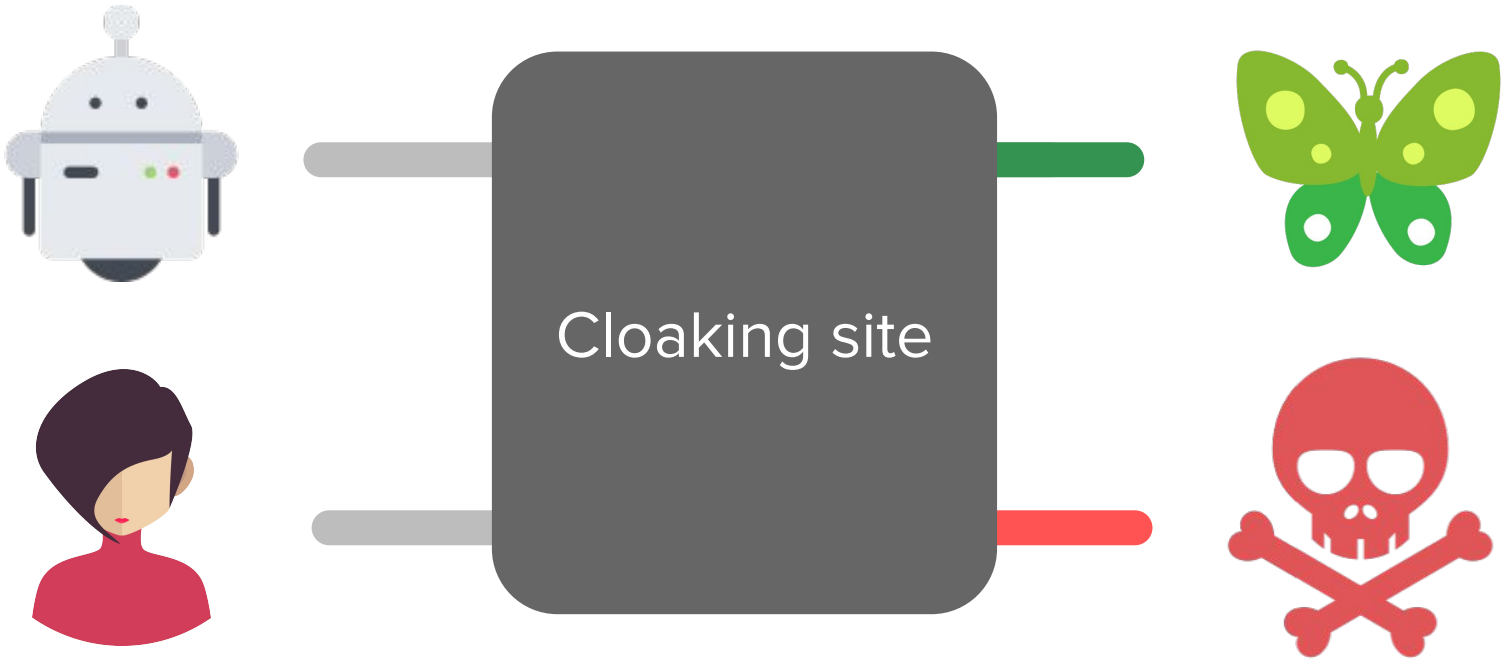
Cloak of Visibility: Detecting When Machines Browse a Different Web

Luca Invernizzi*, Kurt Thomas*, Alexandros Kapravelos[†], Oxana Comanescu*, Jean-Michel Picod*, and Elie Bursztein*

* Google - Anti-fraud and abuse research

[†] North Carolina State University

Web cloaking



Web cloaking

Search

Effective for Search Engine Optimization

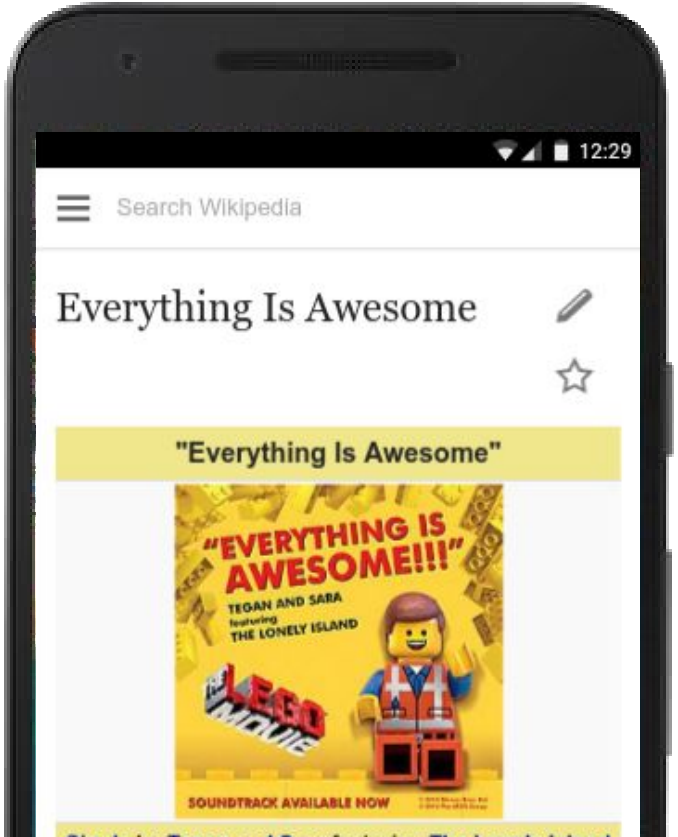
Ads

Effective to infringe policies

~~Malware~~

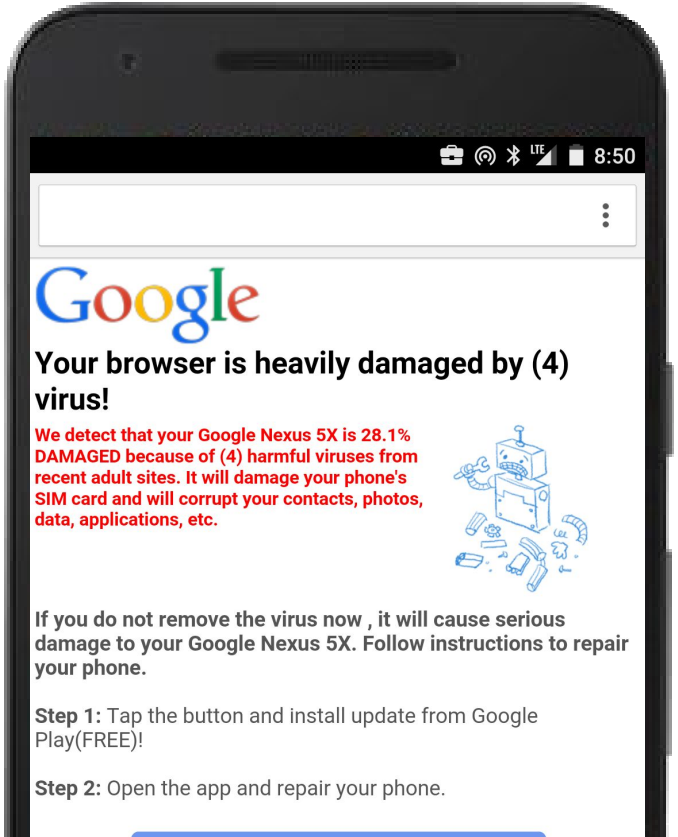
~~Effective to evade security crawlers~~

Responsive design vs cloaking



This is **not cloaking**.

Responsive design vs cloaking



This is **cloaking**.

Research goals

Keep up with arms race

Identify trends

Explore alternatives

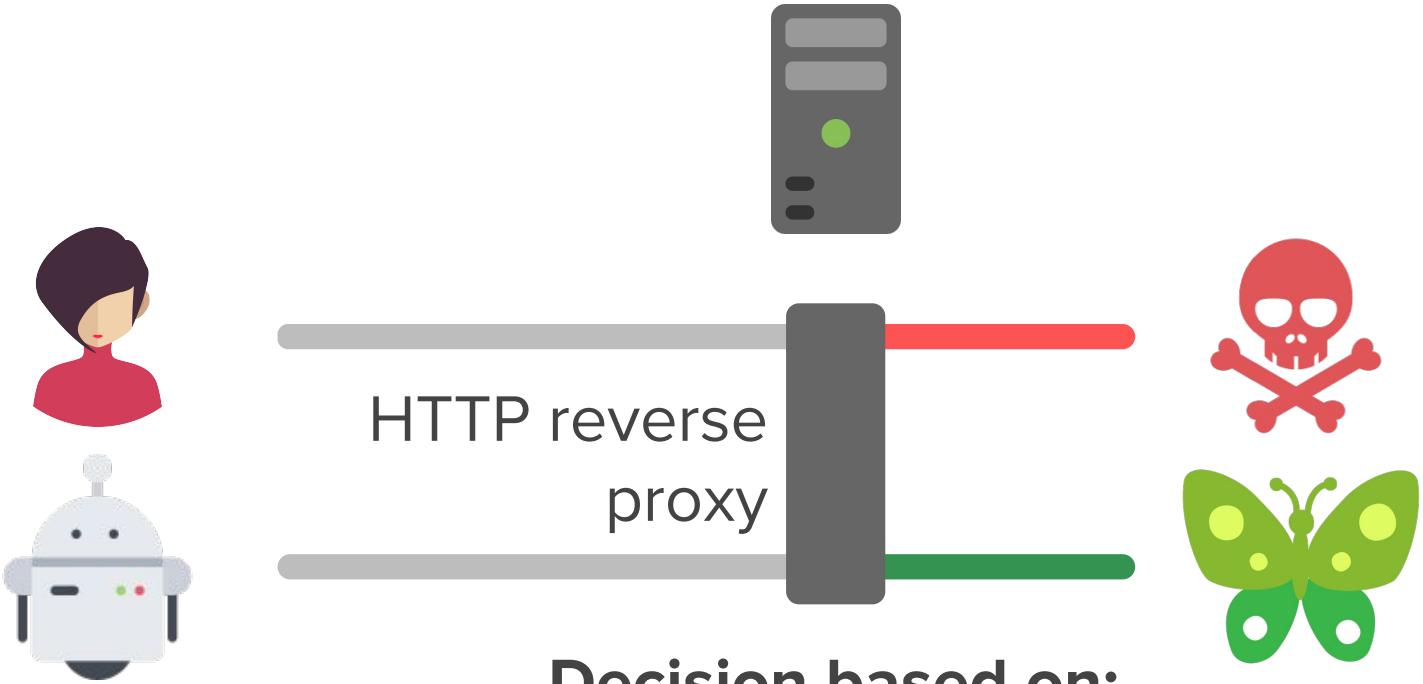
Blackmarket Investigation

Acquired

Top 10

Cloaking software samples

$3500+ cloaking software



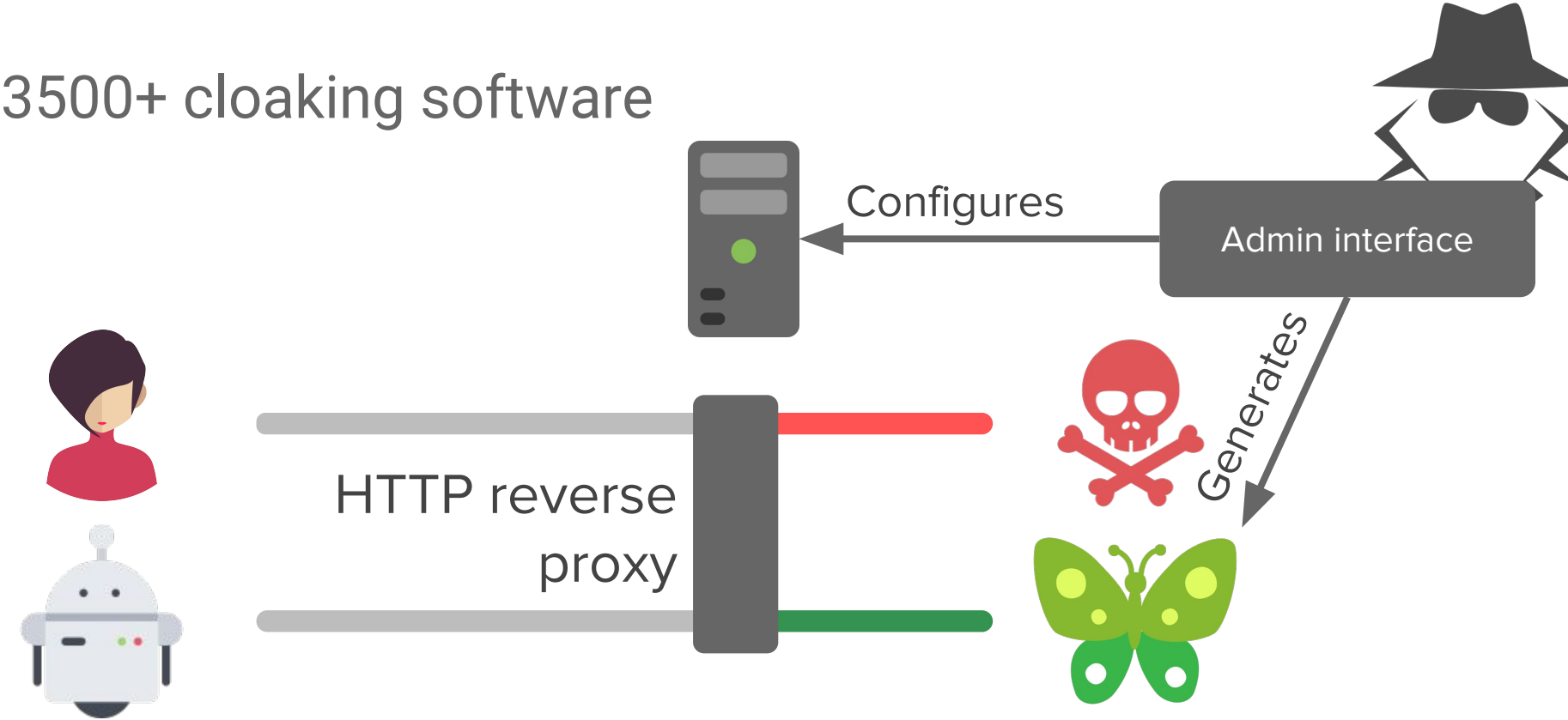
Network

Decision based on:

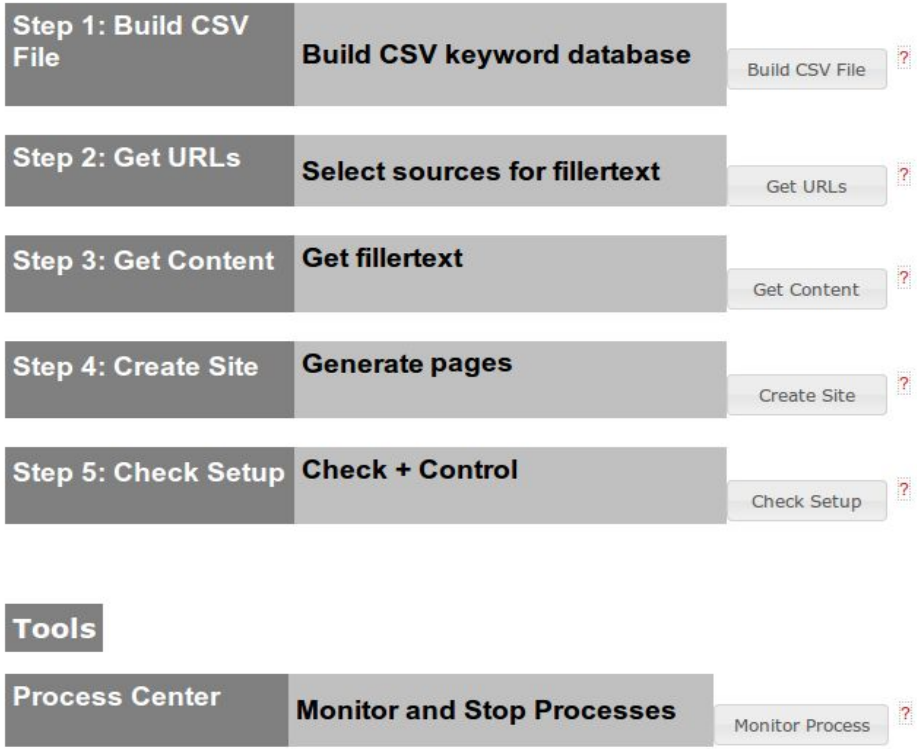
Browser

Browsing context

$3500+ cloaking software



Admin interface



Input

keywords => http://money.site

Features

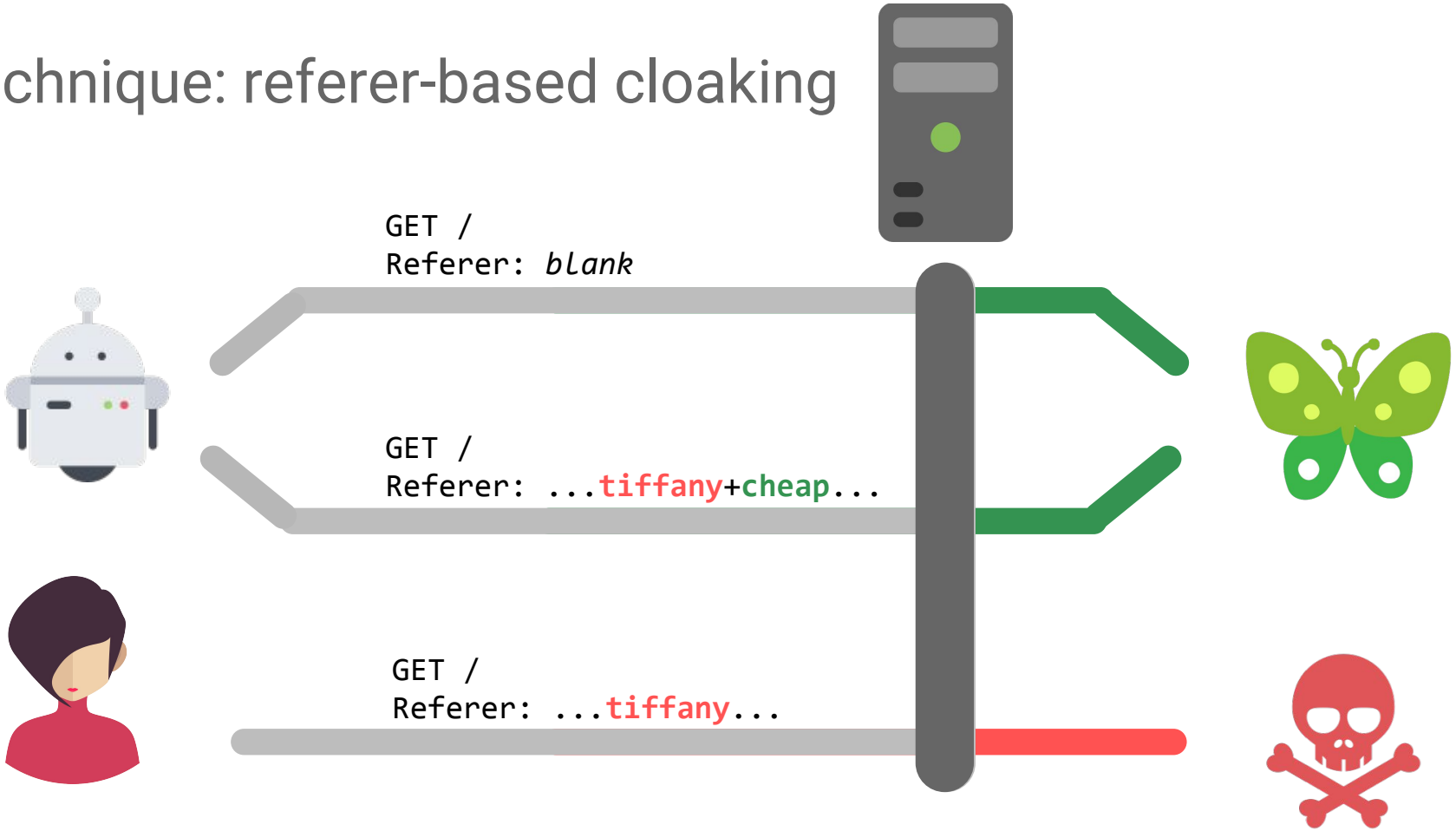
- Find similar sites through SERPs
- Content/Template spinning
- Drip-feeding

Added services

- Plagiarism detection
- SERP ranking

Cloaking techniques

Technique: referer-based cloaking



Technique: IP blacklisting

51m
Blacklisted IPs

30
Security companies

3
Proxy networks

983
Subnets

2
Hacking collectives

122
Entities: companies, universities, registrars

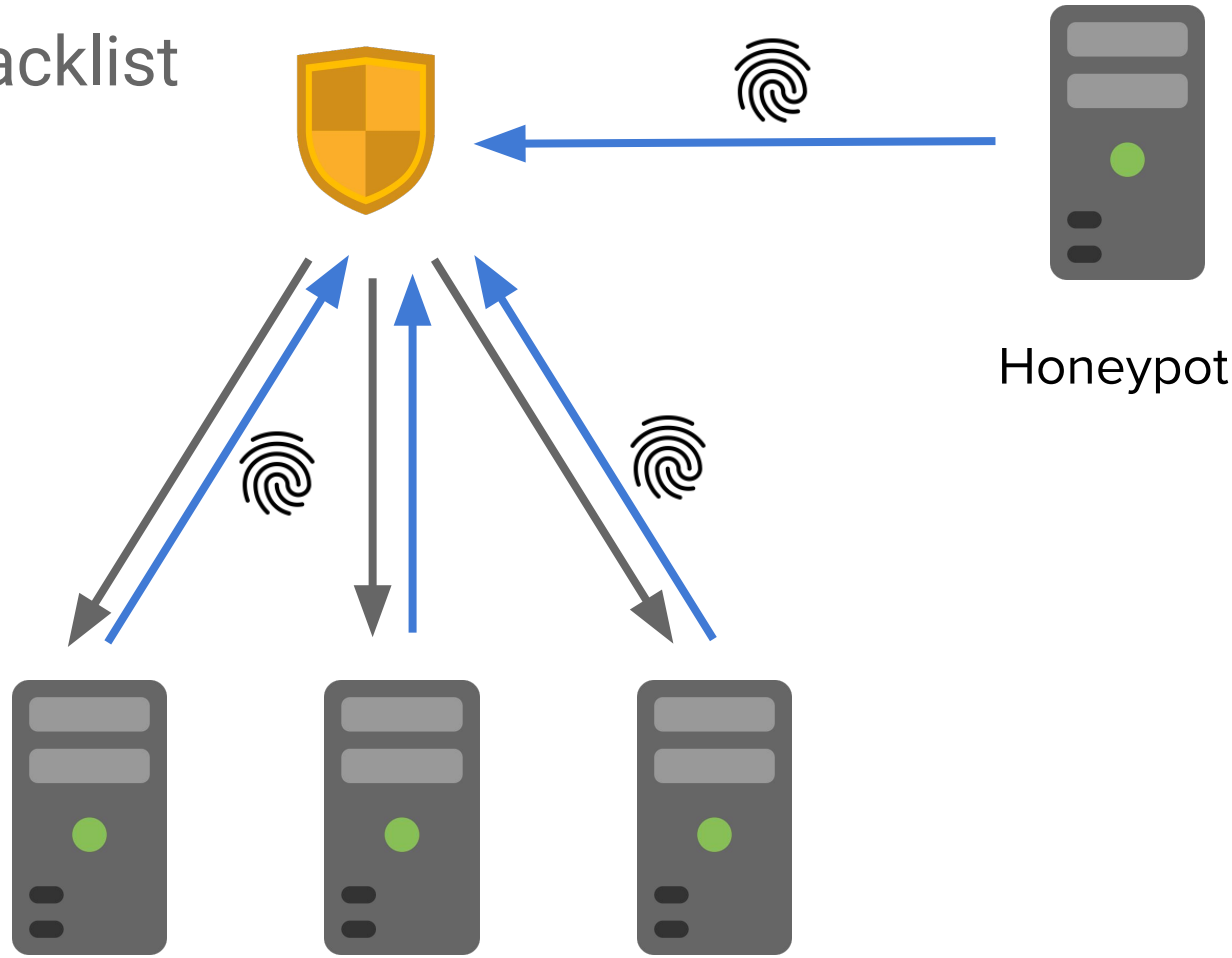
Crowdsourced blacklist

50k

Blacklisted IPs

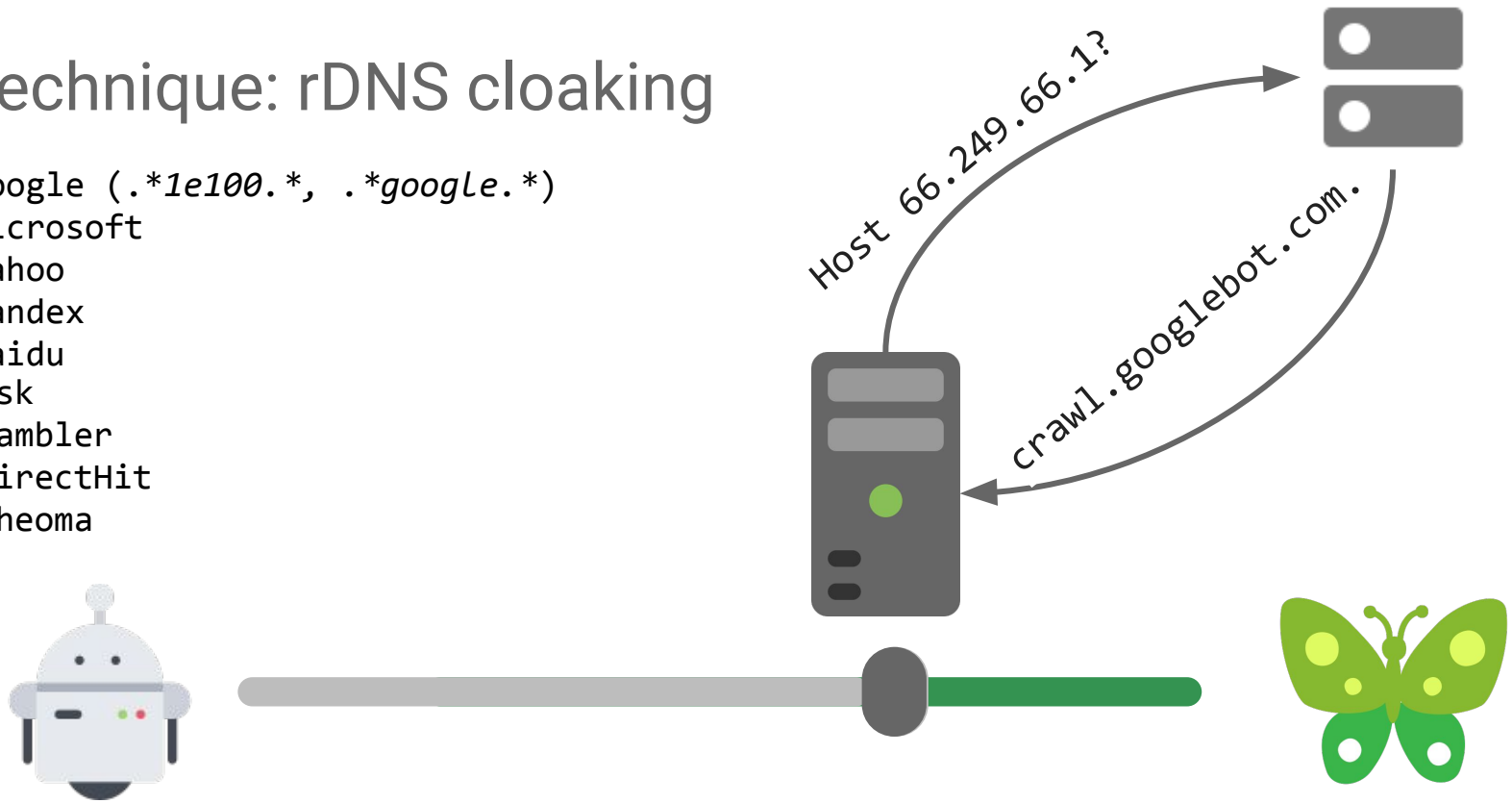
$350+

Subscription



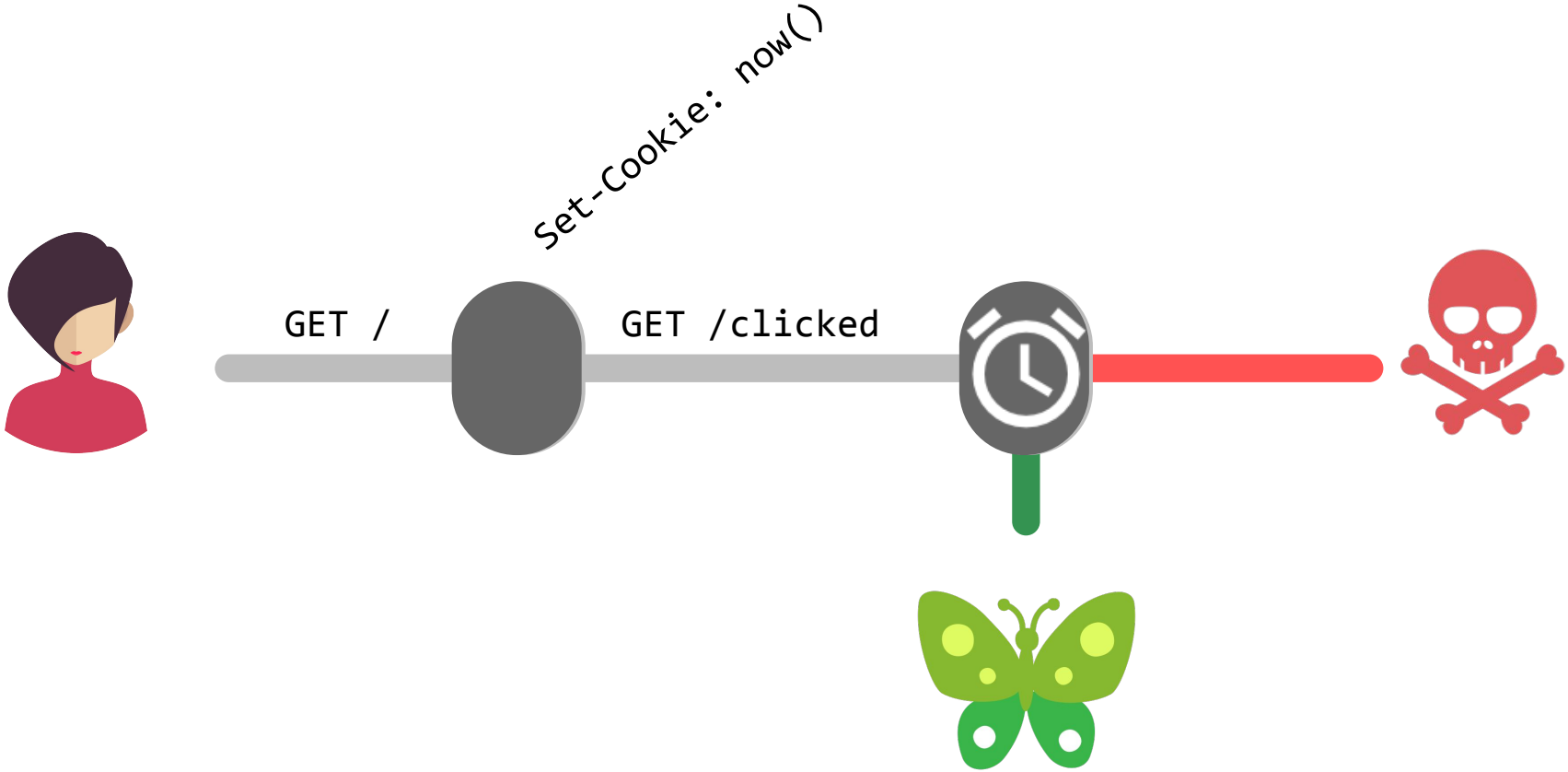
Technique: rDNS cloaking

Google (*.*1e100.*, .*google.*)
Microsoft
Yahoo
Yandex
Baidu
Ask
Rambler
DirectHit
Theoma



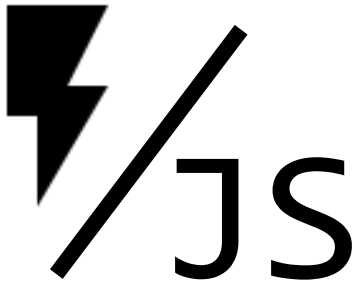
66.249.66.1

Technique: browsing pattern cloaking



More techniques

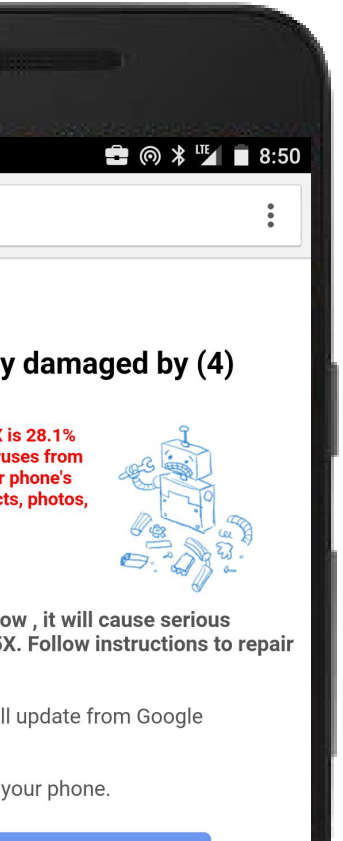
Geolocation: country, city, carrier level.



Flash/JS support & fingerprints

User-Agent

Prevalence and dominant techniques



Is this cloaking?
How do they cloak?

Browser farm

Pretend Google bots

User-Agent: GoogleBot
Referer: blank
Google IP

Simple honey clients

User-Agent: Chrome
Referer: blank, or simple
Cloud provider IPs

Realistic honey clients

User-Agent: Chrome
Referer: context-aware
Residential and mobile IPs

Features

	 **HTML**	 **Image**
Syntactic	Content similarity	Screenshot similarity
Semantic	Topic similarity	Screenshot topic similarity

Classification

82%
True positive rate

.9%
False positive rate

95k labeled samples
75k legitimate websites (Alexa) + 20k cloaked storefronts

Prevalence

4.9%

Cloaking pages in **Google AdWords**, for health and software ads.

11.7%

Cloaking pages in **Google Search**, for luxury storefronts keywords.

Traditional techniques: only IP, Referer, and User-Agent

Search: 1 out of 5
Ads: 1 out of 4

Current techniques: JavaScript support

Search: Half
Ads: 1 out of 4

Current techniques: wait for click

Search: 1 out of 10

Ads: 1 out of 5

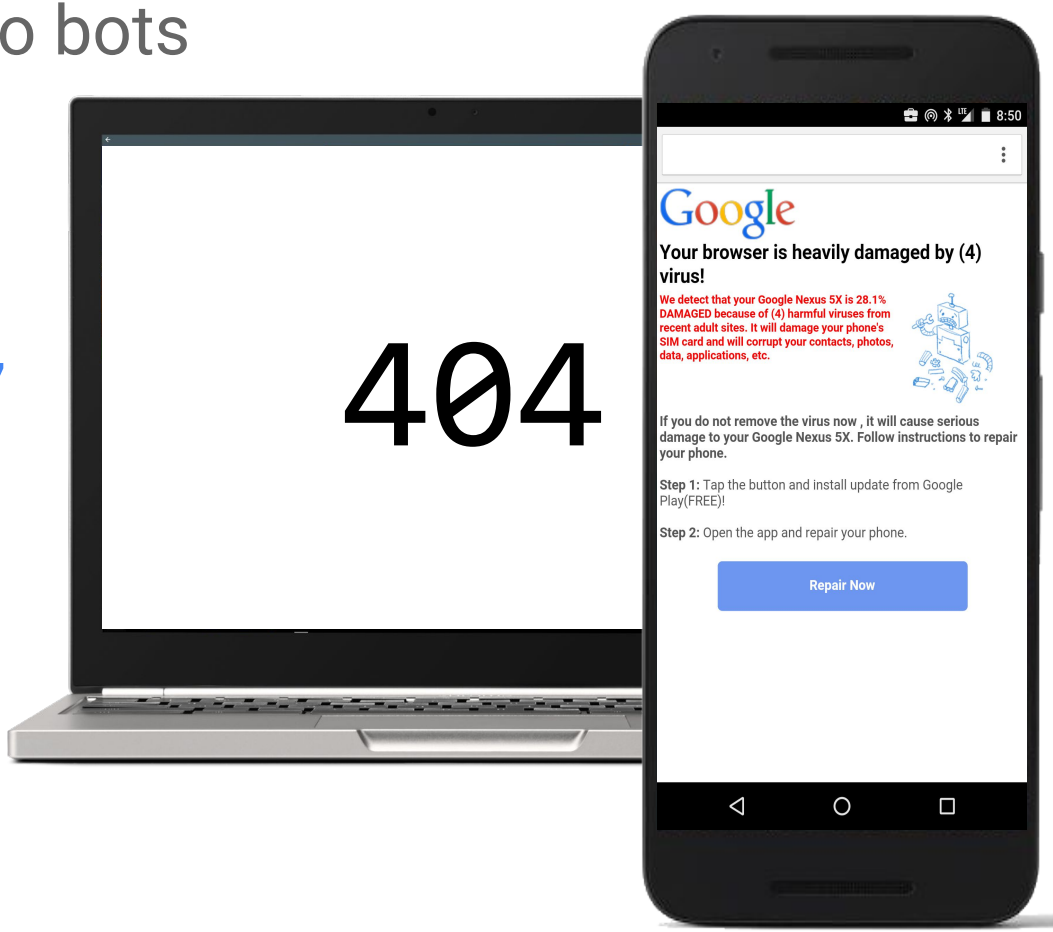
Delivery: same-page cloaking

Search: 1 out of 5
Ads: **2 out of 3**

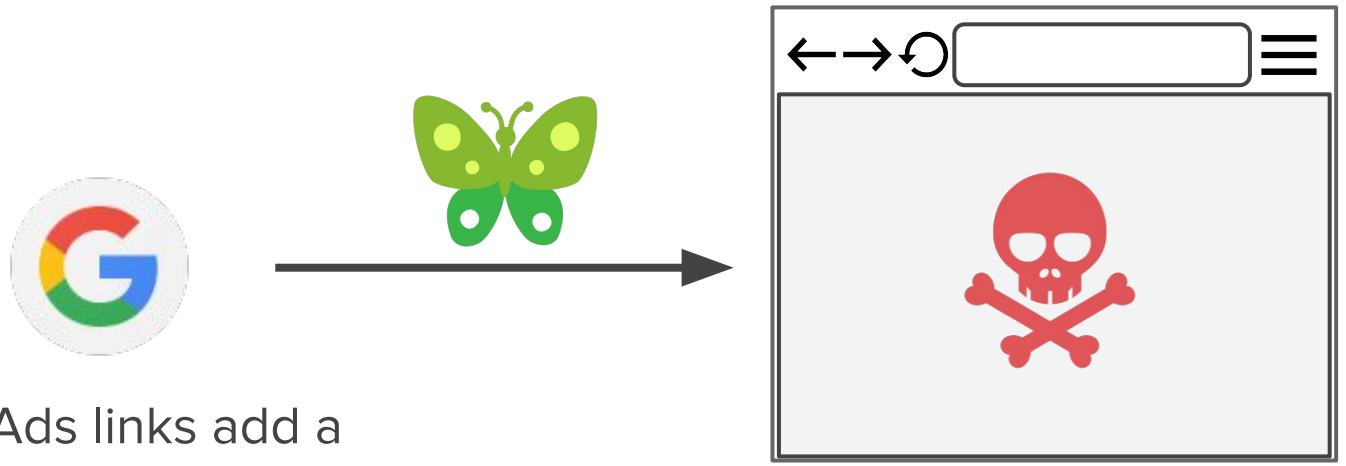
Uncloaked Cloaked

Delivery: 40x/50x errors to bots

Search: 1 out of 7
Ads: 1 out of 8



Future: client-side detection



Search/Ads links add a parameter with the **topics** found by the bot.

Check that the page matches the **same topics.**

Takeaways

Prevalence

5% of ads and 12% of search results for cloaking-prone keywords cloak.

Techniques

IP/`User-Agent`/ `Referer` only gets ⅙ of cloaking.

Moving forward

Client side, semantic features needed for hard cases.

Thank you!

Luca Invernizzi
invernizzi@google.com