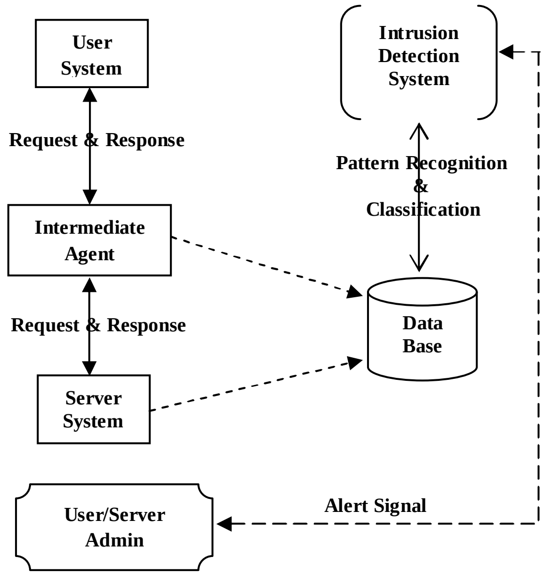

Figure 1: General Structure of Intrusion Detection System

Meanwhile numerous hacking pages are included with original contents as visited and visiting links and reference pages by image or button forms. Once the user clicks those pages, images or the links then automatically hacker will intrude the targeted system with the help of malfunctioned connectivity [11].

All the intrusion attacks are looking for the system files, kernel files and connecting files to corrupt or strip the system information. This corruption obtained in the following ways;

1. Anonymous
2. Misuse
3. Penetration
4. Leakages etc.,

Fig.1 shows that model structure for Intrusion Detection System (IDS). In which, requests and responses are passing from user to server and vice versa through intermediate agent. All the flowing data are matched with databases for pattern checking to analyze the reliability. If any vulnerable pattern found then alert messages will be passing to user/server system administrators.

3. WEB INTRUSION DETECTION SYSTEM (WIDS)

Web intrusion attacks are ensuing on the web applications. Web search engines or Web bots are accessing many websites based on the given input key phrases. All those key phrases are processed on the web crawler's principles and applications. While fetching those web pages, some illegal web documents are also moving to the targeted victims system via visiting links, images and referral links. Web intrusion leads to perform the following type of attacks [12].

1. Cross site scripting
2. SQL injection
3. Session hijacking
4. Remote access
5. HTTP violations etc.,

Web intrusion detection system performs analysis and detection of malware functions and abnormal activities on the web site communications. WIDS generates alert signal to the standalone system and web administrators to take up necessary remedial actions against above said attacks. IDS won't protect the system from attacks instead it gives the warning/alert signals to the users. IDS will monitor the following file system for detecting the molest functionalities [13].

1. Monitoring of user and system activities
2. Auditing of configured files and vulnerabilities
3. Monitoring the integrity of a files systems and critical file systems
4. Pattern analysis on statistical activities
5. Abnormal activities analysis
6. Operating system file analysis
7. Port analysis etc.,

The above said points are the mandatory places to monitor often to check for attacks.

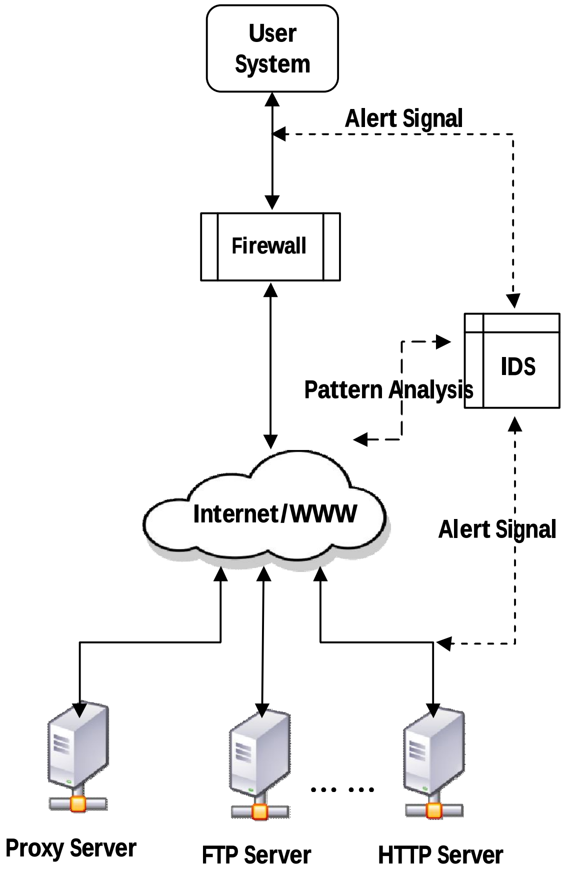

Figure 2: WIDS Model Structure

Fig. 2 shows that model architecture of web intrusion detection system. In the above model, IDS system is working in between firewall system and user system or server system to generate alert signals.

WIDS functioning on the basis of error codes returned from visited web site links in the following ways;

1. Domain Name System (DNS) Error
2. Server Side Error
3. Web Crawler Failures

The above said error messages are generated on the basis error codes of the web pages while crawling the servers and service providers. Every service providers has error codes for the concerned activities such as Not connected, Not Found, Data fragmentation error, Internal/External server error etc. Table 1, shows that some examples for error code values and related error status of the web pages [14].

Table 1: Web Error Status Code and Status Messages

S. No	Status Code	Status Messages
1	Code:0	Connection Time out / Connection Refused / Connection Error / No Response / DNS Lookup Failed
2	Code:301	Moved Permanently
3	Code:302	Moved Temporarily
4	Code:400	Bad Request
5	Code:403	Forbidden
6	Code:404	Page Not Found
7	Code:410	Page Removed
8	Code:429	Too many requested
9	Code:500	Internal Server Error

4. IMPLEMENTATION OF WIDS

WIDS is working on the basis of TCP/IP parameters and its rule violations. Data patterns, classifications, timing of responses, acknowledgements, number of requests at the moment, heavy loads etc are the important parameters to analyze the reliability of a web documents while crawling for multiple pages. Implementation of WIDS is focusing on those criteria's for enhanced result.

Implementation of WIDS

★ Seed URL Key Phrase (SU)

★ Search for Web Page (WP)

★ Retrieve Web Pages for Reference Links (WP(L_i))

$$\boldsymbol{Web_{Page}(SU) = WP(L1) + WP(L2) + \cdots + WP(L_{n-1})} \quad (1)$$

★ Parse Web Page code to extract Outbound URL links (WP(L_i))

$$\boldsymbol{Web_{Page}(SU) = \frac{WP(L1)}{OL(L1)} + \frac{WP(L2)}{OL(L2)} + \cdots + \frac{WP(L_{n-1})}{OL(L_{n-1})}} \quad (2)$$

★ Check for Web Page Freshness/Age to assign score and rank value for concern pages,

$$\boldsymbol{Page_{Rank} = IAV + \sum_{PC=0}^{N} \frac{PRV}{No.of\ OL}} \quad (3)$$

Where,

- IAV = Initial Age Value,
- PC = Number of Pages Visited,
- PRV = Page Rank Value,
- OL = Outbound Links

★ Standardize recorded data set for classification (SD)

$$\boldsymbol{SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}} \quad \textbf{(4)}$$

Where,
N=Number of data values
(x_i– x ’) = Mean value for the previous data values

★ Perform linear separation of data set,

$$\boldsymbol{W^T x + b \geq 1 \ and\ W^T x + b \leq 1} \quad (5)$$

Where,
W = Weight (vector value)

```
        b = Bias value (Scalar value)
★   Look for Misclassification Error,
        $M_{err} = min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi i$        (6)
        Where,
        $min\frac{1}{2}\|w\|^2$
        $= Hypher\ plane\ for\ Linear\ Classification$
        $\xi = Slack\ Variable$
★   If (Outbound links==existed) then
        return OL as Web Page
★   Else update index values for Outbound Links
★   While(Visited OL==Null)
        Update the index checking process
```

5. RESULT AND DISCUSSION

WIDS for the above implementation is working for finding the index values of the every visited web pages then it checks for the security and reliability of the visited page contents. The reliability of a web links are processed on the basis of data patterns from block list and white list[15]. KDD99 dataset is specialized for intrusion detection mechanisms. Above said implementation also working with the help of KDD99 dataset and it yields the accuracy of more than 90% based on security parameters.

Table.2 shows that sample outcome of the above said implementation for few web links. Reliability of outbound link is calculated from the security parameters of TCP/IP protocols, throughput, response times and data loads during communications. Security of a web link processed in three different places such as total security, domain security and transport level security.

WIDS's performance is evaluated by three different categories such as Accuracy, Precision and Recall values [16, 17]. Accuracy is method to find the closeness of the given input data set to existed one. In other hand, error of any measurement is the accuracy of that erroneous measurement.

Error measurement is derived as follows;

$$Err = Observed\ Value - True\ Value \tag{7}$$

And,

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum True\ Positive} \tag{8}$$

Precision is the measurement used to find the spread or range of the results occurred. These parameters are calculated as follows;

$$Precision = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Positive} \tag{9}$$

Recall is the process of retrieval of successful data. In the other hand recall is known as sensitive data and it is the probability of the relevant document retrieved from the given query. Recall is derived as follows;

$$Recall = \frac{True\ Positive}{Total\ Actual\ Positive} \tag{10}$$

Fig. 3, shows that reliability of a web link and the time duration taken for responses from the servers and/or intermediate agents.

Table 2: Reliability and Security level of given in

Web Links	Input	Server	Duration	Transport_ security	Total_ Security	Domain_ Security	TLS	Reliability
L1	https://trypap.com/	Netlify	0.5	404 test page	B	B	A	81.67
L2	http://ihasbucket.com/	NginX	1.5	no_redirect_to _https	C	A	C	48.33
L3	http://hasthelargehadroncol liderdestroyedtheworldyet. com/	Apache	0.375	Unable to connect	C	A	C	48.33
L4	https://chrismckenzie.com/	Nginx	0.625	x_xss_protecti on	B	A	A	88.33
L5	http://endless.horse/	Apache	0.5	TLS certificate does not match the host name	C	A	C	48.33
L6	https://theuselessweb.com/	Netlify	0.375	no_stapled_oc sp	B	A	A	88.33
L7	http://tinytuba.com/	AmazonS 3	5.125	Timeout reached	C	A	C	48.33

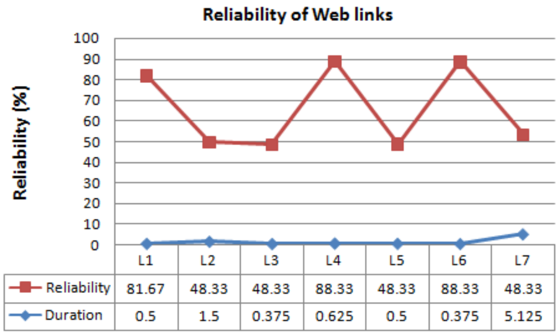

Figure 3: Reliability of Web links

6. CONCLUSION

New technologies and modern tools for attacking the system security for the web document is not limited with few identified parameters. Intrusions are always happens through many hidden loopholes. This research took TCP/IP protocol parameters and time duration for responding to the query based on web error code status of visited and referral links. In future this WIDS mechanism will be applying to alert the victim from unknown attacks with more security parameters.

REFERENCES

1. Liu, W. (2009). **Research on DoS Attack and Detection Programming**. 2009 Third International Symposium on Intelligent Information Technology Application. doi:10.1109/iita.2009.165
2. Liao, H.-J., Richard Lin, C.-H., Lin, Y.-C., & Tung, K.-Y. (2013). **Intrusion detection system: A comprehensive review**. Journal of Network and Computer Applications, 36(1), 16–24. doi:10.1016/j.jnca.2012.09.004
3. A. Arul Prakash, V. Arul, A. Jagannathan, **A Look at of Efficient and more Suitable Load Balancing Algorithms in Cloud Computing**, IJERCSE, ISSN (Online) 2394-2320, Vol 5, Issue 4, April 2018
4. Brad Woodberg, Mohan Krishnamurthy et.al., **Configuring Juniper Networks NetScreen & SSG Firewalls**, Chapter 1 - Networking, Security, and the Firewall, https://doi.org/10.1016/B978-159749118-1/50003-4
5. https://www.thewindowsclub.com/smb-port-what-is-port-445-port-139-used-for
6. Cheng, H., Yan, X., Han, J., & Hsu, C.-W. (2007). **Discriminative Frequent Pattern Analysis for Effective Classification**. 2007 IEEE 23rd International Conference on Data Engineering. doi:10.1109/icde.2007.367917
7. Mohammadi Akheela Khanum, Lamia Mohammed Ketari, **Trends in Combating Image Spam E-mails**, ICFIT 2012, arXiv:1212.1763
8. Biggio, B., Fumera, G., Pillai, I., & Roli, F. (2011). **A survey and experimental evaluation of image spam filtering techniques**. Pattern Recognition Letters, 32(10), 1436–1446. doi:10.1016/j.patrec.2011.03.022
9. Kopelo Letou, Dhruwajita Devi, Y. Jayanta Singh, **Host-based Intrusion Detection and Prevention System (HIDPS)**, IJCA, ISSN:975 – 8887, Volume 69– No.26, May 2013
10. Muhammad K. Asif, Talha A. Khan,Talha A. Taj, Umar Naeem, Sufyan Yakoob, **Network Intrusion Detection and its Strategic Importance**, IEEE Business Engineering and Industrial Applications Colloquium (BEIAC), 2013
11. Shu Wenhui, & Tan, T. D. H. (n.d.). **A novel intrusion detection system model for securing web-based database systems**. 25th Annual International Computer Software and Applications Conference. COMPSAC 2001. doi:10.1109/cmpsac.2001.960624
12. N.Sakthipriya, K.Palanivel, **Intrusion Detection for Web Application: An Analysis**, IJSER, ISSN 2229-5518, Volume 4, Issue 5, May-2013
13. Christopher Kruegel, Giovanni Vigna, William Robertson, "**A multi model approach to the detection of web based attacks**", Journal of Computer Networks - Elsevier 2005
14. https://www.screamingfrog.co.uk/http-status-codes-when-crawling/
15. Suad Mohammed Othman, Fadl Mutaher Ba-Alwi1, Nabeel T. Alsohybe and Amal Y. Al-Hashida, **Intrusion detection model using machine learning algorithm on Big Data environment**, Jouranl of Big Data, Springer, 2018, DOI:.10.1186/s40537-018-0145-4
16. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
17. https://en.wikipedia.org/wiki/Precision_and_recall
18. B.Pandu Ranga Raju, B.Vijaya Lakshmi and C.V.Lakshmi Narayana, **Detection of Multi-Class Website URLs Using Machine Learning Algorithms**, IJATCSE, ISSN 2278-3091, DoI: 10.30534/ijatcse/2020/122922020, 1704 – 1712
19. Fredy Fernandus and Nilo Legowo, **The Effect of Website Design, Website Security, Information Quality, and Perceived Ease of Use on Customer Satisfaction and Online Purchase Intention in Indonesia E-Commerce in Jakarta**, International Journal of Advanced Trends in Computer Science and Engineering, ISSN 2278-3091, Volume 9 No.2, March -April 2020