

Scraping For Malicious Intent

The Moz Top 500 Domains

-(As ranked by Mozscape web index)

Why? Needed a free, big list of current websites that people are actually visiting. (Need much more in hindsight)

The Moz Top 500 Domains

Top 500 Domains		Top 500 Pages				
Rank	Root Domain	Linking Root Domains	External Links	Domain mozRank	Domain mozTrust	Change
1	Facebook.com	7,143,012	4,271,085,373	9.57 <div><div></div></div>	9.36 <div><div></div></div>	0
2	Twitter.com	5,413,688	5,409,190,546	9.60 <div><div></div></div>	9.35 <div><div></div></div>	0
3	Google.com	4,572,849	2,877,006,721	9.36 <div><div></div></div>	9.28 <div><div></div></div>	0
4	Youtube.com	2,692,759	1,678,353,826	9.09 <div><div></div></div>	9.14 <div><div></div></div>	0
5	Linkedin.com	1,793,083	880,269,412	8.97 <div><div></div></div>	8.91 <div><div></div></div>	0
6	Wordpress.org	1,697,617	183,423,120	9.06 <div><div></div></div>	8.70 <div><div></div></div>	0
7	Instagram.com	1,533,396	1,163,216,206	8.89 <div><div></div></div>	8.80 <div><div></div></div>	0
8	Pinterest.com	1,111,548	782,575,001	8.70 <div><div></div></div>	8.62 <div><div></div></div>	0

The Moz Top 500 Domains: Scraping

Straightforward scrape:

- Static page, so
 - BeautifulSoup.findAll("td",{“class”:“url”})
 - tag.a[“href”] = url_to_save
- Save to file and move on

Website of Interest (2 of 2)

urlQuery.net

(a service for detecting web-based malware)

Why? It's a free scanner that doesn't have a captcha to get past :)

urlQuery.net (ex. with Kickstarter.com)

urlQuery

Search

Statistics

About

Login

urlQuery.net is a service for detecting and analyzing web-based malware. It provides detailed information about the activities a browser does while visiting a site and presents the information for further analysis.

Learn about the [advanced settings](#)

Profile URL:

▼ Advanced settings:

User Agent:

Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.13) Gecko/20101203 Firefox/3.6.13 ▼










Referer:

Adobe Reader:

8.0 ▼

Java:

1.6.0_ ▼

Date (CET)	UQ / IDS / BL	URL	IP
2017-05-01 23:37:48	0 - 0 - 4	fest5weight-loss.com/	 64.22.95.115
2017-05-01 23:37:42	0 - 0 - 2	akusajagames.blogspot.co.id/search/label/Asian%20Sex%20Video%20gratis/	 216.58.201.161
2017-05-01 23:37:38	0 - 0 - 0	www.liveinternet.ru/users/raidicrebe1981/post400486891/	 88.212.202.38
2017-05-01 23:37:38	0 - 0 - 0	www.liveinternet.ru/users/carfimacha1970/post401456407/	 88.212.202.35
2017-05-01 23:37:36	0 - 0 - 1	portalsantoandreemfoco.com.br/index.php/entretenimento/34960-horoscopo-do-dia-29-de-janeir (...)	 108.179.253.227
2017-05-01 23:37:34	0 - 0 - 1	afriz.com/wp-content/plugins/aindex.php?email=belinda@mentbros.com	 38.110.76.133
2017-05-01 23:37:31	0 - 0 - 0	www.live-stream4ktv.com/2017/05/01/capitals-vs-penguins/	 164.132.190.162
2017-05-01 23:37:30	0 - 0 - 0	market.android.comhttps:///details?id=com.app.sendspace.mobile	 64.233.161.118
2017-05-01 23:37:27	0 - 0 - 1	download.drp.su/17-online/DriverPack-17-Online.exe	 88.150.137.207

urlQuery.net (ex. with Kickstarter.com)

urlQuery



Search

Statistics

About

Login

Overview

URL	kickstarter.com	
IP	54.230.202.86	
ASN	AS16509 Amazon.com, Inc.	
Location	 United States	
Report completed	2017-05-01 23:43:29 CET	
Status	Report complete.	
urlQuery Alerts	No alerts detected	


Settings

UserAgent	Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.13) Gecko/20101203 Firefox/3.6.13
Referer	
Pool	
Access Level	public

ACCOMPLISH MORE

With fiber Internet

GET MORE PRODUCTIVE >

Spectrum
ENTERPRISE

Intrusion Detection Systems

Snort /w Sourcefire VRT	No alerts detected
Suricata /w Emerging Threats Pro	No alerts detected

Blacklists

Fortinet's Web Filter / fortiguard.com	Added / Verified	Severity	Host	Comment
	2017-05-01	2	dtliltzwpaww.cloudfront.net/s.js	Malware
MDL / malwaremainlist.com	No alerts detected			
DNS-BH / malwaredomains.com	No alerts detected			
mnemonic secure DNS / mnemonic.no	No alerts detected			
OpenPhish / openphish.com	No alerts detected			
PhishTank / phishtank.com	No alerts detected			
Spamhaus DBL / spamhaus.org	No alerts detected			

urlQuery.net (ex. with Kickstarter.com)

← → ↻



urlquery.net/report.php?id=1493672818978

☆







Files Captured

Recent reports on same IP/ASN/Domain







Last 2 reports on IP: 54.230.202.86

Date	UQ / IDS / BL	URL	IP
2017-04-21 01:37:43	0 - 0 - 0	cdn.volunteermatch.org/www/images/emails/affinity_ym/bg_texture_tile.jpg	 54.230.202.86
2017-04-21 01:30:48	0 - 0 - 0	cdn.volunteermatch.org/www/images/affinity_email/guest/corner_top_right.gif	 54.230.202.86

Last 6 reports on ASN: AS16509 Amazon.com, Inc.

Date	UQ / IDS / BL	URL	IP
2017-05-01 23:41:07	0 - 0 - 1	kickstarter.com	 54.230.202.131
2017-05-01 23:39:43	0 - 0 - 1	kickstarter.com	 54.230.202.131
2017-05-01 23:28:22	0 - 0 - 0	groups.diligo.comhttps:///group/phoenixuniversityonline/content/watch-online-stream-magda-linett (...)	 54.148.192.94
2017-05-01 23:27:46	0 - 0 - 0	static-assets.cf.socialware.comhttps:///images/linkedin-profile-icon.png	 54.230.202.179
2017-05-01 23:27:19	0 - 0 - 0	mortgage-brokers.credio.com/ajax_search?_len=20&page=1&app_id=2927&sortdir=ASC& (...)	 52.8.128.181
2017-05-01 23:26:50	0 - 0 - 1	ime.cdn.service.cootek.com/default/market/apk/IME_Dialer.apk	 54.230.202.135

Last 6 reports on domain: kickstarter.com

Date	UQ / IDS / BL	URL	IP
2017-05-01 23:41:07	0 - 0 - 1	kickstarter.com	 54.230.202.131
2017-05-01 23:39:43	0 - 0 - 1	kickstarter.com	 54.230.202.131
2017-05-01 20:31:50	0 - 0 - 1	kickstarter.com	 54.192.3.231
2017-05-01 20:28:42	0 - 0 - 1	kickstarter.com	 54.192.3.231
2017-04-30 09:02:43	0 - 0 - 1	kickstarter.com/	 52.222.157.122
2017-04-29 19:04:12	0 - 0 - 1	kickstarter.com	 54.192.3.231

JavaScript

Executed Scripts (28)

Executed Evals (0)

Executed Writes (1)

#1 JavaScript:Write (size: 34, repeated: 1)

ManageEngine

ServiceDesk Plus

FREE

IT Help Desk Software

Download Now

HTTP Transactions (62)

More complicated scrape

Page redirects on submit:

main page-> queued page-> report page (then JavaScript dynamically creates the page after page outline loaded)

So, use Selenium WebDriver to programmatically submit urls and handle redirects and JavaScript . Then just save the fully rendered HTML file for continued processing.

Continued

Saved HTML file -> BeautifulSoup for elements of interest.
Date elements saved (20):

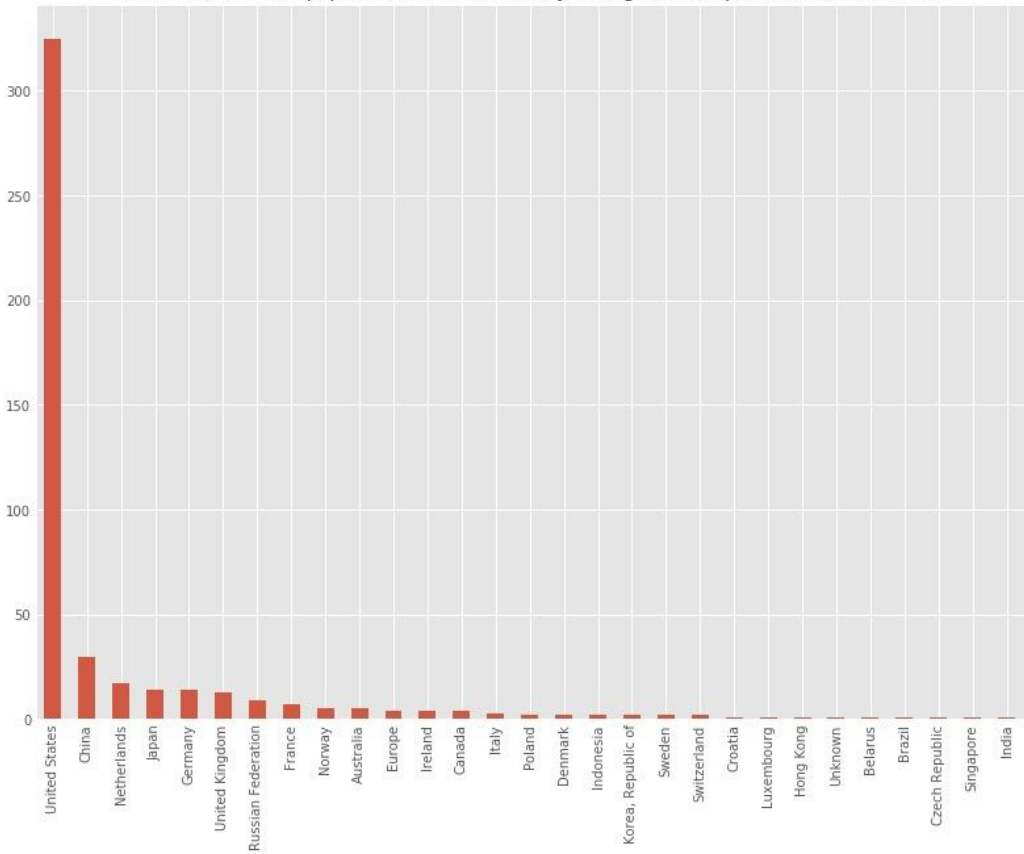
- Url (string)
- IP Address (string)
- ASN (string)
- IP Country (string)
- Report Date (string)
- User Agent (string)
- # Http Tranxs (numeric)
- JS Executed Scripts (numeric)
- JS Executed Writes (numeric)
- JS Executed Evals (numeric)
- ☐ UrlQuery alerts (binary)
- ☐ Snort (binary)
- ☐ Suricata (binary)
- ☐ Fortinet (binary)
- ☐ MDL (binary)
- ☐ DNS BH (binary)
- ☐ MS DNS (binary)
- ☐ Openfish (binary)
- ☐ Phishtank (binary)
- ☐ Spamhaus (binary)

For this sample:

- 476 observations
- Missing observations from 500 due to various reasons:
 - Blocked IP
 - Original url pattern (<http://domain.com>) incorrect
 - domain only had https but did not redirect
 - Not an end user domain (name server,etc)
 - Site down

Basic Analysis

Distribution of most popular domains to country of origin for sample taken on 4/17/2017



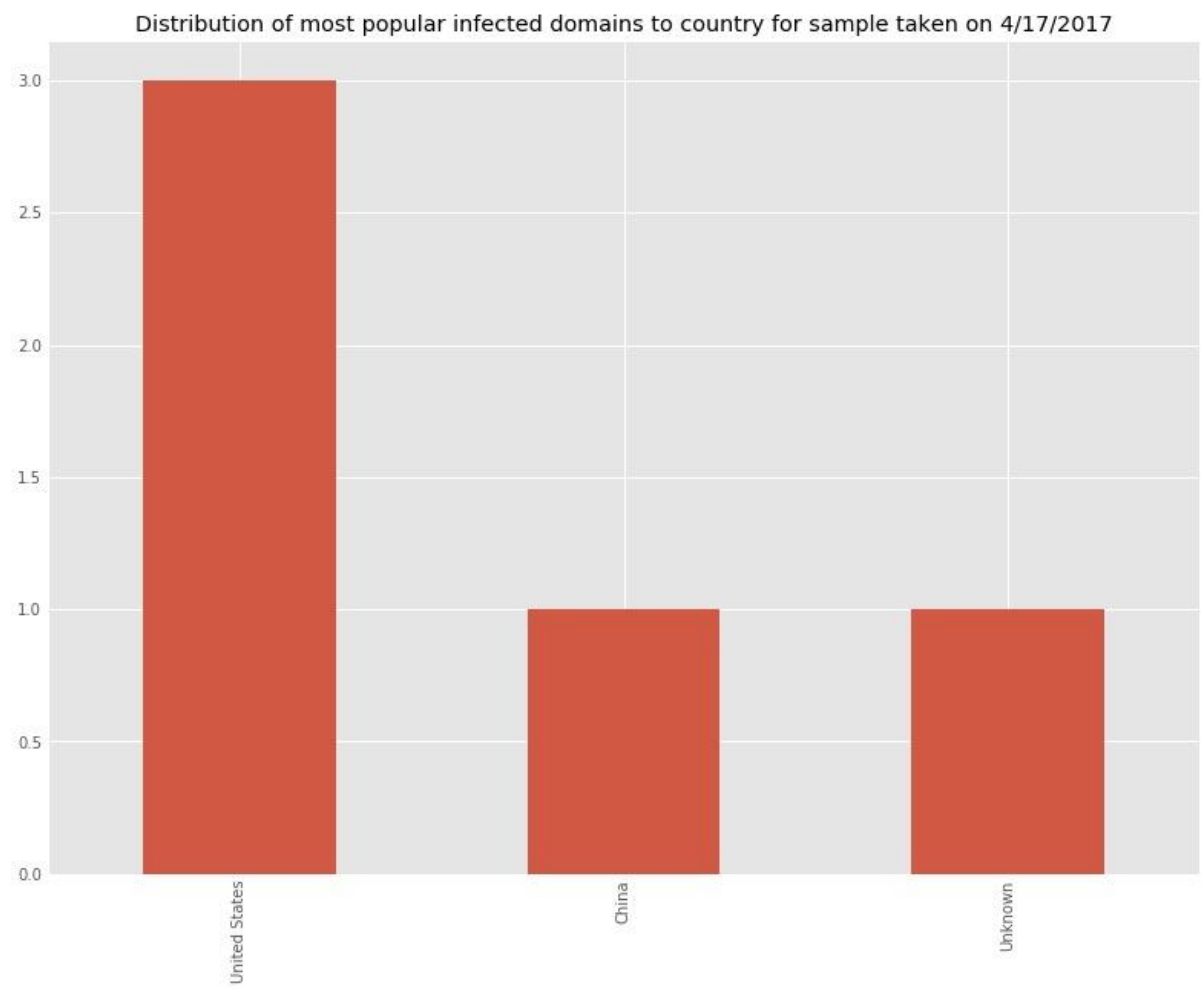
Basic Analysis

```
website[(website["UrlQuery.Alerts"] != 0) | (website["Snort"] != 0) | (website["Suricata"] != 0) |
(website["Fortinet"] != 0) | (website["MDL"] != 0) | (website["DNS.BH"] != 0) |
(website["MS.DNS"] != 0) | (website["Openfish"] != 0) | (website["Phishtank"] != 0) | (website["Spamhaus"] != 0)]
```

Out[64]:

	Url	IPAddress	ASN	IP.Location	Report.Date	UrlQuery.Alerts	User.Agent	Snort	Suricata	Fortinet	MDL	DI
116	http://nih.gov	54.235.145.223	AS14618 Amazon.com, Inc.	United States	2017-04-28 23:02:36 CET	0	Mozilla/5.0 (Windows; U; Windows NT 6.1; en- US...	0	0	1	0	0
303	http://foursquare.com	151.101.192.154	AS6983 Earthlink, Inc.	United States	2017-04-29 17:59:42 CET	0	Mozilla/5.0 (Windows; U; Windows NT 6.1; en- US...	0	0	1	0	0
329	http://technorati.com	208.66.66.66	AS16936 Technorati, Inc.	Unknown	2017-04-29 20:44:30 CET	0	Mozilla/5.0 (Windows; U; Windows NT 6.1; en- US...	0	0	1	0	0
377	http://kickstarter.com	54.192.3.231	AS16509 Amazon.com, Inc.	United States	2017-04-29 19:04:12 CET	0	Mozilla/5.0 (Windows; U; Windows NT 6.1; en- US...	0	0	1	0	0
454	http://51.la	111.74.238.204	AS4134 Chinanet	China	2017-04-29 20:57:14 CET	0	Mozilla/5.0 (Windows; U; Windows	0	0	1	0	0

Basic Analysis



Basic Analysis

```
In [83]: # what's the median value of http transactions for sites flaged as potentially infected?
website[(website["Fortinet"] ==1)][ "HTTP.Tranx"].median(axis=0)
```

Out[83]: 62.0

```
In [84]: # what's the median value of http transactions for sites not flaged as potentially infected?
website[(website["Fortinet"] ==0)][ "HTTP.Tranx"].median(axis=0)
```

Out[84]: 66.0

```
In [86]: # what's the median value of javascript activity for infected vs not infected?
website.groupby("Fortinet")['JS.ES','JS.EE','JS.EW'].median()
```

Out[86]:

	JS.ES	JS.EE	JS.EW
Fortinet			
0	24	0	0
1	23	0	0

```
In [ ]:
```

Future Enhancements

- #! more observations
- Faster processing times (multi-process, multi-thread)
- Textual data from the alerts (what IPs from infection vector, redirects?)
- Could JavaScript complexity be a tell? Or maybe JavaScript DNA?