An industrial perspective on web scraping characteristics and open issues

Elisa Chiapponi*, Marc Dacier†, Olivier Thonnard‡, Mohamed Fangar‡, Mattias Mattsson‡, Vincent Rigal‡
* EURECOM, France, † RC3, CEMSE - KAUST, Kingdom of Saudi Arabia, ‡ Amadeus IT Group, France
elisa.chiapponi@eurecom.fr, marc.dacier@kaust.edu.sa,
{olivier.thonnard,mohamed.fangar,mattias.mattsson,vincent.rigal}@amadeus.com

***Abstract*—An ongoing battle has been running for more than a decade between e-commerce websites owners and web scrapers. Whenever one party finds a new technique to prevail, the other one comes up with a solution to defeat it. Based on our industrial experience, we know this problem is far from being solved. New solutions are needed to address automated threats. In this work, we will describe the actors taking part in the battle, the weapons at their disposal, and their allies on either side. We will present a real-world setup to explain how e-commerce websites operators try to defend themselves and the open problems they seek solutions for.**

***Index Terms*—Web Scraping, Residential IP Proxy**

I. Introduction

In today's world, everyone uses the Internet to buy all kinds of goods, from shoes to books, from jewelry to event tickets. One could think that only two parties take an active part in this process: the user interested in the goods and the e-commerce platform providing the products. However, this is not the case.

In most scenarios in which a lucrative e-commerce activity is involved, scraping bots are part of the scene. Web scraping consists of the periodical or continuous retrieval of accessible data and/or processed output contained in web pages [1]. This activity is usually conducted by a scraper behind a network of bots. The goal of this actor is to obtain the e-commerce prices for different purposes which all lead to a gain for him.

On the e-commerce side, web scraping has large business implications. To protect themselves from these malicious actors, e-commerce platforms take advantage of anti-bot solutions. Over the years, anti-bot solutions and bots have engaged into an endless arms race, leading to a continuous production of detection techniques followed by evasion ones.

Recently, scrapers have brought to the scene another player to facilitate their activities: Residential IP Provider (RESIP) companies. These players claim to provide a vast network of residential IP addresses to their clients and some of them provide tools to developers to build stealthy bots.

As of today, the battle between e-commerce and bots is still ongoing, but the bots are on the winning side. In this work, we want to share an industrial viewpoint on this ongoing battle and its implications, hoping to encourage researchers to seek new solutions.

In section II, we will describe at a high level the players in this scenario. Then, in section III, we will show a real-world case study: the fight between one of the major players in e-commerce for travel and scraping bots.

II. The players

A. E-commerce websites

E-commerce websites are platforms in which goods owners show their products and customers can perform online purchases. In 2021, online sales amounted to 4.9 trillion U.S. dollars worldwide and trends show that this number will increase in the next years [2].

These services come with costs, such as maintaining the servers in which the information resides, retrieving the data, performing calculations, and rendering the final page to the client. Usually, the business model of e-commerce websites takes into account that only a few items among the displayed ones will end up being bought. Hence, the revenue for the sale of an item covers the costs associated with the maintenance of the whole infrastructure. Naturally, price scraping affects this business model. According to DataDome, an anti-bot company [3], the loss could be up to 10% of the revenue of a website. The income reduction can be direct and/or indirect. The direct one is caused by the dramatic rise in the number of requests which is not followed by an increase in the number of purchases. Moreover, bots interfere with the metrics used to evaluate e-commerce business. The indirect loss of revenue is caused by scraping bots representing large portions of the traffic towards a website and causing congestion and slow connections. This situation can reduce the number of legitimate users reaching the website and/or downgrade their user experience. In both cases, the company could lose a potential client.

Initially, e-commerce websites answered to these attacks with layer 3 network mitigations, such as IP blocking and IP-based rate limiting. With the bots becoming more and more sophisticated, in-house solutions were not enough anymore, thus e-commerce websites bought anti-bot solutions.

B. Scrapers

Web scraping is performed for different purposes such as content reselling, statistics modifications, and competitors' price monitoring. The information gathered through this practice can be used directly by the scrapers or can be sold for an income.

As documented by the anti-bot company Imperva [4], scrapers target a wide range of markets and they change their activities according to the global situation. With the COVID-19 pandemic reducing incomes from the traditional sectors,

e.g. tickets for events, bots modified their actions targeting websites selling medical devices, such as masks. Furthermore, as stated by the anti-bot company Akamai [5], bots profit from peaks of traffic due to holidays, e.g. the Lunar New Year, to increase their traffic without being noticed.

Bots evolved from simple scripts to browser emulation frameworks e.g. Scrapy, Phantom JS. They started using automated browsers, e.g. Selenium, and implementing cookies and Javascript support. In this way, they made it more difficult to distinguish them from real browsers. Furthermore, they started reacting to CAPTCHAS [6], creating infrastructures able to forward them to real people paid to resolve them (CAPTCHA farms) [7]. These workers are required to solve the tests in a time comparable to the one used by direct users, to prevent the anti-bot solutions from recognizing them. Another way consists in forwarding the CAPTCHAS to unaware users of other websites and letting them solve them [8].

Finally, scrapers have started to use the services of so-called Residential IP (RESIP) providers. These parties offer millions of residential IP addresses that scrapers can use as exit point of their requests. The advantages for the scrapers are multiple: no need to have a private distributed infrastructure, no possibility to directly trace back the activity to the initiator, and access to a pool of IPs with a good reputation.

C. Anti-bot companies

Anti-bot companies emerged in the late 2010s. Their products are positioned in front of websites to protect them. Their goal is to detect and mitigate bot traffic. Each company uses different technologies.[1]

Initially, bot detection was performed based on IP reputation, HTTP-based rate-limiting, and HTTP header anomaly detection. Then, browser fingerprinting became the favorite choice [10]. With bots avoiding these detection methods, JavaScript (JS) and cookies challenges were introduced. These tests run JS in the requester browser and collect information from it. Simple bots could not solve these challenges because they did not have JS and cookie storage support. Scrapers implemented their activities with automated browsers able to run JS and store cookies to avoid detection. Over the years, fingerprinting and challenges have evolved, including automated browser detection and checking for human interactions, e.g. mouse movement. CAPTCHAS have been adopted to stop bots until the advent of CAPTCHA farms. Machine learning algorithms have been implemented to better distinguish between user and bot interactions. Lately, a common approach is to try to waste the scrapers' time, hoping to produce losses in their revenues. New solutions are crypto challenges that make bots waste CPU cycles in solving them. Other techniques such as *tarpit* consist of slowing down the bot connections or giving them open connections with no responses.

Moreover, at present, we can categorize anti-bot solutions in two main detection approaches, the *knowledge-* and *behaviour-*based ones [11]. The first approach consists in recognizing the fingerprints of scrapers by studying the HTTP headers and by grouping requests according to specific parameters. Then, specific rules are written to block further requests matching these traffic subsets. In the second approach, machine learning is privileged and the detection is performed by detecting outliers. Requests which HTTP headers and/or payloads differing significantly from the ones issued by known human beings are considered to be coming from bots and are answered with a countermeasure.

[1]We provide a high-level overview of anti-bot solutions. We refer the interested reader to Azad et al. [9] for a thorough introduction.

D. RESIP companies

In the past, scraping was conducted mainly by leveraging data center machines and compromised machines. Recently, RESIP companies started changing this trend. As displayed in a blog post of the company DataDome [12], RESIP IPs counted for almost 30% of the malicious bot traffic at the end of the year 2019.

These companies announce on their websites to have access to tens of millions of residential IP addresses and allow scrapers to use their network upon payment. The scraper just sends his request to a so-called super proxy which then forwards the request to a residential machine in its network. The request is then sent to the target website with the IP of that machine.

Moreover, some RESIP companies offer automated services able to rotate among different fingerprints, perform CAPTCHA solving and JS rendering automatically. This helps overcome the anti-bot detection systems and allows scrapers who are not proficient developers to easily conduct their campaigns.

As explained in the works of Mi et al. [13], [14], RESIP infrastructures are built taking advantage of mobile SDKs included by developers in all kinds of applications in exchange for a fee per installed app. A percentage of these networks are composed of infected devices, such as IoT ones. A recent DataDome blog post [15] points out that browser extensions are used as a vehicle of proxy activity and that a recent trend consists in building mobile proxy networks using dedicated hardware. In this way, scrapers create large clusters of SIM cards and use them to route the traffic towards mobile networks.

RESIP connections are a big problem because they often offer IPs used by legitimate users (eg., their mobile phone IPs). Thus, the risk of blocking a connection from a genuine customer is high and this prevents e-commerce companies to put in place strict policies against those IP addresses.

III. Case study: Amadeus IT Group

A. Overview

Amadeus IT Group later referred to as *Amadeus*, is a Global Distribution System (GDS) and is one of the world's leading technology companies for the travel industry. The products made by this company are used around the globe by more than 480 airlines, 128 airports, 300 hotel chains, and the large majority of travel players. In 2019, more than 646 million bookings were processed by Amadeus and 1.9 billion

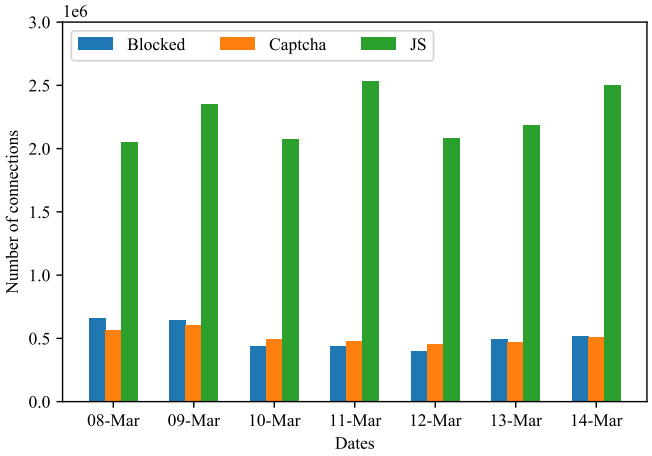

Fig. 1. Daily amount of countermeasures between 08/03/22 and 14/03/22.

passengers were boarded thanks to its portfolio of IT solutions. [16].

Among the various products offered by the company, there are solutions specifically built for airlines to let passengers make bookings on airline websites. These products share a common back-end, in which Amadeus calculates for each customer request the possible flight routes and their corresponding price. These fares are computed in real-time and based on a large number of parameters e.g. origin-destination, departure and arrival dates, travel classes, passenger types, etc, but also the availability of seats, the period of the year, and many other business rules. Thus, every generated response comes with a high computation cost.

Clearly, price scraping increases dramatically the costs for providers like Amadeus and its customers. Moreover, one of the Key Performance Indicators (KPIs) for an airline is the look-to-book ratio. This value is the ratio between the number of searches made by website visitors and the actual number of bookings made on the airline website. The high volume of traffic generated by bots towards airlines' websites is highly corrupting this important indicator. As highlighted by Imperva [4], the travel industry is one of the most targeted sectors and the one in which the percentage of sophisticated bots is predominant: 59,7% of the overall bot traffic.

Scrapers abuse airline domains to automatically collect and take advantage of the displayed fares. Based on the purpose of their action, we can divide these bots into three categories:

1) *competitive intelligence companies* that scrape directly or through a third-party organization to collect fare intelligence about airlines and provide this data as a service to competitors;
2) *aggregators*, metasearch players which find the best solutions for users across different airline domains;
3) *online travel agencies* which aggregate content and create new offers.

Since 2015, Amadeus is using a ruled-based anti-bot solution from a third-party company to protect more than 80 airline companies, corresponding to more than 200 websites. Every week more than 100 new custom rules are put in place to mitigate the scraping traffic and, every month, around 140 million requests trigger these rules. In the month of February 2022, on average, 8 intense bot investigations on specific airlines were running every week to mitigate the huge amount of bots requests. In the same month, considering all the domains protected by the anti-bot solution, that product adjudicated that 41% of the attempted connections to our servers were issued by bots. Fig. 1 shows the countermeasures served to bots in a representative, yet relatively quiet, week of March 2022. More than 22 million rules were triggered and the plot shows how the bot traffic is constant and not concentrated on specific days.

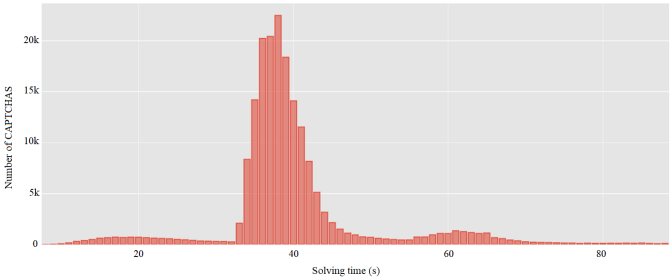

Fig. 2. Solving time of CAPTCHAS (2018).

Today, bots respond immediately to new custom rules. Once enough requests are detected as coming from a scraper, they are blocked or receive a countermeasure by our team. In the past years, it was taking around 24h for the scraper to see the problem and change parameters to avoid detection. Nowadays, we can see bots changing behavior in a couple of hours after a rule has been put in place. We assume that scrapers have at their disposal analysts that monitor the bots 24/7 and act manually to bypass the custom rules as soon as the countermeasure is put in place.

B. CAPTCHAS *solving time*

When CAPTCHAS were first introduced, these tests were able to block bots. However, bots targeting Amadeus started using automated browsers able to break simple ones. Then, scrapers began redirecting the tests to CAPTCHA farms. Initially, the plot of the Gaussian curve of the bots' CAPTCHAS solving time was clearly different from the one of the real users. Thus, it was easy to find a threshold above which the requests were classified as coming from bots. Fig. 2 shows the solving time of CAPTCHAS of a specific airline in 2018. Normal users used to take 20 seconds, on average, to answer the test. We can see high peaks of activity with solving times of around 40s and 60s. We recognized these as bot activities and put a threshold lower than 40s to block them.

Today, the CAPTCHA farm solving times are comparable to the human ones and it is not possible anymore to distinguish

between the two categories. This shows how scrapers are adaptable and flexible to quickly react to countermeasures.

C. RESIP *activities*

Recently, scrapers started to take advantage of RESIP services. We identified this problem by noticing that a growing fraction of traffic flagged by the anti-bot solution was coming from residential Internet Service Providers (ISPs). In the 30 days between February 13 and March 15 2022, Amadeus blocked more than 22M connections, considering all the protected airlines. Dividing these connections by ISP of origin, we saw that a few providers accounted for the majority of the blocked connections. There were 43 ISPs of which more than 50K connections were blocked. This corresponded to 84% of the blocked traffic. Among these providers, 13 organizations were identified as mostly providing IP addresses belonging to residential use. Their traffic amount to 12% of the blocked traffic of the period. This percentage could look small, but we have to keep in mind that Amadeus' goal is to reduce to zero the probability of false positives. A decision to block a RESIP connection is only taken when the confidence is very high that it is a malicious, or a bot initiated, one. This data implies that the traffic received by bots through residential ISPs is a larger portion than this and thus shows how wide is their usage.

Moreover, we have personal experience of RESIP forwarding requests to us. During an investigation, one of the authors found out that we received many requests from IP addresses registered by one specific RESIP company. Most likely, there had been a problem in their setup and they were using their machines instead of residential ones to proxy the traffic.

The use of RESIP services is a big problem for e-commerce because it increases the risk of false-positive when serving countermeasures. RESIP companies claim to have access to millions of residential IP addresses which are shared between legitimate users and bots. In this scenario the risk of blocking a real user becomes high. However, recent works question these numbers and bring hope to the detection of these players. In 2019, Mi et al. [13] showed that the number of collected RESIP IPs was not in line with the values advertised by the companies' websites. Additionally, they displayed how two proxy providers shared part of their pools.

In our past work [17], for 56 days, we gathered bot requests, supposedly coming from RESIP services, targeting an airline domain. The requests were issued from almost 14K different IPs and 30% of them made requests in different days. This is not coherent with picking IPs from the very large pool advertised by RESIP services. We applied two mathematical models to the data to estimate the size of the pool from which these IPs were obtained. The first approach models the IP assignation with Uniform, Gaussian and Beta distributions. For each distribution, we calculated the size of the pool from which the collected IPs were taken. In the second approach, we found the curve that best fitted the cumulative distribution of new daily IPs and we projected it in time to find a plateau. Both approaches showed that the pool was likely much smaller than the numbers advertised by RESIP companies. Further analysis is needed but if these results were confirmed, it would be possible to detect scraping bots through the detection of the IPs of RESIP services.

IV. Conclusions

In this paper, we have presented how web scraping affects e-commerce websites and the current techniques used by scrapers and anti-bot solutions to overcome one another. We have shared the difficulties faced by e-commerce websites as a result of the adaptability and flexibility of sophisticated scraping bots. We have decided to give light to this problem so the research community can tackle it. Among other possible research avenues, we believe that the growing number of RESIP IP addresses used by bots requires the development of specific detection methods to defeat them.

References

[1] C. Watson and T. Zaw, "OWASP Automated Threat Handbook Web Applications," OWASP Foundation, Tech. Rep., 2018.

[2] https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/, accessed: 2022-03-10.

[3] https://datadome.co/cost-of-bot-calculator/ , accessed: 2022-03-14.

[4] Imperva, "Bad Bot Report 2021," Imperva, Tech. Rep., 2021.

[5] https://www.akamai.com/blog/security/china-and-japan-holiday-botnets, accessed: 2022-03-14.

[6] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Advances in Cryptology — EUROCRYPT 2003*, E. Biham, Ed. Springer, 2003, pp. 294–311.

[7] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G. M. Voelker, and S. Savage, "Re: Captchas: Understanding captcha-solving services in an economic context," in *Proc. USENIX Security 2010*, USA, p. 28.

[8] M. Egele, L. Bilge, E. Kirda, and C. Kruegel, "Captcha smuggling: Hijacking web browsing sessions to create captcha farms," in *Proc. of the 2010 ACM SAC conf.*, 2010, p. 1865–1870.

[9] B. Amin Azad, O. Starov, P. Laperdrix, and N. Nikiforakis, "Web runner 2049: Evaluating third-party anti-bot services," in *Proc. of DIMVA 2020*.

[10] A. Vastel, W. Rudametkin, R. Rouvoy, and X. Blanc, "FP-Crawlers: Studying the Resilience of Browser Fingerprinting to Block Crawlers," in *Proc. of MADWeb'20*, San Diego, United States.

[11] H. Debar, M. Dacier, and A. Wespi, "A revised taxonomy for intrusion-detection systems," in *Annales des télécommunications*, vol. 55, no. 7. Springer, 2000, pp. 361–378.

[12] https://datadome.co/bot-management-protection/one-third-bad-bots-using-residential-ip-addresses/, accessed: 2022-03-14.

[13] X. Mi, X. Feng, X. Liao, B. Liu, X. Wang, F. Qian, Z. Li, S. Alrwais, L. Sun, and Y. Liu, "Resident evil: Understanding residential ip proxy as a dark service," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 1185–1201.

[14] X. Mi, S. Tang, Z. Li, X. Liao, F. Qian, and X. Wang, "Your Phone is My Proxy: Detecting and Understanding Mobile Proxy Networks," in *Proc. of NDSS 2021*.

[15] https://datadome.co/bot-detection/how-proxy-providers-get-residential-proxies/, accessed: 2022-03-13.

[16] "Amadeus Global Report 2020," Amadeus IT Group, Tech. Rep., 2021.

[17] E. Chiapponi, M. Dacier, O. Catakoglu, O. Thonnard, and O. Todisco, "Scraping airlines bots: Insights obtained studying honeypot data," *Intl. Journal of Cyber Forensics and Advanced Threat Investigations*, vol. 2, no. 1, pp. 3–28, 2021.