HTTP DATASET CSIC 2010

HTTP DATASET CSIC 2010

- **What is this?**: The HTTP dataset CSIC 2010 contains thousands of web requests automatically generated. It can be used for the testing of web attack protection systems. It was developed at the “Information Security Institute” of CSIC (Spanish Research National Council).

- **Motivation**: A current problem in web attack detection is the lack of publicly available data sets to test WAFs (Web Application Firewalls). The DARPA data set [1, 2] has been widely used for intrusion detection. However, it has been criticized by the IDS community [3]. Regarding web traffic, some of the problems of the DARPA data set are that it is out of date and also that it does not include many of the actual attacks. Because of that, it is not appropriate for web attack detection. The problem of data privacy is also a concern in the generation of publicly available data sets and is probably one of the reasons why most of the available HTTP data sets do not target real web applications. Because of these reasons, we decided to generate our own HTTP data set CSIC 2010 and we present it here.

- **Dataset Description**: The HTTP dataset CSIC 2010 contains the generated traffic targeted to an e- Commerce web application developed at our department. In this web application, users can buy items using a shopping cart and register by providing some personal information. As it is a web application in Spanish, the data set contains some Latin characters.

  The dataset is generated automatically and contains 36,000 normal requests and more than 25,000 anomalous requests. The HTTP requests are labeled as normal or anomalous and the dataset includes attacks such as SQL injection, buffer overflow, information gathering, files disclosure, CRLF injection, XSS, server side include, parameter tampering and so on. This dataset has been successfully used for web detection in previous works [4, 5, 6, 7, 8, 9].

  The traffic is generated following the next steps:

First, real data are collected for all the parameters of the web application. All the data (names, surnames, addresses, etc.) are extracted from real databases. These values are stored in two databases: one for the normal values and other for the anomalous ones. Additionally, all the public available pages of the web application are listed.
Next, normal and anomalous requests are generated for every web page. In the case that normal requests have parameters, the parameter values are filled out with data taken from the normal database randomly. The process is analogous for anomalous requests, where the values of the parameters are taken from the anomalous database.

Three types of anomalous requests were considered:

1) Static attacks try to request hidden (or non-existent) resources. These requests include obsolete files, session ID in URL rewrite, configuration files, default files, etc.

2) Dynamic attacks modify valid request arguments: SQL injection, CRLF injection, cross-site scripting, buffer overflows, etc.

3) Unintentional illegal requests. These requests do not have malicious intention, however they do not follow the normal behavior of the web application and do not have the same structure as normal parameter values (for example, a telephone number composed of letters).

The attacks were generated with the help of tools such as Paros [10] and W3AF[11].

The WAFs where this dataset was used [4,5,6,7] follow the anomaly approach, i.e. the normal behavior of the web application is defined and the behavior apart from that are considered anomalous. Therefore, in this approach only normal traffic is needed for the training phase.

The dataset is divided into three different subsets. One subset for the training phase, which has only normal traffic. And two subsets for the test phase, one with normal traffic and the other one with malicious traffic.

- **Download**:

  Normal Traffic (Training)

Normal Traffic (Test)

Anomalous Traffic (Test)

- **Authors**: Carmen Torrano Giménez, Alejandro Pérez Villegas, Gonzalo Álvarez Marañón.

- **References**:

[1] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. Kendall, D. McClung, D. Webber, S. Webster, D. Wyschograd, R. Cunninghan, and M. Zissman. Evaluating Intrusion Detection Systems: The 1998 DARPA off-line intrusion detection evaluation. In Proc. of DARPA Information Survivability Conference and Exposition (DISCEX00), Hilton Head, South Carolina, January 25-27. IEEE Computer Society Press, Los Alamitos, CA, 1226 (2000).

[2] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba and K. Das. The 1999 DARPA Off-Line Intrusion Detection Evaluation. In Proc. Recent Advances in Intrusion Detection (RAID2000). H. Debar, L. Me, and S. F. Wu, Eds. Springer-Verlag, New York, NY, 162182 (2000).

[3] J. McHugh. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. In Proc. of ACM Transactions on Information and System Security (TISSEC) 3(4), pp. 262-294 (2000).

[4] A. Perez-Villegas, C. Torrano-Gimenez, G. Alvarez. Applying Markov Chains to Web IntrusionDetection. In Proc. of Reunión Española sobre Criptología y Seguridad de la Información (RECSI 2010), pp. 361-366. Publicaciones urv. Tarragona (España), 7-10 Septiembre (2010).

[5] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez. An anomaly-based approach for intrusion detection in web traffic. Journal of Information Assurance and Security, vol. 5, issue 4, pp. 446-454. ISSN 1554-1010 (2010).

[6] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez, A Self-

Learning Anomaly-Based Web Application Firewall. In Proc. of 2nd International Workshop in Computational Intelligence in Security for Information Systems (CISIS 09). Advances in Intelligent and Soft Computing, vol. 63, pp. 85-92, Springer-Verlag. A. Herrero, P. Gastaldo, R. Zunino, E. Corchado, editores. Burgos (España), 23-26 Septiembre (2009).

[7] C. Torrano-Gimenez, A. Perez-Villegas, G. Alvarez, An Anomaly-based Web Application Firewall. In Proc. of International Conference on Security and Cryptography (SECRYPT 2009), pp. 23-28. INSTICC Press. E. Fernández-Medina, M. Malek, J. Hernando, editores. Milán (Italia), 7-10 Julio (2009).

[8] H. Nguyen, C. Torrano-Gimenez, G. Álvarez, S. Petrovic, K. Franke, Application of the Generic Feature Selection Measure in Detection of Web Attacks. In Proc. of International Workshop in Computational Intelligence in Security for Information Systems (CISIS 11), LNCS 6694, pp. 25–32. Editor Á. Herrero and E. Corchado, Springer-Verlag. Torremolinos, Málaga (España), Junio (2011).

[9] C. Torrano-Gimenez, H. Nguyen, G. Álvarez, S. Petrovic, K. Franke, Applying Feature Selection to Payload-Based Web Application Firewalls. In Proc. of International Workshop on Security and Communication Networks (IWSCN 11), pp. 75-81. Editor Patric Bours. Gjøvic (Noruega). ISBN: 978-82-91313-67-2. 18-20 Mayo (2011).

[10]Chinotec Technologies Company: Paros - for web application security assessment. http://www.parosproxy.org/index.shtml (2004).

[11] Andrés Riancho: Web Application Attack and Audit Framework. http://w3af.sourceforge.net (2007).

23204

SS