Semantic Architecture for Web application Security

Abdul Razzaq, Ali Hur, H Farooq Ahmad, Muddassar Masood

School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology, Islamabad, Pakistan

Abstract

Growth of web applications has facilitated the humanity almost in all aspects of life especially e-health, e-business and e-communication but this application are exposed for web attacks, unauthorized access, evil intentions and treacherous engagements. Various strategies have been formulated over a period of time in the form of intrusion detection system, encryption devices, and firewalls but still proved to be ineffective. In this paper, we have proposed a system having semantic architecture that is capable of performing detection semantically in the context of HTTP protocol, the data, and the target application. The knowledgebase of the system is the ontological representation of communication protocol, attacks data and the application profile that can be refined and expanded over time. Unlike traditional signature base approach, the semantic architecture analysis the HTTP request with the help of semantic rules and inferred knowledge after reasoning of knowledgebase through Inference engine. Non signature based approach of the system enhance the capability of the system to detect the unknown attacks with low false positive rate. The system is evaluated by comparing with existing open source solutions and showing significant improvement in term of detection ability with low alarm rate.

Keywords: *Semantic architecture, Application security, Semantic security*

1. Introduction

The world trend towards using web is increasing day by day. This increasing dependency on the web has been seen for the last few years. There are 2,095,006,005 users on March 31, 2011, around the globe [1]and Web is not only serving the purpose of allowing access to the wealth of information but is also being used as the most appropriate medium for today's businesses and ecommerce activities. Web being used as the major platform for the flow of sensitive information using sophisticated technologies also exposed to illegal activities committed via these technologies and increasing security concerns for the organizations as well as for the individuals. According to various authentic sources like (MITRE, OWASP, WHITE HAT, ACUNETIX) their survey shows that about 75 % of information security attacks are being launched on application layer and growing exponentially with sophisticated methodologies to launch polymorphic attacks.

In 2010, Symantec encountered more than 286 million unique variants of malware. Symantec recorded more vulnerability in 2010 than in any previous year since starting this report. Furthermore, the new vendors affected by vulnerability rose to 1,914, a 161% increase over the prior year [2] same year. A detailed look propagation mechanism is given in figure 1:

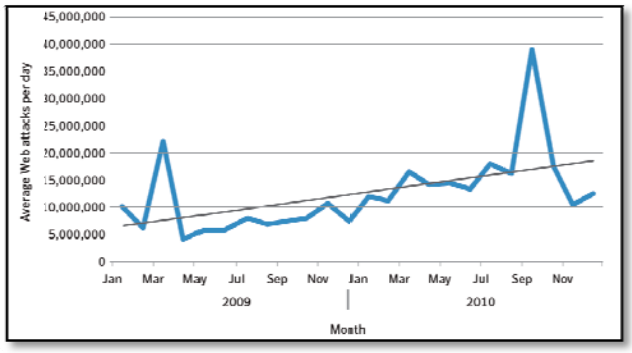

Fig 1. Average attacks per day by month (2009-2010)- Symantec

According to White Hat report winter 2011[3], during 2010, the average website had 230 serious vulnerabilities and most websites were exposed to at least one serious vulnerability every day of 2010. Only 16% of websites were vulnerable less than 30 days of the year overall.
The application layer has most of the vulnerabilities because it has been deprived of the research focus and the heterogeneity of emerging technologies at

application layer making the application security more complex. There are various technologies from different vendors for implementing same standards, e.g. CGI common gateway interface is the standard mechanism for specifying the working of dynamic web application but different technologies like ASP and ASP.net from Microsoft, JSP from Sun Java, AJAX, PHP to name a few exists to implement the same in different ways and hence result in increasing complexities and increasing security concerns. Different security techniques at the application layer have been deployed but are proved to be useless.

The most of the conventional techniques in the form of intrusion detection systems, encryption devices, and firewalls are found ineffective for securing web applications. The proposed system has semantic architected that is specifically designed for web applications. In order to mitigate application level attacks the system is intelligent enough to understand the context of the content and is able to filter that content on the basis of its consequences onto the target application. The information only regarding attacks could not achieve the purpose. The application's context and the contextual understanding of the data are also required. The proposed system can capture the context of attacks, the target application and the underlying protocol used. Capturing the context of the domain and the relationship between the entity bodies enable the system to reason and generate more rules and assertions.
The proposed system has been implemented by using the Resource Description Framework Schema (RDFS) [4] with OWL-DL, JENA framework [5], and Protégé [6] as ontology development tool. For inference upon knowledgebase Pallet [7] reasoned has been used. For system development environment JAVA Net Bean has been used.

This paper is organized as follows. In section II, related work pertaining to security architecture, the usage of ontology in information security and various application security solutions has been discussed. The section III, cover the system architecture and in section IV, the evaluation results of the system has been shown. Finally section V concludes the paper.

2. Related Work

Various security mechanisms in the form of firewalls, anomaly detection systems and intrusion detection systems have been deployed and various security architectures have been proposed but are ineffective in providing a solution at application level. Brief overview of existing solutions and their issues are highlighted in this section.
A Security Architecture [8, 9] has highlighted the issues of a significant gap between the speed with which companies are moving, and the speed of IT security. These emphases the enterprise security with centralized policy but only provides the basic general guidelines and lacks the technical aspects.

Signature-based systems have the signatures of known attacks and apply pattern matching algorithms [10] for attack detection. Signature-based detection may be performed at the network level and also at the application level by analyzing network traffic. The signature-based Intrusion Detection System (IDS) are relatively ineffective in mitigating zero day attacks. Due to a perpetual change in the variety of attacks, signature-based IDSs face the challenge of continually updating the database of signature rules, which is hectic and time consuming. Bro [www.bro-ids.org] and Snort are well-known and popular network IDS, but have a limitation on their effectiveness of analyzing encrypted communications (e.g., HTTPS traffic) [11]. In these IDSs it is not possible to write rule on outbound traffic. Context-based application IDSs [12] face the problem that the signatures have to be manually formulated and they also lack effectiveness against zero day attacks.

Application level anomaly-based IDS systems are more effective in detection of novel attacks as compared to signature-based IDSs, but are less efficient due to a significantly higher rate of false positives [13]. Data Mining Methods for Anomaly Detection [14] provide the framework for web application attacks based on the statistical techniques. However, this framework lacks semantic knowledge needed to analyze the malicious payload on contextual basis, and thus fails by assigning equal probabilities to both equal length attack strings and benign strings.
Ontologies-based IDS solutions are used in information security. Raskin et al. [15] developed the ontology for data integrity of web recourses and advocate the use of ontologies for information security. Landwehr et al. [16] present a categorized taxonomy of intrusion according to location, means and genesis. Ning et al. [17] considered a hierarchical model for the specifications of attacks and modeled the thorough examination of attack characteristics and attributes. McHugh [18] focused on the attacks classification according to the protocol layer and Guha [19] emphasized upon the analysis of each layer of the TCP/IP protocol stack to serve as the foundation for attack taxonomy. Undercoffer et al. [37] J. Undercoffer, J. Pinkston, A. Joshi and T.

Finin, A Target-Centric ontology for intrusion detection, IJCAI Workshop on Ontologies and Distributed Systems, IJCAI'03, August (2003).[20] and [21] have defined the ontology for intrusion detection for network layer attacks. Denker et al [22] drive the control access through ontology developed in DAML+OIL[23] but these ontologies have not been fully utilize due to simple representation of attack attributes thus they are inefficient for intrusion detection. Publication [24] defined system architecture at some level of abstraction but lack the results with some state of the art solution.

Mod Security [30] is an open source, free web application firewall (WAF) that works on Apache system. Mod Security lacks semantics to understand the contextual nature of attack vector and unable to detect, session id brute forcing, authentication brute-forcing and HTML hidden field manipulation attacks [25].

Tammo Krueger et al [26] stated that the internet world comprises of hundreds of heterogeneous application set that includes Information portals, social network sites, blogs, content management systems, e-commerce, web email, groupware have millions of hits per day and thousands of users using them. According to the authors of the paper signature based system have failed to cope with the attack vectors variations that these applications are hit with each day.

3. System Architecture

Semantic architecture is designed keeping in view the complexity of various component involved, and for future expansion. System is built on low coupling and high cohesion principle. The system architecture elaborates the proposed methodology by giving the details of system components.
The system is consists of four major components

1) HTTP Interceptor as an external interface of the system;
2) Logging and Monitoring for audits and learning from the request patterns;
3) Rule-based Analyzer; and
4) Knowledgebase as a managed repository of attack and protocol ontological models.

HTTP interceptor is used for the interception of the web application traffic. Analyzer is the most central component of the system. It further consists of multiple sub-components which are responsible for rule generation, rule caching, and request analysis. Knowledgebase is an ontology which generates rules at run time and it will send it to analyzer which does analysis on the basis of rules generated. Logging module responsible for logging the different HTTP packets which maintain access and denied logs and make the system generate repots according to the administrator needs. Figure 2 as shows the layout and interaction of the system components whereas figure.3 shows the technical view of system architecture.

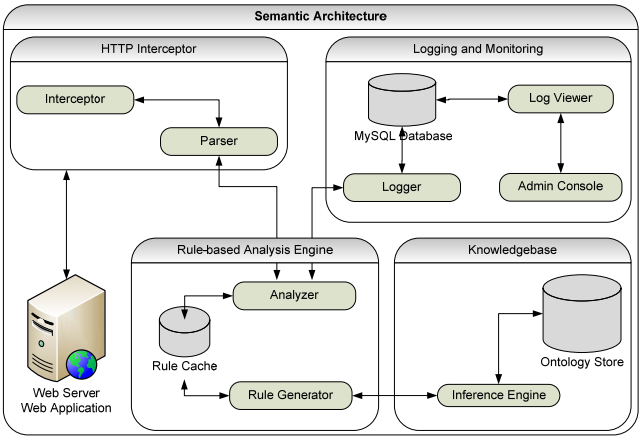

Fig 2. System Architecture

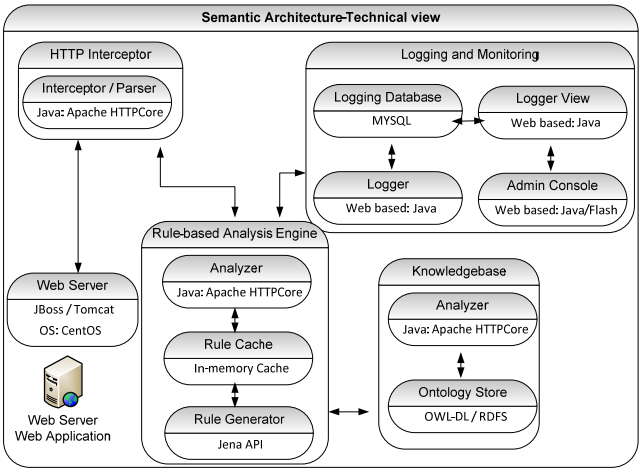

Fig 3. System Architecture- Technical overview

3.1 HTTP Interceptor

HTTP Interceptor intercepts HTTP packets from network stream and then parses it into the different fragments these fragments are send to system analyser for further analysis. HTTP Interceptor is divided into two sub-modules i.e. Interceptor and Parser.

Interceptor operates on port 80 and 443 which is basically used for HTTP and HTTPS traffic. It takes the HTTP traffic from the network stream and sends the packets to the Parser.

Parser used for parsing HTTP/HTTPS packets and then sends it to the analyzer module. It provides support to the analyser module by extracting those parts of HTTP packet that are most probable for attacks. Parser parse the HTTP request in a standardized message format, according to the information (ontology) saved in the knowledge base. It divides HTTP request in three chunks i.e. Body, Query String and Headers. For further analysis parsed contents are sent to the system Analyzer, which checks either the content is malicious or benign.

3.2 Logging and Monitoring

Logging provides the record of events and processes. Each event recorded is entered in log. Logs are monitored by administrator to check changing states, exceptions, unusual behaviour, and other events.

Logging is majorly done for detection of attacks, errors and application misuse. Logs can be used for further analysis e.g. generating run time charts like pie chart, Gantt chart and bar chart. The Logging module is responsible for logging messages or whole session. System maintains access log, access content log, access header log, response log, response header log, infected log, infected header log, and audit log. Access log stores information of each request received by the system and infected log stores information pertaining to the infected messages detected by the analyzer. Audit Log maintains the user log that store information about logged in and logged out of various users. Audit Log also store information about change in configuration settings.

Log store is the central store of storing all the traffic, categorized as infected traffic or access traffic. The log store is build using the open source database MYSQL 5.1. The main reasons of selecting MYSQL 5.1 are that it provides memory cache for fast and efficient data manipulation. It is also efficient performance in terms of parallel processing, bulk data handling and thread concurrency support.

Log viewer is helpful for viewing reports and charts generated by system. Administrator can view the information stored in log in the form of charts like pie, Gantt, bar chart and in the form of tables.

The Administrative Console provides an interface to the administrator. The administrator can configure the system or write commands. All the alerts are also displayed on admin console. This console is the presentation layer of the reporting and monitoring module.

3.3 Rule-based Analysis Engine

The rule based analysis engine is main component of the system which is further sub divided into analyser, rule cache and rule generator.

The Analyzer analyzes the HTTP packet thoroughly and makes a decision whether data is malicious or benign by using the rules generated by the rule generator. On the basis of rules it differentiates between valid and invalid request. The decision will be based on inferred data after the inference process upon knowledgebase. If the request is benign it passes to the web application otherwise an error is generated through the Admin Console.

Rule Cache is the temporary storage of rules at run time. It stores the rules form the rule generation module. These rules would be applied upon the parsed requests for the detection of malicious activities.
The Rule Generator dynamically generates the rules on the basis of ontological models stored in the knowledge base. The Rule generator is responsible for creating semantic-based rule objects. The rule specifies the action after the detection of any malicious activity. The rule generator queries the inferred model, through inference upon the Knowledgebase and then generates rules specific to the user input applied upon the specific portion of the HTTP message. The rule actions include, discard the message, generate alert, log the current message, ignore the action or allow the message.

3.4 Knowledgebase

This module after the inference process sends the inferred knowledge to Rule generator module which dynamically generates rules. This module contains the ontology which helps in decision making of valid and invalid HTTP \ HTTPS requests.

The basic purpose of this inference engine is to formulate new decisions. Inferences engine first inferences the Knowledgebase and extract the required information and pass it to the Rule Generator through the Jena API. Inference engine responsibilities are validation and verification of

ontological model and passing the extracted information to the rule generator that generates the rules for attack detection.
The Knowledge Base contains the data in the form of ontological models. Ontology is specialization of conceptualization that has all the concepts of different web application attacks. It also contains the information of the HTTP/HTTPS protocols in the form of ontological models.

4. Evaluation

Evaluation of the system has been carried out by using OWASP's WebScarab [28] and WebGoat [27] tools for analysis purpose. System is compared and tested with Mod Security (open source)] and Profense Web Application Firewall (trial version), state of the art technologies for XSS and SQL attacks. Result shows that proposed system having higher detection rate with minor false positive.

The system is also effective and efficient in multiple ways. Firstly it analysis specific portion of user request where attack is possible thus avoid the sequential search of entire user request. Only focusing on specific portion of input where attack or exploit is possible would reduce the space reduction and save the processing time.

Secondly Semantic rules can be mention attack metadata such as severity level of the attack, the consequences of the attack on the system throughput and efficiency, and the contents of the alert message. Alert generation is modeled in the action part and is processed by the analysis engine.

The system is evaluated through comparison with other sate of the art solutions on the basis of detection rate. The system has assumption as criteria mention in [29] while calculating the detection rate and false alarm rate:

False Positive = *FP*: the total number of records that are classified as anomalous
False Negative = *FN*: the total number of anomalous records that are classified as normal
Total #Normal = *TN*: the total number of normal records
Total #Attack = *TA*: the total number of attack records
Detection Rate = *[(TA-FN)/TA]*100*
False Alarm Rate = *[FP/TN]*100*

Table1. False Alarm and Detection Rate of the proposed System

Web Application Attacks	False Alarm Rate %	Detection Rate%
Cross Site Scripting(XSS)	2.4	94.93
SQL Injection(SQLi)	2.3	95.39
Cross site Request Forgery(CSRF)	1.2	88.54
Predictable Resource Location	1.4	92.94
CRLF Injection	2.2	80.34
Cross User Defacement	2.3	83.29
Denial of Service (Dos)	1.2	84.44
Hidden Field Manipulation	2.1	83.45

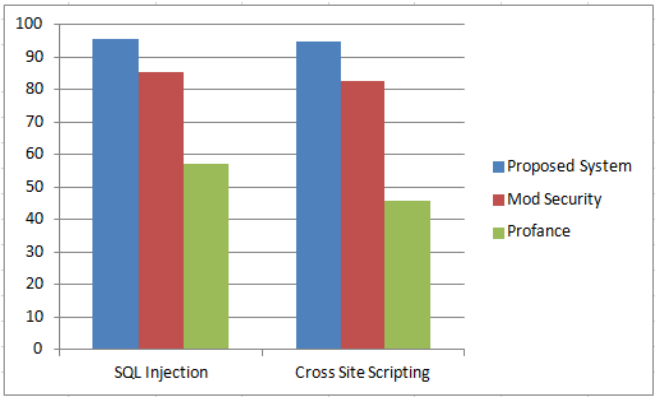

Fig 4. Comparison of detection rate of XSS and SQLi

5. Conclusion

In this paper the brief overview of various security techniques have been presented and concluded that due to increasing security concern for web applications, future survival of e-business organizations depends on the effective security measures at application level. We have critically studied the existing techniques and figured out that, a semantic based architecture capable of making intelligent decision based on the context of the target domain. The proposed system proposes an innovative approach of applying semantics in web application security. It opens up a completely new dimension to web application security by creating the synergy between information security and semantic systems. The basic theme is to capture the context of the web applications at a certain level of abstraction as ontological models pertinent to attack scenarios, communication protocols, vulnerabilities, consequences, and mitigation actions.

Acknowledgements

This research is funded by National ICT R&D Fund Pakistan under the ICT-Related Development and Research Grant on Project "Semantics based Web Application Security: Concept, Design and Implementation"

References

[1] World Internet Usage and Population Statistics March 31, 2011, http://www.internetworldstats.com/stats.htm

[2] Symantec Internet Security Threat Report (2010), https://www4.symantec.com/mktginfo/downloads/21182883_GA_REPORT_ISTR_Main-Report_04-11_HI-RES.pdf

[3] White hat report on web application security (2010), https://www.whitehatsec.com/assets/WPstats_winter11_11th.pdf?doc=WPstats_winter11_11th

[4] Resource Description Framework (RDF) Schema Specification 1.0, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[5] Jena – A Semantic Web Framework for Java. [Online] Available: http://jena.sourceforge.net/

[6] The Protégé Ontology Editor and Knowledge Acquisition System. http://protege.stanford.edu/

[7] Pellet: The Open Source OWL 2 Reasoner. [Online] Available: http://clarkparsia.com/pellet

[8] Andreas M. Antonopoulos SVP and Founding Partner, "Dynamic, Flexible Security Architecture", Nemertes Research 2011. http://www.cisco.com/en/US/prod/collateral/vpndevc/dynamic_flexible_security_architecture.pdf

[9] Gunnar Peterson "Security Architecture Blueprint" Arctec Group 2007 , http://arctecgroup.net/pdf/ArctecSecurityArchitectureBlueprint.pdf

[10] R. Boyer and J. Moore, "A fast string- searching algorithm" Communication of the ACM, 1997.

[11] Xin Zhao and Atul Prakash. WSF: "An HTTP-level Firewall for Hardening Web Servers".

[12] A. Anitha and V. Vaidehi. "Context based Application Level Intrusion Detection". IEEE: International conference on Networking and Services (ICNS'06) 2006.

[13] Frank S. Rietta. Application Layer Intrusion Detection for SQL Injection. ACM SE'06 March 10 12, 2006, Melbourne, Florida, USA. In Symposium on Applied Computing (SAC), ACM Scientific Press, March 2002.

[14] Xiao-Feng Wang, Jing-Li Zhou, Sheng-Sheng Yu, and Long-Zheng Cai. "Data Mining Methods for Anomaly Detection of HTTP Request Exploitations," Springer-Verlag Berlin Heidelberg 2005.

[15] V. Raskin, C.F. Hempelmann, K.E. Triezenberg, Nirenburg, "Ontology in Information Security: "A Useful Theoretical Foundation and Methodological Tool," Proceedings of the 2001 Workshop on New Security Paradigms (NSPW-2001), pp. 53-59, 2001.

[16] C. E. Landwehr, A. R. Bull, J. P. McDermott, and W. S. Choi. Taxonomy of Computer Program Security Flaws. ACM Computing Surveys, 26(3):211 – 254, September 1994.

[17] P. Ning, S. Jajodia, and X. S. Wang. Abstraction-Based Intrusion in Distributed Environments. ACM Transactions on Information and Systems Security, 4(4):407 – 452, November 2001

[18] J. McHugh. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPAIntrusion Detection System Evaluations as Performed by Lincoln Laboratory. ACM Transactions on Information and System Security, November 2000.

[19] B. Guha and B. Mukherjee. Network Security via Reverse Engineering of TCP Code: Vulnerability Analysis and Proposed Solutions. In IEEE Networks, pages 40 – 48. IEEE, July/August1997.

[20] J. Undercoffer, J., Pinkston, A. Joshi, T. Finin, "Target-Centric Ontology for Intrusion Detection," IJCAI Workshop on Ontologies and Distributed Systems (IJCAI'03), August, 2003.

[21] XML-Signature SYNtax and Processing, http://www.w3.org/TR/xmldsig-core/, 2002.

[22] G. Denker, L. Kagal, T. Finin, M. Paolucci, K. Sycara, "Security for DAML Web Services: Annotation and Matchmaking," The Semantic Web (ISWC 2003), LNCS 2870, Springer, , 2003.

[23] DAML+OIL.: www.daml.org/2000/12 /daml+oil.dam.

[24] Abdul Razzaq, Ali Hur, Muddassar Masood, Khalid Latif, H Farooq Ahmad,Hironao Takahashi "Foundation of Semantic Rule Engine to Protect Web Application Attacks" ISADS-2011.

[25] K. K. Mookhey, Network Intelligence "Detection and Evasion of Web Application Attacks," :www.nii.co.in.

[26] Tammo Krueger "TokDoc: A Self-Healing Web Application Firewall" SAC'10 March 22-26, 2010, Sierre, ACM Switzerland

[27] WebGoat. : http://www.owasp.org/index.php/Category:OWASP_WebGoat_Project

[28] WebScarab. http://www.owasp.org/index.php/Category:OWASP_WebScarab_Project

[29] S.T. Sarasamma, Q.A. Zhu, and J. Huff, „Hierarchel Kohonenen Net for Anomaly Detection in Network Security", IEEE Transaction on Systems, Man, and Cybernetics-Part B: Cybernetics, 35(2) ,2005,pp.302-312.

[30] Mod Security, http://www.modsecurity.org/

Abdul Razzaq: PhD scholar at National university of Science and Technology (School of Electrical Engineering and Computer Science) Pakistan. MSc Mathematics (1992), Master in Information Technology (2004), Master of Science - IT (2009). The pioneer of Semantics base Web Application Firewall (SWAF). Presently Team Lead in the ICT R&D Project "Semantics based Web Application Security: Concept, Design and Implementation". Various publications in the domain of semantic based web application security. Patent filed on Web Application Firewall. Current research interests include the formal modeling of Cyber Security, Semantic Systems and Vulnerability Analysis.

Ali Hur: Master of Science - IT (2009). The pioneer of Semantics base Web Application Firewall (SWAF). Patent filed on Web Application Firewall. Presently Team Lead/ System Architect in the ICT R&D Project "Semantics based Web Application Security: Concept, Design and Implementation". Current research interests include Web Security, Semantic Systems and Architecture.

H Farooq Ahmad: PhD - Distributed Computing, Software Agents – Tokyo Institute of Technology, Tokyo, Japan, Japan (2002). Associate Professor - School of Electrical Engineering and Computer Science (NUST), Pakistan. Patent filed on Web Application Firewall. Project Director in the ICT R&D Project "Semantics based Web Application Security: Concept, Design and Implementation". Expertise in Distributed Computing, Semantic Systems, Semantic Security, Semantic Health.

Muddassar Masood: Master of Science - IT candidate at School of Electrical Engineering and Computer Science, Pakistan. Senior researcher in the ICT R&D Project "Semantics based Web Application Security: Concept, Design and Implementation". Current research interests include Web Security and System Architecture.