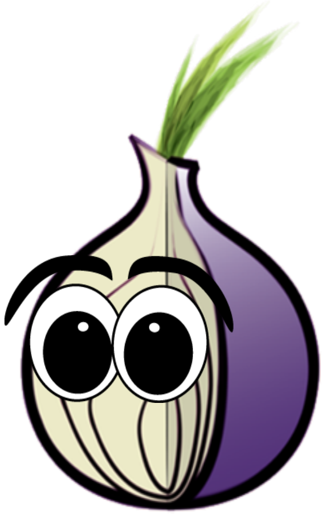
WWW 2017 DeustoTech EURECOM Sophia Antipolis

The Onions Have Eyes:



A Comprehensive Structure and Privacy Analysis of Tor Hidden Services

Iskander Sanchez-Rola, Davide Balzarotti, Igor Santos

Tor Hidden Services

- Provides anonymity through the **onion routing protocol**
- Tor has the largest number of users among the different types of Darknets

  Over 7000 relays

- Are used to provide access to different applications

  Such as chat, email, or websites

Motivation

- **Previous studies** about Tor hidden services have been focused on:

  Relay Analysis and Routing Analysis (e.g., Sanatinia et al. 2016)

  Criminal activity (e.g., Ciancaglini et al. 2015, Soska et al. 2015)

  Some studies about connectivity (OnionScan, 2016 & Deeplight, 2016)

Lack of a complete application-level structure analysis like in Surface Web

Lack of a complete privacy analysis

Our Work

The MOST complete exploration and crawl of Tor hidden services to date

- Comprehensive structure and privacy analysis
- Not only limited to home pages

  According to our data, home pages contain only:

  11% of links, 30% resources,

  21% of the scripts and 16% of tracking

- We crawl more than 1.5M of unique onion URLs

Analysis Platform (in a nutshell)

The ephemeral and isolated nature of onion sites makes crawling a challenge.

1) We manually collected a .onion URLS comprising 195,748 domains from 25 public forums and directories.

2) We implemented a specific crawler for web Tor hidden services

3) We perform a **structure analysis** regarding different connection types: links, resources, and redirections

4) We inspect the **privacy implications** of the connections and perform a measurement study of **web tracking** in Tor Dark Web

Design of the crawling phase

Crawler implementation based on PhantomJS

- Modified to hide its automatic nature from sites
- Can deal with script obfuscation (modification of JSBeautifier)

Two modes

- Collection mode
- Connectivity mode

Crawler - Collection mode

Data Retrieved

HTML headers , Redirections (+type)

HTML content, Scripts and Links

Crawling Strategy & Boundaries

3 levels of depth

10 links per each level → Prioritize : keywords & (link size + position)

Modifies the “referrer” to mimic user navigation

Crawler - Connectivity mode

Retrieved Data

Links (all of them: visible or invisible)

Not position ones: “#” or files (e.g., pdf, images)

Crawling Strategy & Boundaries

No limit in depth or links visited

Avoid the so called calendar effect: 10,000 URLs per each domain

Goal: capture the remaining structure not previously crawled

Size & Coverage

Domains Data

198,050 domains gathered → 7,257 were active domains

Confirmation of the ephemeral nature of onion sites

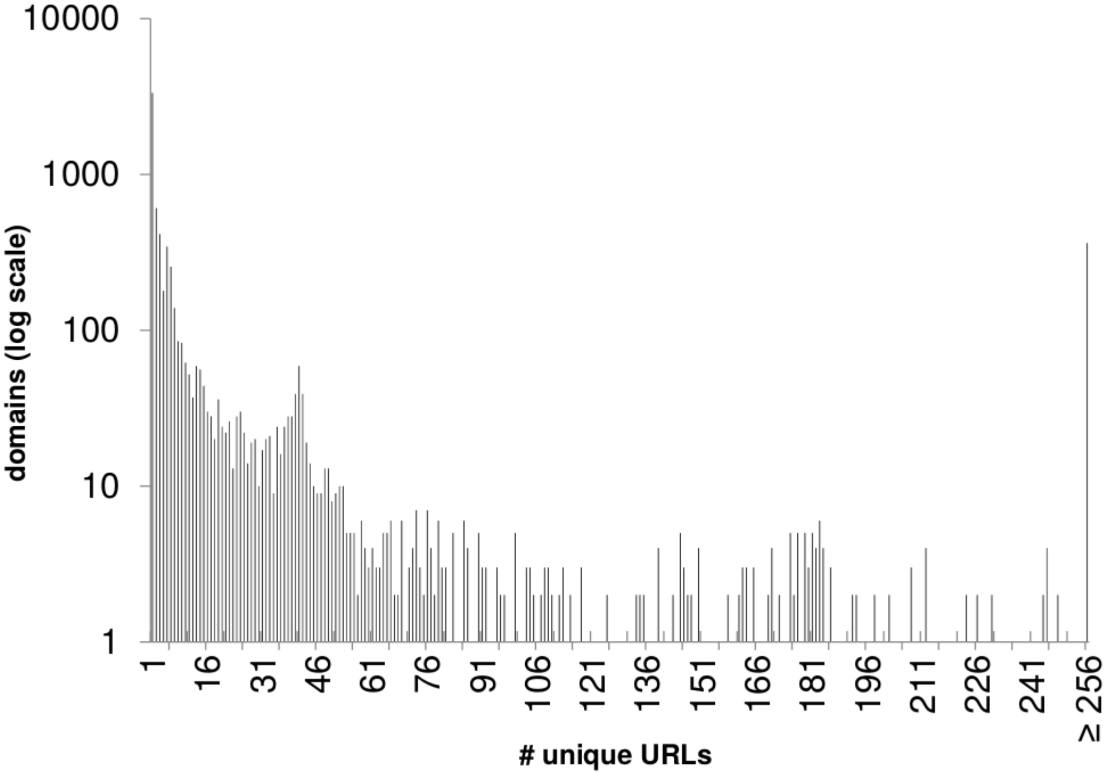
3 more crawling attempts (days and month of difference)

81.07% were completely crawled by the collection mode

18.49% were added by the connectivity mode

0.54% contained more than 10,000 URLs

Onion Domains/URL Distribution



46.07% of the domains contained just one URL

>80% of the domains less than 17 URLs

Language & Categories - Methodology

Languages

We use the Google Translate API

Categories

1) Translate the HTML plain text with Google Translate API

2) Remove stop words + stemming

3) Model as Bag of Words (Vector Space Model)

4) Clustering process with *Affinity Propagation*

5) Manual inspection of the clusters to find the category

Language Distributions

Language	% Domains
English	73.28%
Russian	10.96%
German	2.33%
French	2.15%
Spanish	2.14%

Ranking is similar to the surface web, with the omission of Japanese

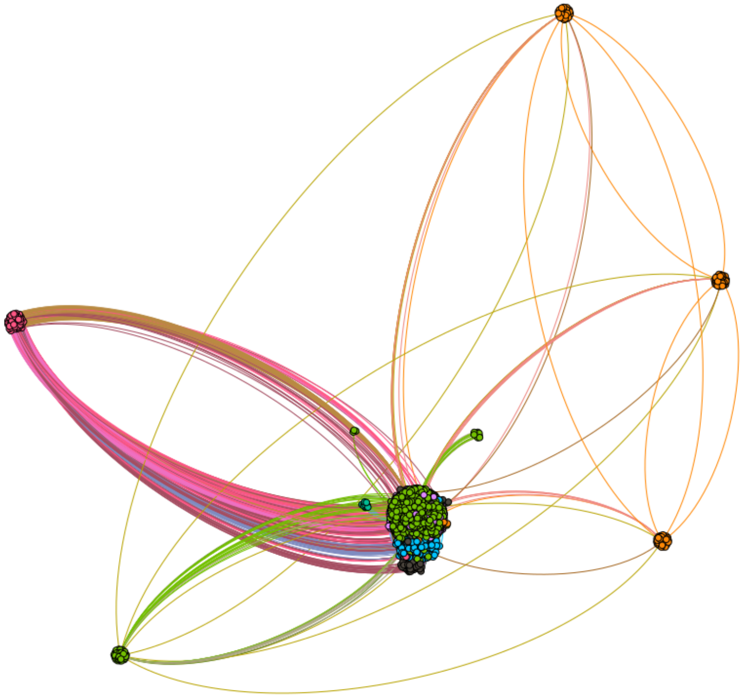
The ranking is different to other studies (Deeplight)

Category Distributions

Category	% Domains
Directory/Wiki	63.49%
Default Hosting Message	10.35%
Market/Shopping	9.80%
Bitcoins/Trading	8.62%
Forum	4.72%
Online Betting	1.72%
Search Engine	1.30%

15.4% of the domains belonged to more than 1 category

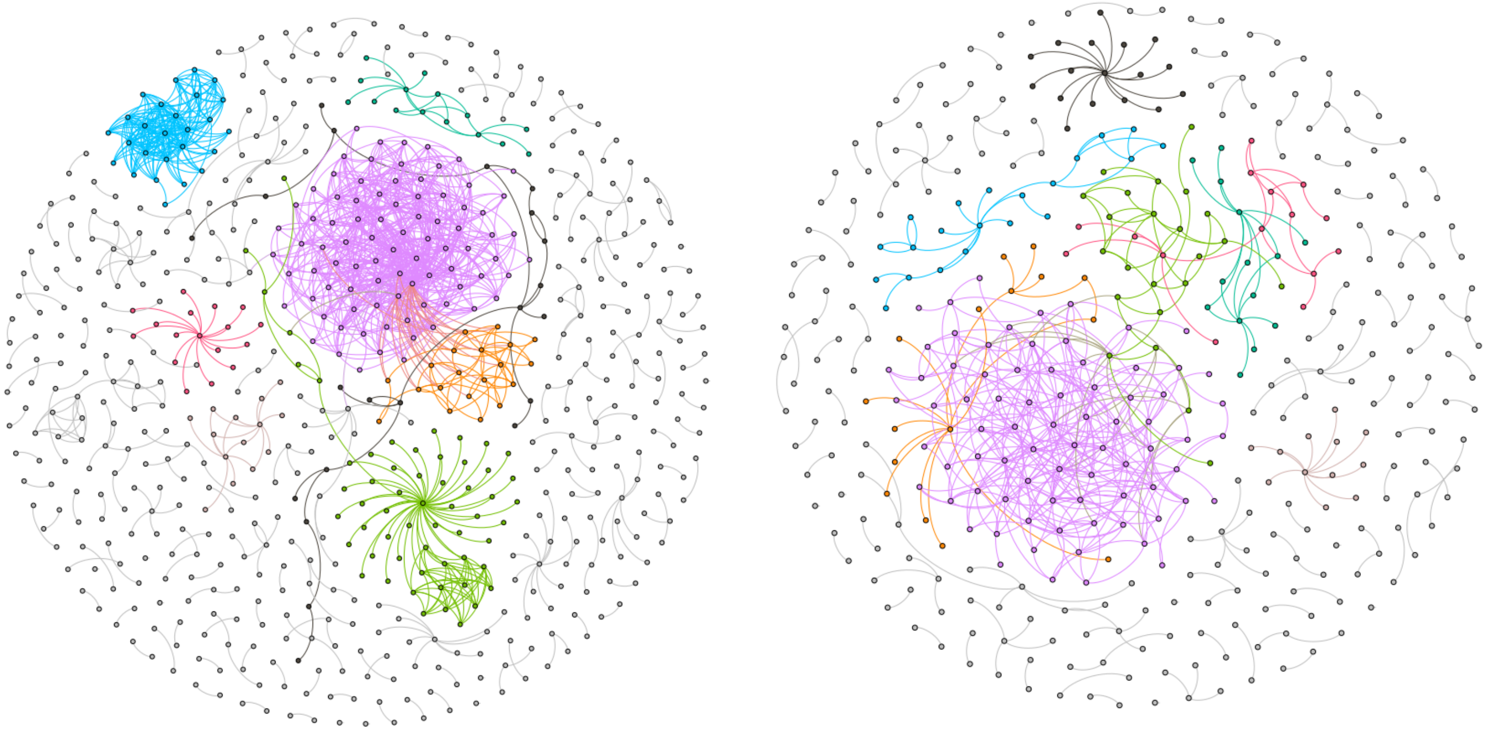
Structure Analysis - Links



Highly connected but sparse (>60,000 connections)

10% were complete isolated and not reachable → 90% are

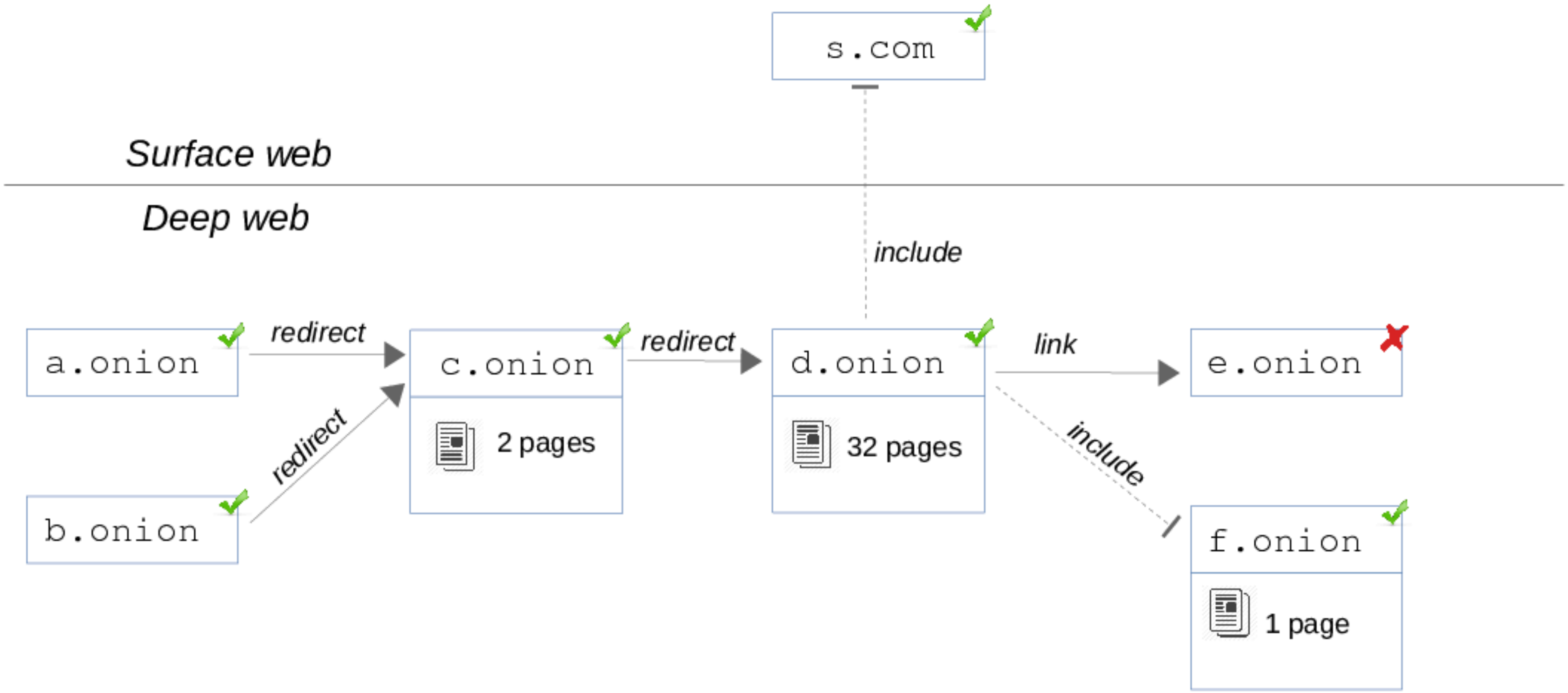
Structure Analysis – Resources and Redirections



82.83% and 84.88% of the nodes are strongly connected

Also highly connected but smaller networks of connections than links

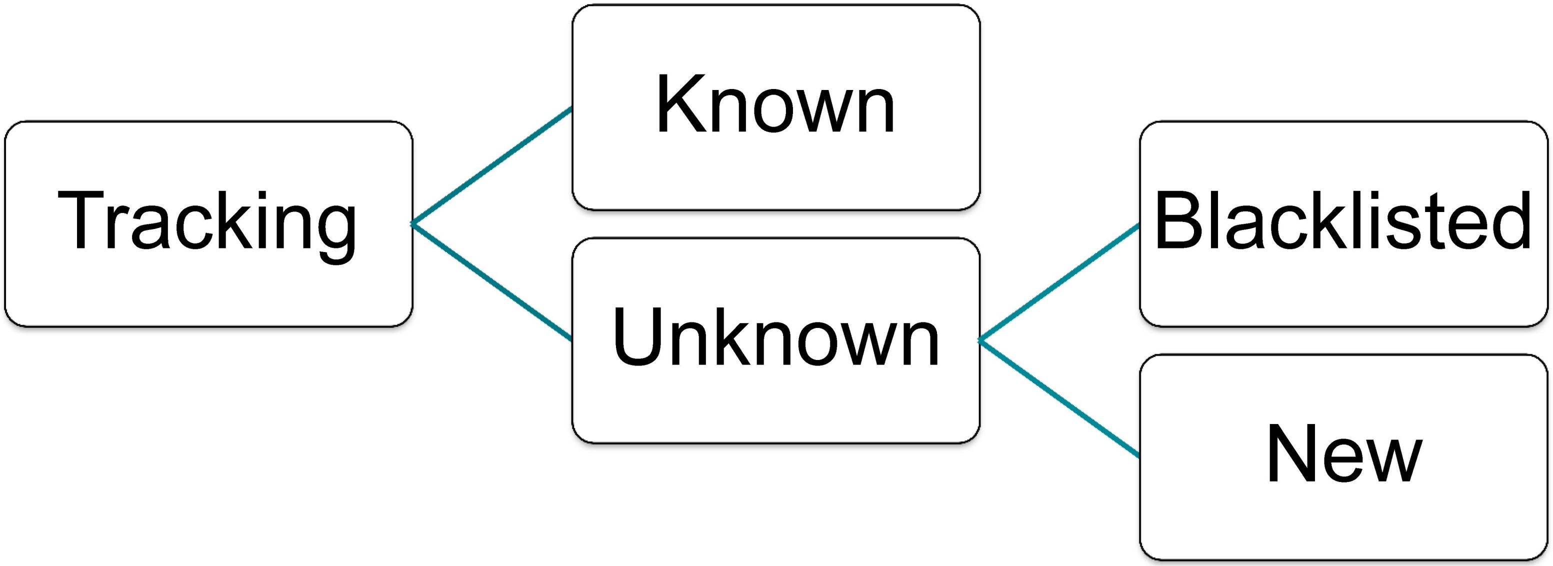
Privacy Analysis - Dark-to-Surface Leakage



21% of the sites import resources from the surface

Google alone can monitor the 13% of the Tor hidden services

Privacy Analysis - Web Tracking



Tracking

Known

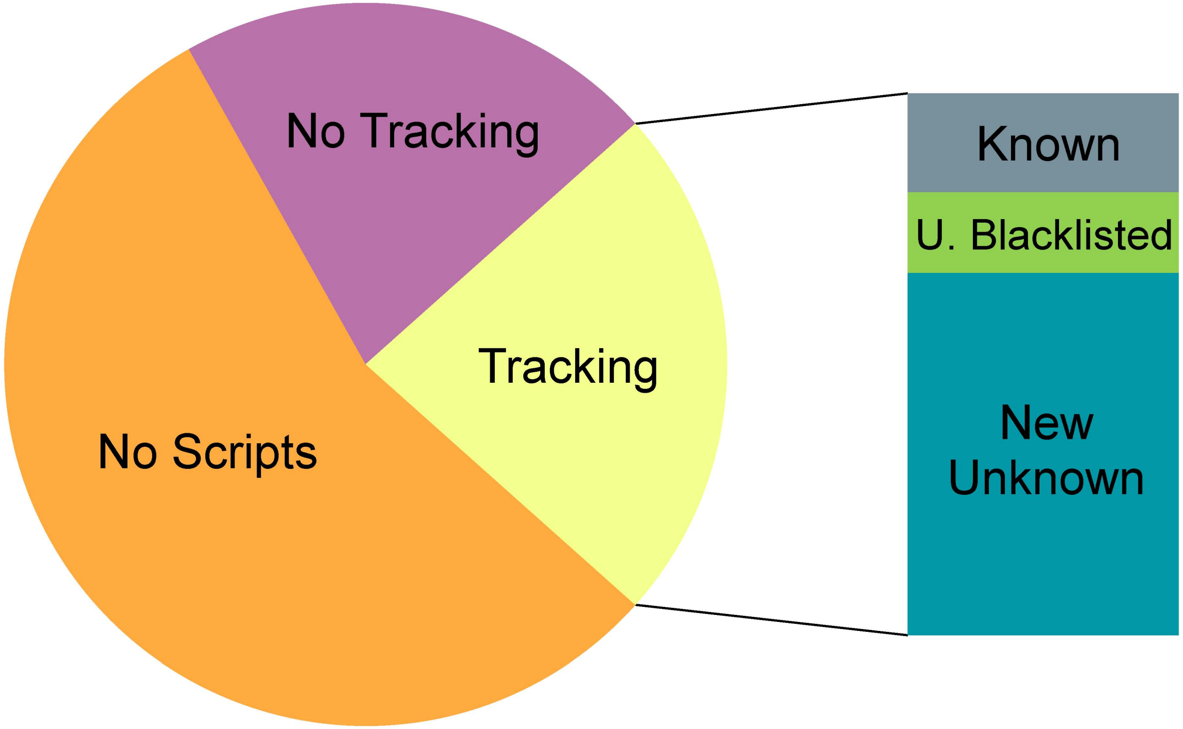
Unknown

Blacklisted

New

TrackingInspector is used to analyze scripts

Privacy Analysis - Web Tracking - Prevalence



Privacy Analysis - Web Tracking - Specifics

Type	% Tracking Scripts
Statistics	17.10%
Stateless Tracking	15.04%
Advertisement	10.48%
Web Analytics	10.08%
Stateful Tracking	7.22%

10% of the tracking scripts were unique

32.50% of the tracking came from surface web

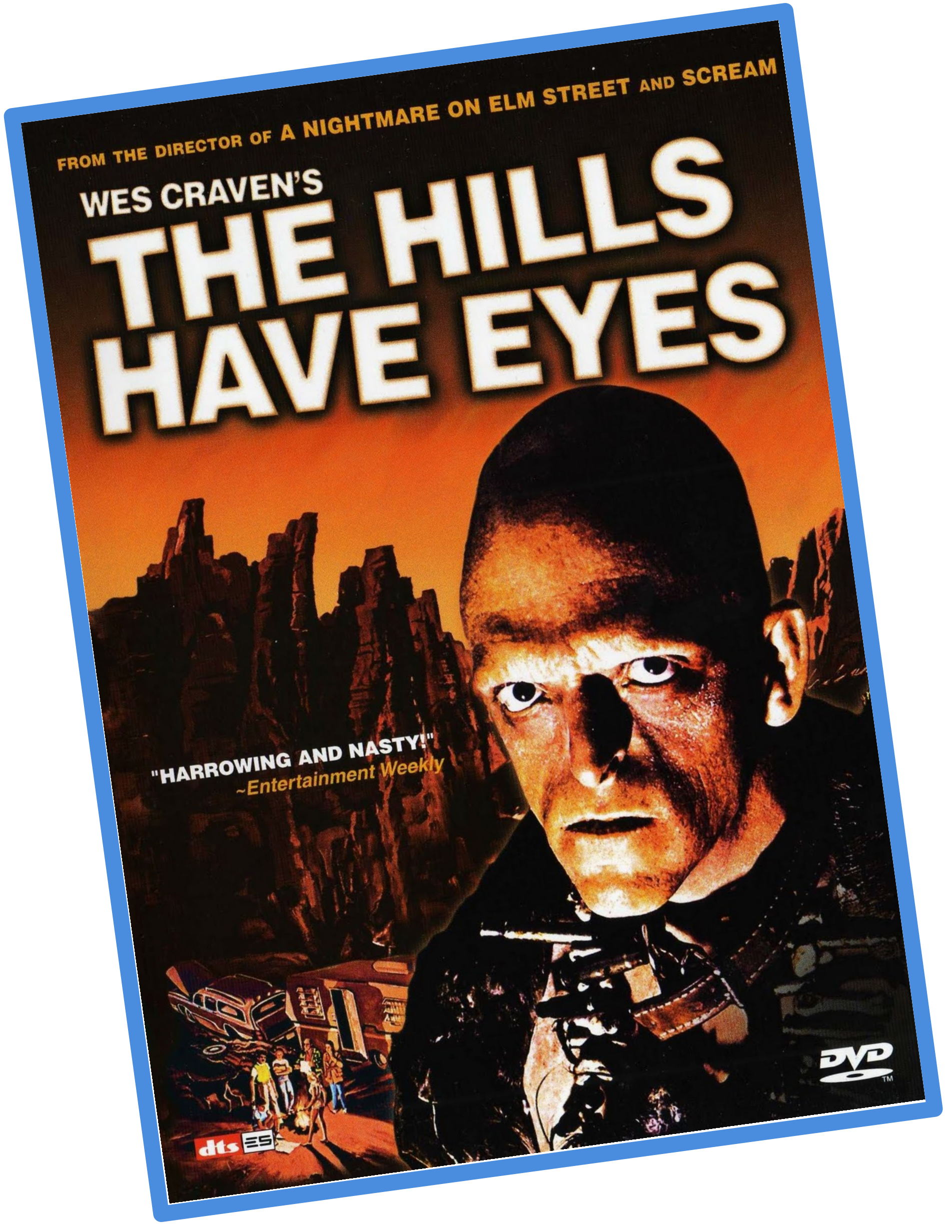
Privacy Analysis - Tracking Hiding techniques

- **Obfuscated** tracking exists in the dark web: 0.61% of the scripts did
- **Script embedding** is highly used (16.28%) and with a large number of techniques, e.g.:

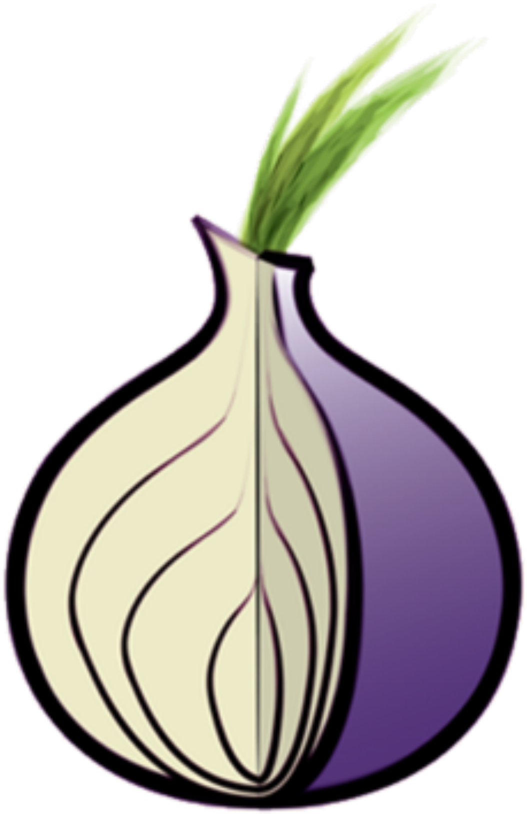
  dota.js → canvas fingerprinting

  analytics.js → the usual Google tracking
- New technique: **intermediate tracking** in redirections: 1.67%

We already knew that the hills have eyes...

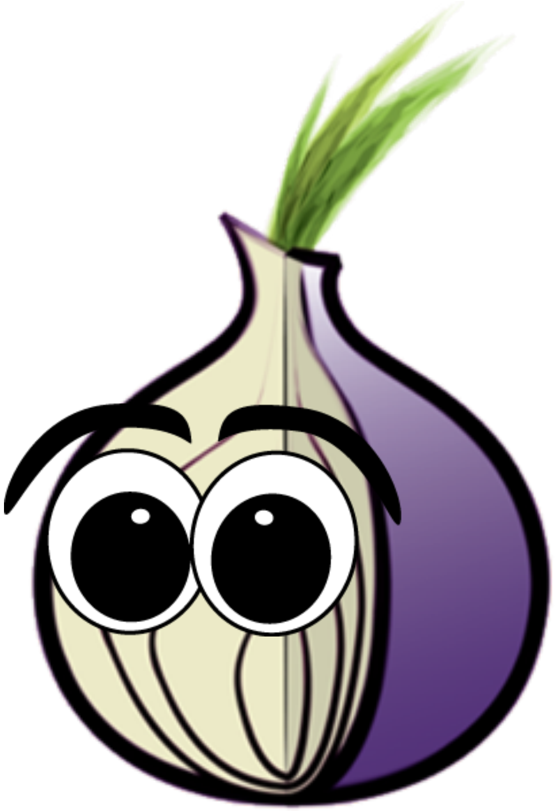


but we didn’t expect onions to have them too…



but they do...

The Onions Have Eyes



iskander.sanchez@deusto.es
iskander-sanchez-rola.github.io