Using Diverse Detectors for Detecting Malicious Web Scraping Activity

Pedro Marques[1,3], Zayani Dabbabi[2], Miruna-Mihaela Mironescu[2], Olivier Thonnard[2], Alysson Bessani[1], Frances Buontempo[3], Ilir Gashi[3]
[1]LaSIGE, Faculdade de Ciencias, Universidade de Lisboa, Portugal
[2]Amadeus, France
[3]Centre for Software Reliability, City, University of London, United Kingdom
pedro.dm.marques@gmail.com; anbessani@ciencias.ulisboa.pt; {Zayani.Dabbabi, Miruna-Mihaela.Mironescu, Olivier.Thonnard}@amadeus.com; {frances.buontempo, ilir.gashi.1}@city.ac.uk

***Abstract*— We present ongoing work about how the use of diverse tools may help with detecting malicious web scraping behavior. We use a real dataset of Apache HTTP Access logs for an e-commerce application provided by Amadeus, a large multinational IT provider for the global travel and tourism industry. Two tools have been used to detect scraping activities based on the HTTP requests: a commercial tool from Distil Networks, and an in-house tool called Arcane. Preliminary results suggest there is considerable diversity in alerting behavior of these tools.**

Keywords - security assessment; software diversity; security tools; botnet detection.

I. BACKGROUND

Web scraping is the process of using bots to extract content and data from a website[1]. There are many legitimate use cases of web scraping, such as a search engine bots crawling a site, analyzing its content and then ranking it; price comparison sites deploying bots to auto-fetch prices and product descriptions for seller websites etc. However, web scraping is also used for illegal purposes. Use cases of illegal use include undercutting of prices, theft of copyrighted content etc. In price scraping, a perpetrator typically uses a botnet from which to launch scraper bots to inspect competing business databases. The goal is to access pricing information, undercut rivals and boost sales. Attacks frequently occur in industries where products are easily comparable and price plays a major role in purchasing decisions. Victims of price scraping can include travel agencies, ticket sellers and online electronics vendors. Large multi-nationals such as Amadeus, who are an IT provider for the global travel and tourism industry, are prime targets for this type of malicious activity. To protect themselves from these types of attacks, organizations use specialized software that can monitor for suspicious activity, attempt to separate bot traffic from human traffic, use IP reputation websites to block activities from suspicious IP addresses, monitor the behavior of visitors in the way in which they interact with the website to check for abnormal browsing patterns etc. Amadeus use a commercial tool from Distil Networks[2] and an in-house tool they have developed called Arcane. Both of these tools monitor the same application layer interactions to monitor for malicious web scraping behavior. An interesting question is how diverse these tools are in their detection behavior. In this paper we present some preliminary results based on the behavior of these tools when analyzing a subset of Amadeus traffic over a one week period in March 2018. The data is not labelled yet, which means we cannot present the data in terms of the usual measures for binary classifiers (e.g. Sensitivity and Specificity[3]), though this is the intended next step in our research.

The rest of the paper is organized as follows: Section II presents related work. Section III briefly explains the dataset. Section IV presents preliminary results of the analysis of this dataset and Section V outlines the next steps of our research.

II. RELATED WORK

There have been several works that have looked at ways in which malicious web crawling and scraping can be detected (e.g. [1], [2], [3]) but none that we are aware of that has looked at combining multiple diverse detectors.

The security community is well aware of diversity as potentially valuable [4], [5]. Discussion papers argue the general desirability of diversity among network elements, like communication media, network protocols, operating systems etc., but only sparse research exists on how to choose diverse defenses (some examples in [6], [7] [5, 8]).

III. DATASET

The dataset consists of Apache HTTP Access logs for an e-commerce application. The dataset covers a period of 8 days: from March 11th to March 18th 2018. Two tools (namely Distil Network and Arcane) have been used to detect scraping activities based on the HTTP requests.

IV. RESULTS

Table 1 below shows the total number of HTTP requests (1,469,744) in the analyzed dataset, and the totals HTTP requests alerted from the two tools.

Table 1 – HTTP requests alerted by the two tools

Total HTTP requests	**1,469,744**
HTTP request alerted as malicious by **Distil**	1,275,056
HTTP request alerted as malicious by **Arcane**	1,240,713

[1] https://www.incapsula.com/web-application-security/web-scraping-attack.html

[2] https://www.distilnetworks.com/

[3] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

We analyzed the similarity and diversity in the alerting behavior by the two tools. Table 2 shows the breakdown of the HTTP request that were alerted by both tools, by neither, or by only one of the tools respectively. We notice the *similarity* in the alerting behavior by the two tools (both tools alert on than 1.2 million of these HTTP requests), but there is also *diversity* in the alerting behavior: 43,648 HTTP requests are alerted by Distil only, and 9,305 are alerted by Arcane only.

Table 2 – Diversity in the alerting behavior by the two tools

HTTP request alerted as malicious by:	Count
Both Distil and Arcane	1,231,408
Neither	185,383
Arcane Only	9,305
Distil Only	43,648

We also analyzed the breakdown of these alerts, for both of these tools, based on the HTTP status[4] of each request. Table 3 contains the results for both tools in total, while Table 4 contains the breakdown of the alerts by HTTP request only for those request alerted by one of the tools.

Table 3 – Alerted requests by HTTP status – overall counts

Arcane		Distil	
HTTP status	**Count**	**HTTP status**	**Count**
200 (OK)	1,204,241	200 (OK)	1,239,079
302 (Found)	34,561	302 (Found)	34,832
204 (No content)	1,560	204 (No content)	1,018
400 (Bad request)	256	400 (Bad request)	73
304 (Not modified)	76	404 (Not found)	32
500 (Internal Server Error)	11	304 (Not modified)	15
404 (Not found)	8	500 (Internal Server Error)	6
		403 (Forbidden)	1

V. Conclusions So Far and Next Steps

The preliminary analysis conducted so far reveals that there is diversity in the alerting behavior of these tools. Whether this diversity is beneficial or not depends on whether the alerts that the tools are generating are True Positives or False Positives. Also, and perhaps more importantly, we also need to label whether the HTTP requests on which the tools for no alert are True Negatives or False Negatives. The Amadeus team is currently working on labelling the dataset, as well as providing new datasets, to enable this type of analysis. Based on the analysis using labelled data we can derive conclusions on whether diversity is useful in this context under different adjudication schemes (e.g. 1-out-of-2, raise an alarm as long as either tool does so; 2-out-of-2, only raise an alarm if both tools do so etc.). We can also analyze the trade-offs between false positives and false negatives when deploying the tools in parallel (both tools monitor all the traffic) versus serial configurations (one tool monitors and filters the traffic that need to be also analyzed by the second tool). We also plan to look more closely in to the alerts generated by only one of the tools, to get a better understanding on the possible reasons why a given tool is more appropriate to detect certain behaviors, and thus how diversity could enhance the detection rate. These results can prove valuable to Amadeus in their endeavor to protect their networks from malicious web-scraping activity.

Table 4 - Alerted requests by HTTP status for those request that were alerted by only one tool

Arcane only		Distil only	
HTTP Status	**Count**	**HTTP status**	**Count**
200 (OK)	7,693	200 (OK)	42,531
204 (No content)	956	302 (Found)	592
302 (Found)	321	204 (No content)	414
400 (Bad request)	247	400 (Bad request)	64
304 (Not modified)	76	404 (Not found)	31
404 (Not found)	7	304 (Not modified)	15
500 (Internal Server Error)	5	403 (Forbidden)	1

Acknowledgment

This work was supported in part by the EU H2020 framework DiSIEM project, and the UK EPSRC project D3S.

References

1 Stevanovic, D., An, A., and Vlajic, N.: 'Feature evaluation for web crawler detection with data mining techniques', Expert Syst. Appl., 2012, 39, (10), pp. 8707-8717

2 Stassopoulou, A., and Dikaiakos, M.D.: 'Web robot detection: A probabilistic reasoning approach', Computer Networks, 2009, 53, (3), pp. 265-278

3 Al-Bataineh, A., and White, G.: 'Analysis and detection of malicious data exfiltration in web traffic', 2012, pp. 26-31

4 Littlewood, B., and Strigini, L.: 'Redundancy and diversity in security', Springer-Verlag, 2004, pp. 423-438

5 Garcia, M., Bessani, A., Gashi, I. et al. 'Analysis of operating system diversity for intrusion tolerance', Software: Practice and Experience, 2014, 44, (6), pp. 735-770

6 Sanders, W.H., Cukier, M., Webber, F. et al. 'Probabilistic Validation of Intrusion Tolerance', in DSN 2003, pp. 615-624

7 Gupta, V., Lam, V., Ramasamy, H.V. et al: 'Dependability and Performance Evaluation of Intrusion-Tolerant Server Architectures', in LADC 2003, Springer, pp. 81-101

8 Bishop, P., R.B., Gashi, I., and Stankovic, V.: 'Diversity for Security: A Study with Off-the-Shelf Antivirus Engines', in IEEE ISSRE 2011, pp. 11-19

[4] https://developer.mozilla.org/en-US/docs/Web/HTTP/Status