Ads networks are following you, follow them back

(The web is even worse than you thought)

CIRCL
Computer Incident
Response Center
Luxembourg

Quinn Norton - @quinnnorton
Raphaël Vinot - @rafi0t

https://www.circl.lu

2018-03-15

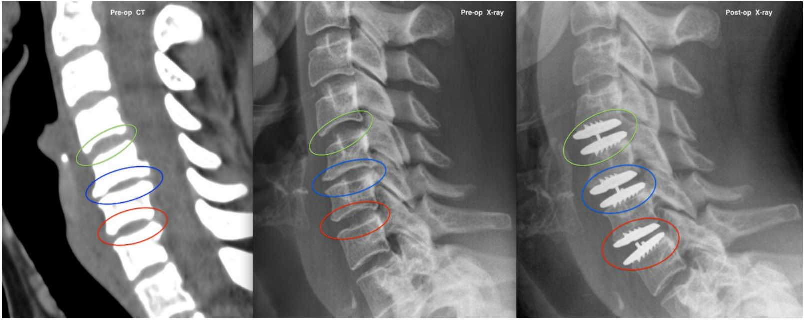
Who are we

Quinn Norton

- Freelance journalist & writer
- Former (kinda) UI/UX
- Infosec trainer

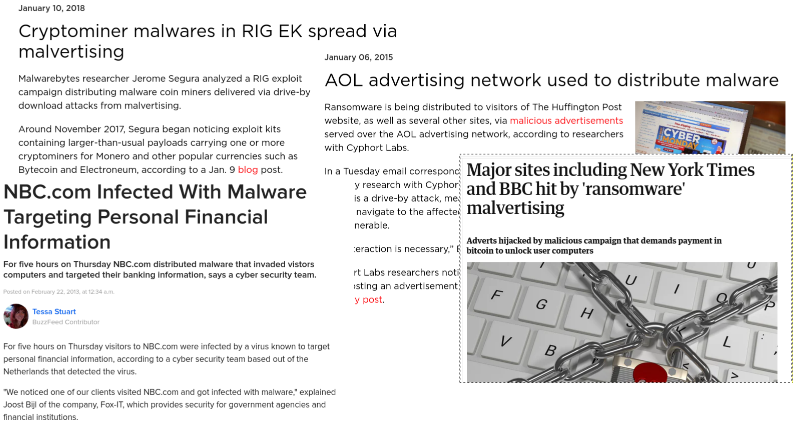
Raphaël Vinot

- Incident responder @ CIRCL.lu
- Developer
- Infosec trainer


Pre-op CT
Pre-op X-ray
Post-op X-ray

Hyd

Origin of the project



January 10, 2018

Cryptominer malwares in RIG EK spread via malvertising

Malwarebytes researcher Jerome Segura analyzed a RIG exploit campaign distributing malware coin miners delivered via drive-by download attacks from malvertising.

Around November 2017, Segura began noticing exploit kits containing larger-than-usual payloads carrying one or more cryptominers for Monero and other popular currencies such as Bytecoin and Electroneum, according to a Jan. 9 blog post.

NBC.com Infected With Malware Targeting Personal Financial Information

For five hours on Thursday NBC.com distributed malware that invaded vistors computers and targeted their banking information, says a cyber security team.

Posted on February 22, 2013, at 12:34 a.m.

Tessa Stuart
BuzzFeed Contributor

For five hours on Thursday visitors to NBC.com were infected by a virus known to target personal financial information, according to a cyber security team based out of the Netherlands that detected the virus.

"We noticed one of our clients visited NBC.com and got infected with malware," explained Joost Bijl of the company, Fox-IT, which provides security for government agencies and financial institutions.

January 06, 2015

AOL advertising network used to distribute malware

Ransomware is being distributed to visitors of The Huffington Post website, as well as several other sites, via malicious advertisements served over the AOL advertising network, according to researchers with Cyphort Labs.

In a Tuesday email correspond
y research with Cyphor
is a drive-by attack, me
navigate to the affecte
nerable.

:eraction is necessary," R

rt Labs researchers noti
›sting an advertisement
y post.

Major sites including New York Times and BBC hit by 'ransomware' malvertising

Adverts hijacked by malicious campaign that demands payment in bitcoin to unlock user computers

The lawyers' reply

¯_(ツ)_/¯

"*long look at each other* *pause* yeeeeahhhh..... *shrug* Can you help us?"

Our answer

(͡° ͜ʖ ͡°)

looked at each other *looked back at them* and said "...We'll get back to you on that"

Current situation

- Very complex and huge websites (often close to 10mb for the front page)
- Extremely dynamic
- Dozens of 3rd party components
- ... which may pay the bills, or keep the site going
- No tools to audit such a website (please prove me wrong)

Day to day CERT work

- Phishing websites are super common
- They are also often relatively simple
- ... unless they're not (i.e. dynamically generated JS, chained redirects)
- Reproducing is painful (i.e. User Agent, timing, source IP)
- We like to have the newest browser, using an older one is annoying

Requirements

- Complete emulation of a browser (JS, iFrames, redirects, cookies, headers)
- Keep the dataset for analysis later, screenshot of the page, full HTML
- Easy to deploy
- Flexible way to pass parameters to the query
- Legit browser, not IE6 in virtualbox
- Something a human can use efficiently

Splash and Scrapy

- Instrument a recent webkit (Chrome/Chromium)
- Let you define a user-agent
- Can take a screenshot of the website
- Comes in a docker image
- Killer feature: Returns a HTTP Archive (HAR)

Available as a standalone python3 module for your own project:
`https://github.com/viper-framework/ScrapySplashWrapper`

HTTP Archive

- List all the requests and all the responses
- Including headers, cookies, and redirects
- But also every body of every response
- ...and that means hundreds of unique entries

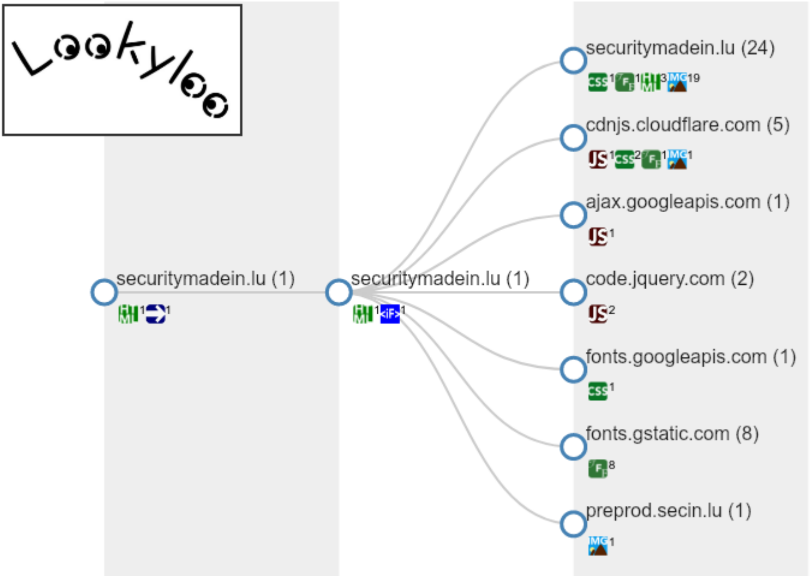
Ben Watts – https://www.flickr.com/photos/benwatts/4087289013

Digging into the HAR file

Two things stand out and look like a good starting point:

- redirectURL (the location key in the HTTP header)
  - URL1 redirects to URL2
- The referrer key in the HTTP headers
  - All the URLs with the referrer key set are loaded from that one

Sounds like we could built a tree, right?


Lookyloo
securitymadein.lu (1)
1 1
securitymadein.lu (1)
1 <iF> 1
securitymadein.lu (24)
CSS 1 F 1 3 IMG 19
cdnjs.cloudflare.com (5)
JS 1 CSS 2 F 1 IMG 1
ajax.googleapis.com (1)
JS 1
code.jquery.com (2)
JS 2
fonts.googleapis.com (1)
CSS 1
fonts.gstatic.com (8)
F 8
preprod.secin.lu (1)
IMG 1

The beautiful things you find on webpages

Turns out the redirected URL can be any of these:

- Full URL
- URL **without the scheme** (http/https will be guessed)
- The path, **with or without "/"**
- **Just the parameters** (";..." attached to the path of the caller)
- **Just the query** ("?..."attached to the parameters)
- ...**port number** (just to mess with you)

And of course, the referrer header can be, and often is, stripped out.

T.J. Hawk – `https://www.flickr.com/photos/102627552@N04/25440096000`

iFrames to the rescue

Turns out iFrames didn't stay in the 90s. They...

- Can load more iFrames
- Can redirect to other pages, containing more iFrames
- Can contain JavaScript
- Can set/read cookies

Splash saves them in a tree-like format, so that's easy to attach.

The final touch: regexes!

No hellscape^Wsoftware project is complete without regexes, right?

- Search in each body for URL-like strings
- Lookup against the HAR entries
- Attach in tree when possible

.... And the few URLs I wasn't able to attach anywhere are connected to the root node as "orphans"

Tree capabilities

- Not reinventing the wheel: use ETE Toolkit (phylogenetic trees library)
- Each node has features: type of content, cookies, headers, full body
- Possible to search each features individually
- Get ancestors and children

I heard you like trees

Problem with the current tree:

- Too many URLs
- URLs are way too verbose
- Impossible to display efficiently

So let's make moar trees:

- Aggregate by hostname
- Aggregate features accordingly (cookies, content type)

Now available in a standalone python3 module:
`https://github.com/viper-framework/har2tree`

Aaand the web interface (aka The Glue)

- Overview of the hostnames
- Overview of what is loaded by which domain
- Collapse parts of the tree
- Expand hostnames to see the full URLs
- See details of each URL
- Download body loaded by a specific query

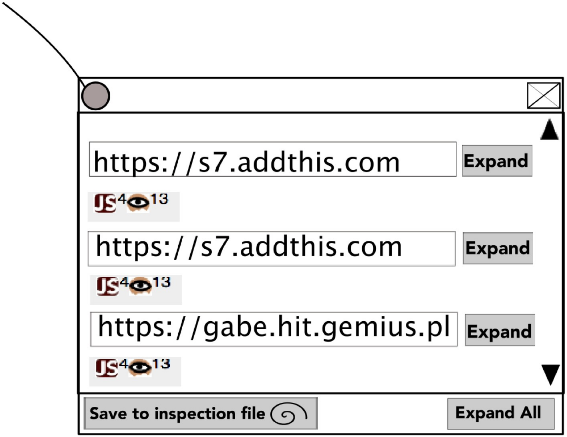
DEMO

`https://github.com/CIRCL/lookyloo`

`https://lookyloo.circl.lu`

Next steps

- New expansion box (Within existing trees)



https://s7.addthis.com

https://s7.addthis.com

https://gabe.hit.gemius.pl

Next steps

- Add more meta informations in the icons (iFrame, missing referer, content types)
- Automatic lookups against 3rd party services (VT, MISP, Phishtank)
- Compare runs with different User agents
- Add the possibility to crawl a website when logged-in
- Detect cookies set and read by different actor

References - Q&A

- Scrapping module: `https://github.com/viper-framework/ScrapySplashWrapper`
- Tree generator: `https://github.com/viper-framework/har2tree`
- Web interface: `https://github.com/CIRCL/lookyloo`
- Demo instance: `https://lookyloo.circl.lu`
- Contact: raphael.vinot@circl.lu - @rafi0t