Bot Detection and Traffic Analysis

Toma Taylor Makoundou | August 2022

Outline

- Identify_google_bot Signals
- Identify_bad_bot_traffic Signals
- Identify_human_traffic Signals
- Interesting Findings

identify_google_bot Signals

Toma Taylor Makoundou | August 2022

identify_google_bot Signals

This functions detects legitimate google bots traffic.

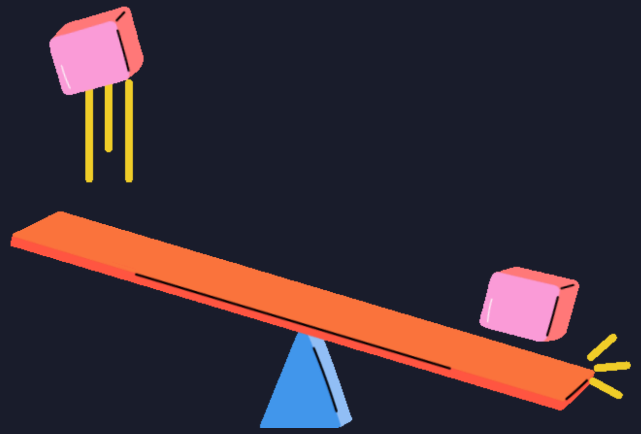
Heuristic Algorithm

identify_google_bot = google_bots_signals

identify_google_bot Signals

- **Signal 1**: Googlebot in User-Agent
- **Signal 2**: "GOOGLE" in apiIpAutonomousSystemOrganization
- **Signal 3**: Use of Legitimate IP addresses
- **Signal 4**: fingerprintRequestJsWebGlRend fingerprintRequestJsWebDriver fingerprintRequestJsHardwareConcu are null or NaN for legitimate google bots.
- **Signal 5**: apiEndpoint must be http.
- **Signal 6**: fingerprintAccept that does contain "text/html" at the very least

identify_google_bot Signals Considerations



Timestamps

Using timestamps as part of traffic analysis yields to more accurate behavioral analysis.

HTTP version

Google bots always access a site using http1.1. It does not access 2.0 unless supported and does not access a website using HTTP 0.x.

Reverse DNS Lookups

reverse DNS lookups help to deterministically verify the source IP of the requests.

Complete IP addresses

The complete IP addresses need to be provided in order to deterministically confirm the source of the request.

identify_bad_bot_traffic Signals

Toma Taylor Makoundou | August 2022

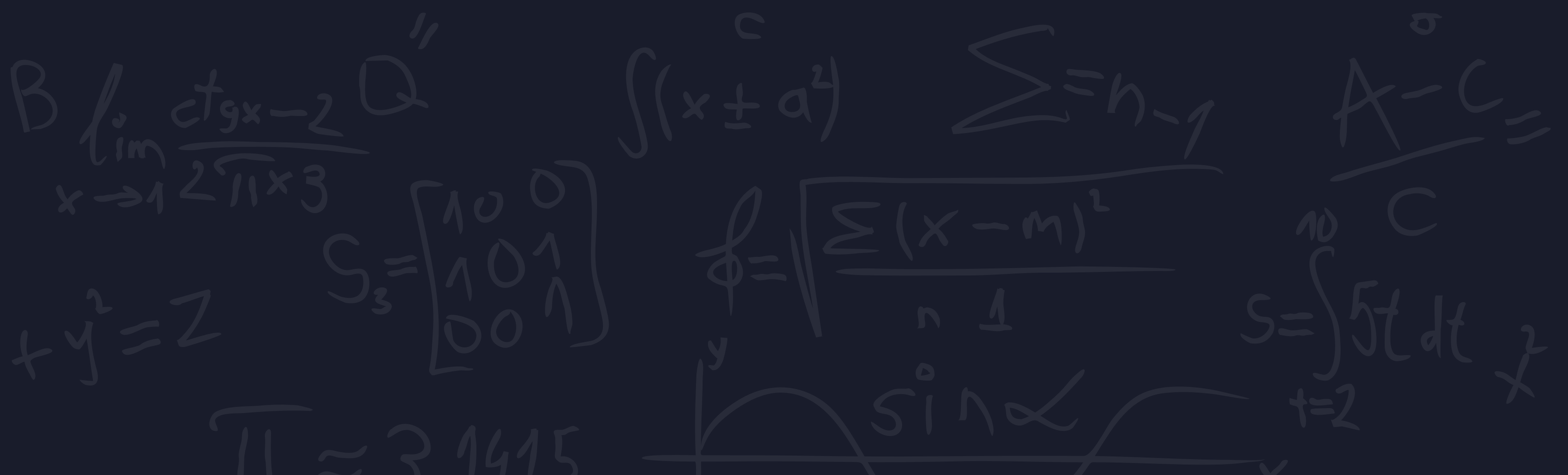
identify_bad_bot_traffic Signals

This function detects activities from non-identified bots, fake google bots, known bad bots and the user of libraries and net-tools.

Heuristic Algorithm

identify_bad_bot_traffic=(¬identify_google_bot)^(bad_bots_signals)

,where bad_bots_signals = (fake_google_bots_signals)^(non_identified_bots_signals)^(libraries_and_net_tools)^(path_traversal_attacks)



identify_bad_bot_traffic Signals

The signals used to eliminate Google Bots traffic:

- **Signal 1**: Googlebot NOT in User-Agent
- **Signal 2**: "GOOGLE" NOT in apiIpAutonomousSystemOrganization
- **Signal 3**: Do NOT Use Google IP addresses
- **Signal 4**: fingerprintRequestJsWebGlRend fingerprintRequestJsWebDriver fingerprintRequestJsHardwareConcu are NOT null or NaN for legitimate google bots.

identify_bad_bot_traffic Signals

Non-identified Bots Signals

- Signal 1: no user-agent

Fake bots Signals

- Signal 1: the user-agent value contains Googlebot
- Signal 2: the IP address does not belong to Google
- (Signal 1)^(Signal 2)

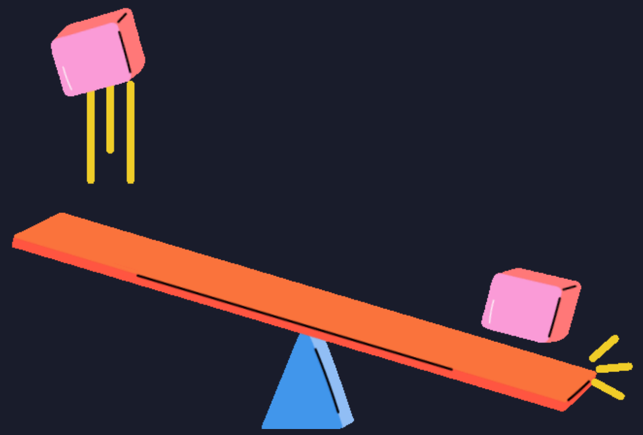
identify_bad_bot_traffic Signals

Known bad bots Signals
- Signal 1: the user-agent dynamically matches a string in a publicly available list of bad bots.

Libraries and net tools Signals
- Signal 1: the string curl is present with its corresponding version
- Signal 2: the string python is present with its corresponding library name and version
- Signal 3: the string "Postman" is present with its corresponding version

Path traversal attacks Signals
- Signal 1: fingerprintRequestUrl containing "/../"

identify_bad_bot_traffic Signals Considerations



Timestamps

Using timestamps as part of traffic analysis yields to more accurate behavioral and temporal analysis.

Dynamic IP Verification

https://original-domain.com/bots/ip/<ip_address>

Other Known Attack Schemes

SQLi, XSS, and more

Bad Bot Definition

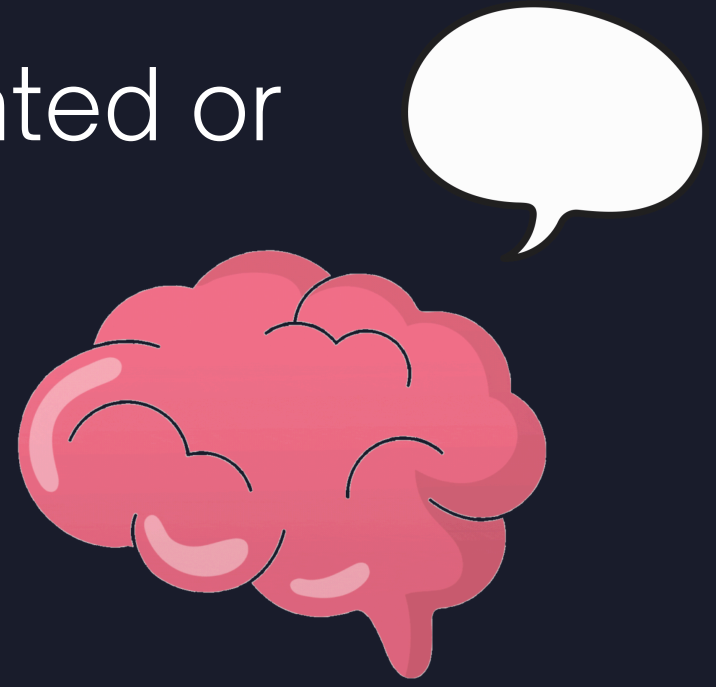
product-specific, context-dependent, ever-evolving

identify_human_traffic Signals

Toma Taylor Makoundou | August 2022

identify_human_traffic Signals

This function detects human activity by eliminating unwanted or malicious traffic.



Heuristic Algorithm

(¬identify_google_bot)^(¬identify_bad_bot_traffic)

=(¬google_bots_signals)^(¬fake_google_bots_signals)^(¬non_identified_bots_signals)^(¬libraries_and_net_tools)^(¬path_traversal_attacks)

identify_human_traffic Signals

Negative of Fake Bots Signals

- Signal 1: the user-agent value contains Googlebot
- Signal 2: the IP address does not belong to Google
- Signal 3: existing user-agent

Negative of Known Bad Bots Signals

- Signal 1: the user-agent does not dynamically match a string in a publicly available list of bad bots.

identify_human_traffic Signals

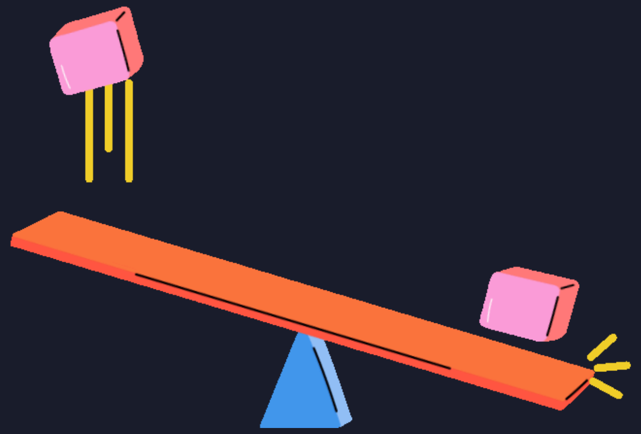
Negative of libraries and net tools Signals

- Signal 1: the string curl is not present with its corresponding version in fingerprintUserAgent
- Signal 2: the string python is not present with its corresponding library name and version in fingerprintUserAgent
- Signal 3: the string Postman is not present with its corresponding version in fingerprintUserAgent

Negative of Path Traversal Attack Signals

- Signal 1: the string "/../" is not in fingerprintRequestUrl

identify_human_traffic Signals Considerations



Human Traffic is Complex

Human traffic depends on and is influenced by many external factors.

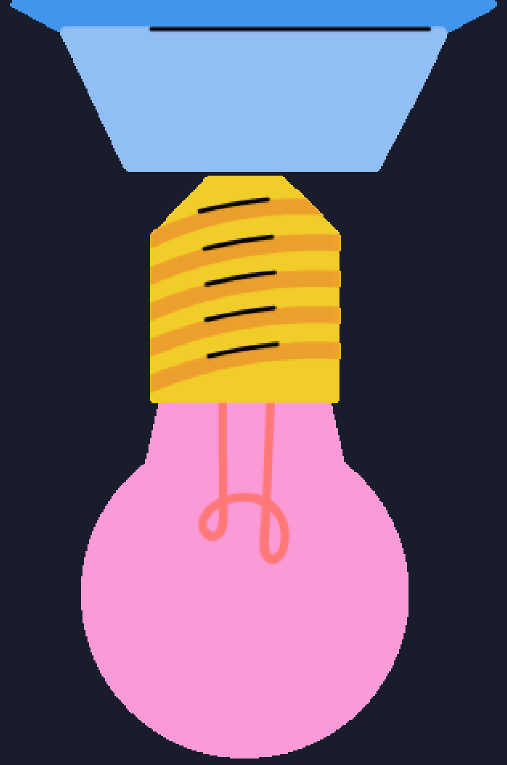
Nature of the Api Endpoint

A RESTful APIs and websites expect different types of traffic.

Timestamps

Timestamps can help differentiate automated traffic from human activity via trend analysis.

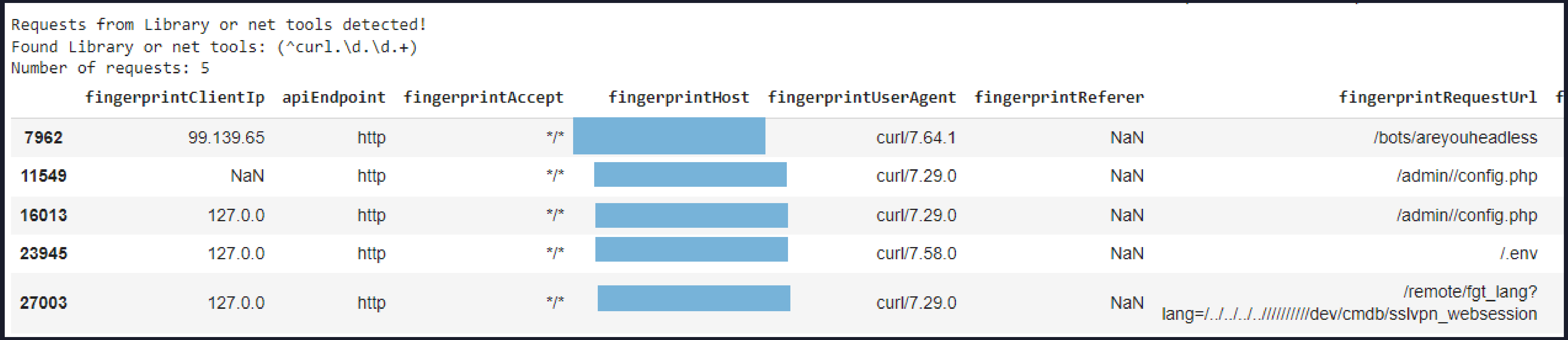
Toma Taylor Makoundou | August 2022



Interesting findings

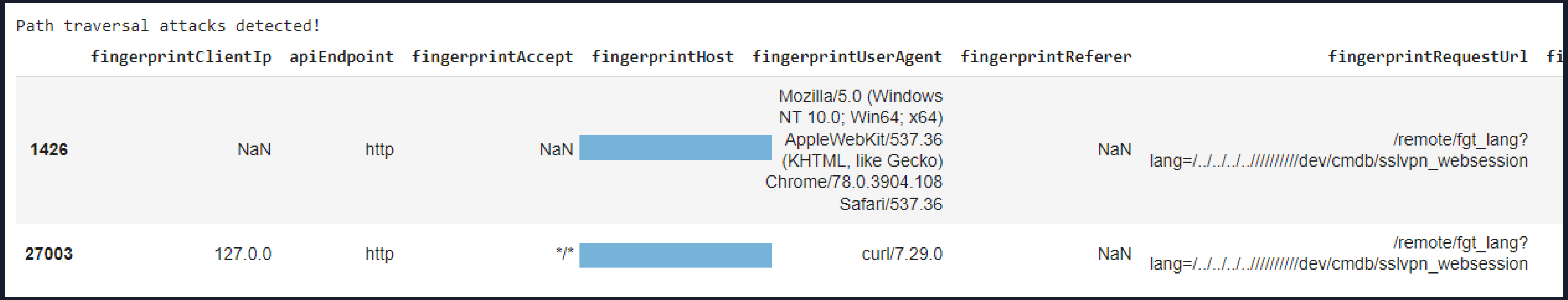
Attempt to exploit CVE-2018-13379

- **Net tool Name:** Curl
- **Malicious activity:**
  - probes admin pages
  - attempts to exploit CVE-2018-13379



```
Requests from Library or net tools detected!
Found Library or net tools: (^curl.\d.\d.+)
Number of requests: 5
```

	fingerprintClientIp	apiEndpoint	fingerprintAccept	fingerprintHost	fingerprintUserAgent	fingerprintReferer	fingerprintRequestUrl	f
7962	99.139.65	http	*/*	[redacted]	curl/7.64.1	NaN	/bots/areyouheadless	
11549	NaN	http	*/*	[redacted]	curl/7.29.0	NaN	/admin//config.php	
16013	127.0.0	http	*/*	[redacted]	curl/7.29.0	NaN	/admin//config.php	
23945	127.0.0	http	*/*	[redacted]	curl/7.58.0	NaN	/.env	
27003	127.0.0	http	*/*	[redacted]	curl/7.29.0	NaN	/remote/fgt_lang? lang=/../../../..//////////dev/cmdb/sslvpn_websession	



```
Path traversal attacks detected!
```

	fingerprintClientIp	apiEndpoint	fingerprintAccept	fingerprintHost	fingerprintUserAgent	fingerprintReferer	fingerprintRequestUrl	fi
1426	NaN	http	NaN	[redacted]	Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.108 Safari/537.36	NaN	/remote/fgt_lang? lang=/../../../..//////////dev/cmdb/sslvpn_websession	
27003	127.0.0	http	*/*	[redacted]	curl/7.29.0	NaN	/remote/fgt_lang? lang=/../../../..//////////dev/cmdb/sslvpn_websession	

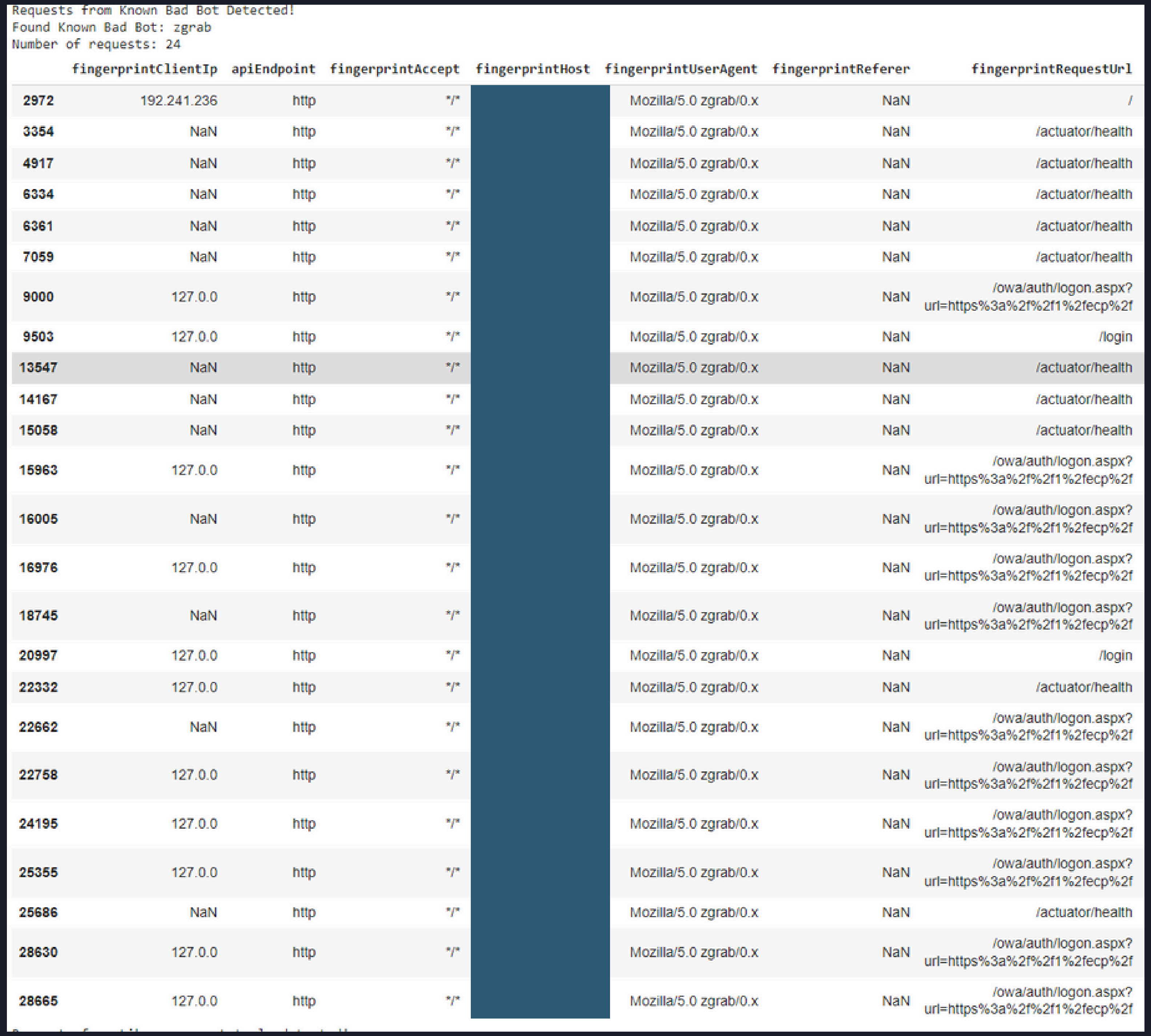
Attempt to exploit CVE-2018-13379

- CVSS 3.x score: 9.8
- A path traversal vulnerability in the FortiOS SSL VPN web portal.
- When successfully exploited, the vulnerability allows an attacker to access Fortinet FortiOS, leak files and read login/passwords in clear text.

- The exploit is publicly available
  - https://gist.github.com/code-machina/bae5555a771062f2a8225fd4731ae3f7
  - https://www.exploit-db.com/exploits/47288

Other Malicious Known Bad Bot Requests

Bad Bot Name: Moblie Safari
Malicious Activity:
- attempt to perform a WordPress 5.1.1 Slider Revolution 4.6.5 UpdateCaptionsCSS Remote Content Injection
- probes for environment variables
- probes easy-wp-smtp plugin



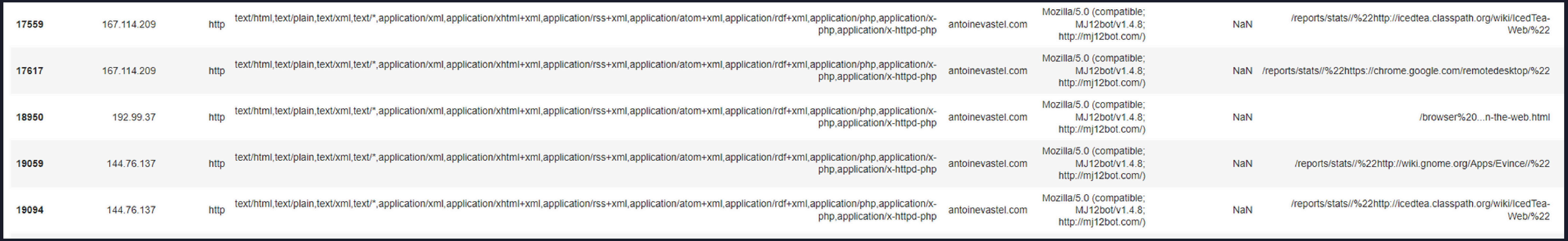
```
Requests from Known Bad Bot Detected!
Found Known Bad Bot: zgrab
Number of requests: 24
```

	fingerprintClientIp	apiEndpoint	fingerprintAccept	fingerprintHost	fingerprintUserAgent	fingerprintReferer	fingerprintRequestUrl
2972	192.241.236	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/
3354	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
4917	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
6334	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
6361	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
7059	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
9000	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
9503	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/login
13547	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
14167	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
15058	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
15963	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
16005	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
16976	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
18745	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
20997	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/login
22332	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
22662	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
22758	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
24195	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
25355	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
25686	NaN	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/actuator/health
28630	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f
28665	127.0.0	http	*/*	[redacted]	Mozilla/5.0 zgrab/0.x	NaN	/owa/auth/logon.aspx?url=https%3a%2f%2f1%2fecp%2f

Other Malicious Known Bad Bot Requests

Bad Bot Name: MJ12bot

Malicious Activity: Attempt to access remote servers

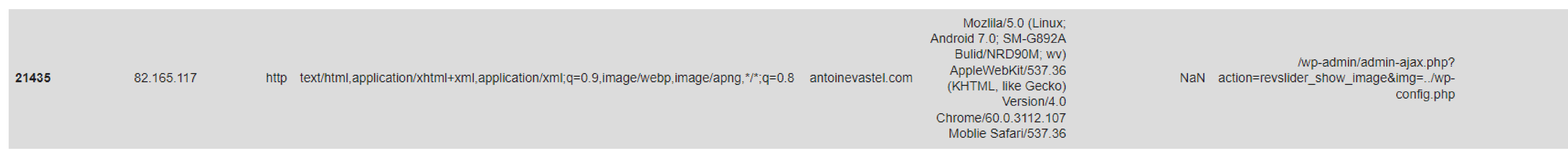


17559	167.114.209	http	text/html,text/plain,text/xml,text/*,application/xml,application/xhtml+xml,application/rss+xml,application/atom+xml,application/rdf+xml,application/php,application/x-php,application/x-httpd-php	antoinevastel.com	Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)	NaN	/reports/stats//%22http://icedtea.classpath.org/wiki/IcedTea-Web/%22
17617	167.114.209	http	text/html,text/plain,text/xml,text/*,application/xml,application/xhtml+xml,application/rss+xml,application/atom+xml,application/rdf+xml,application/php,application/x-php,application/x-httpd-php	antoinevastel.com	Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)	NaN	/reports/stats//%22https://chrome.google.com/remotedesktop/%22
18950	192.99.37	http	text/html,text/plain,text/xml,text/*,application/xml,application/xhtml+xml,application/rss+xml,application/atom+xml,application/rdf+xml,application/php,application/x-php,application/x-httpd-php	antoinevastel.com	Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)	NaN	/browser%20...n-the-web.html
19059	144.76.137	http	text/html,text/plain,text/xml,text/*,application/xml,application/xhtml+xml,application/rss+xml,application/atom+xml,application/rdf+xml,application/php,application/x-php,application/x-httpd-php	antoinevastel.com	Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)	NaN	/reports/stats//%22http://wiki.gnome.org/Apps/Evince//%22
19094	144.76.137	http	text/html,text/plain,text/xml,text/*,application/xml,application/xhtml+xml,application/rss+xml,application/atom+xml,application/rdf+xml,application/php,application/x-php,application/x-httpd-php	antoinevastel.com	Mozilla/5.0 (compatible; MJ12bot/v1.4.8; http://mj12bot.com/)	NaN	/reports/stats//%22http://icedtea.classpath.org/wiki/IcedTea-Web/%22

Other Malicious Known Bad Bot Requests

Bot Name: MicroMessenger
Malicious Activity: bots searching for vulnerable plugins

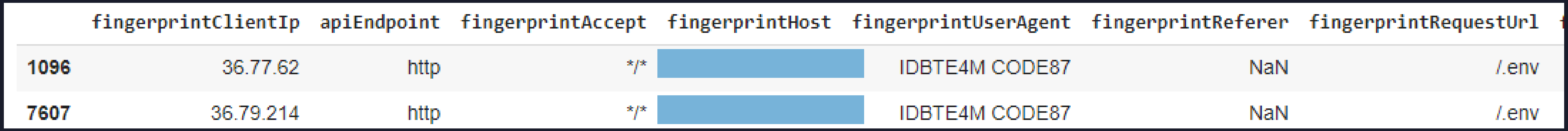
Source:
https://core.trac.wordpress.org/ticket/48049



21435	82.165.117	http	text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8	antoinevastel.com	Mozlila/5.0 (Linux; Android 7.0; SM-G892A Bulid/NRD90M; wv) AppleWebKit/537.36 (KHTML, like Gecko) Version/4.0 Chrome/60.0.3112.107 Moblie Safari/537.36	NaN	/wp-admin/admin-ajax.php?action=revslider_show_image&img=../wp-config.php

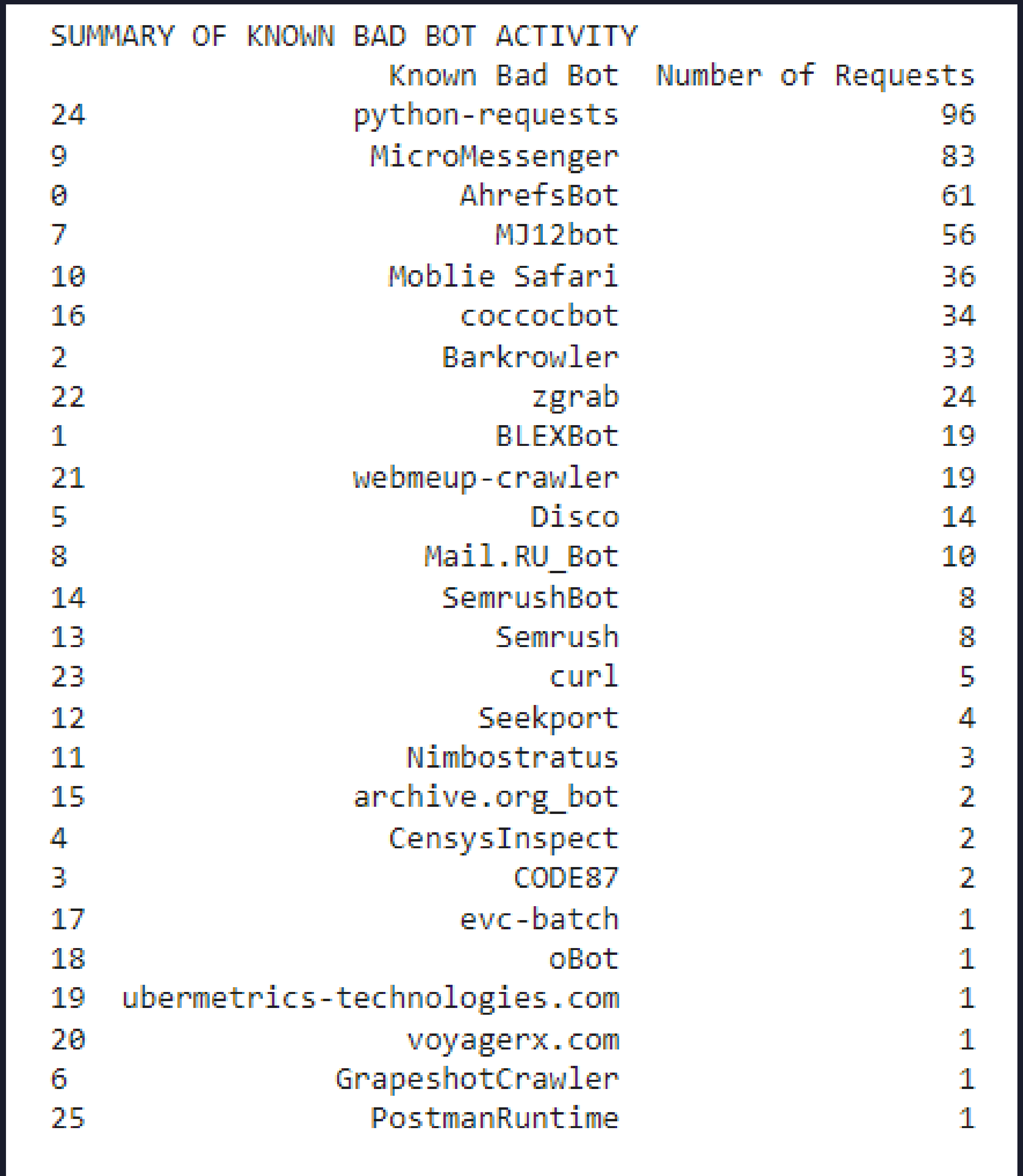
Other Malicious Known Bad Bot Requests

Bot Name: CODE87
Malicious Activity: attempt to remotely enumerate environment variables on antoinevastel.com



	fingerprintClientIp	apiEndpoint	fingerprintAccept	fingerprintHost	fingerprintUserAgent	fingerprintReferer	fingerprintRequestUrl	f
1096	36.77.62	http	*/*	[redacted]	IDBTE4M CODE87	NaN	/.env	
7607	36.79.214	http	*/*	[redacted]	IDBTE4M CODE87	NaN	/.env	

Known Bad Bot Activity Summary



```
SUMMARY OF KNOWN BAD BOT ACTIVITY
                   Known Bad Bot  Number of Requests
24               python-requests                  96
9                 MicroMessenger                  83
0                      AhrefsBot                  61
7                        MJ12bot                  56
10                 Moblie Safari                  36
16                     coccocbot                  34
2                     Barkrowler                  33
22                         zgrab                  24
1                        BLEXBot                  19
21               webmeup-crawler                  19
5                          Disco                  14
8                    Mail.RU_Bot                  10
14                    SemrushBot                   8
13                       Semrush                   8
23                          curl                   5
12                      Seekport                   4
11                  Nimbostratus                   3
15               archive.org_bot                   2
4                  CensysInspect                   2
3                         CODE87                   2
17                     evc-batch                   1
18                          oBot                   1
19  ubermetrics-technologies.com                   1
20                  voyagerx.com                   1
6               GrapeshotCrawler                   1
25                PostmanRuntime                   1
```

- **Top 3 Known Bad Bots**
  - python-requests*
  - MicroMessenger
  - AhrefsBot

*may be a false positive

References

- https://developers.google.com/search/blog/2015/01/crawling-and-indexing-of-locale
- https://developers.google.com/search/docs/advanced/crawling/googlebot
- https://raw.githubusercontent.com/mitchellkrogza/hnginx-ultimate-bad-bot-blocker/master/_generator_lists/bad-user-agents.list

Thank you for listening!

Toma Taylor Makoundou | August 2022