

Methodological issues in:

“Empirically Characterizing Domain Abuse and the Revenue Impact of Blacklisting”

Stefan Savage
University of California, San Diego

Empirically Characterizing Domain Abuse and the Revenue Impact of Blacklisting

This paper has two different kinds of results

- **Characterization of various counterfeit pharma sites/domains**

How are they getting traffic, how much revenue per domain, etc?

- **How much did blocklisting impact that revenue?**

Can we use it to reason about how well blocklisting can help?

What are the kinds of data needed for this analysis?

- How did a given site attract traffic?
- How much money did a given site make at time t ?
- When was a given site blocklisted?

Found data: pharmaleaks

Sometimes unique data just *shows up*

- Usually, a byproduct of building networks *outside academia*

Leaked ground truth data

PharmaLeaks: Understanding the Business of Online Prescription Drugs

Damon McCoy, Andrew Pitts, Brian Krebs, Geoff...

Department of Computer Science, George Mason University, International Center for Cyber Security

Abstract

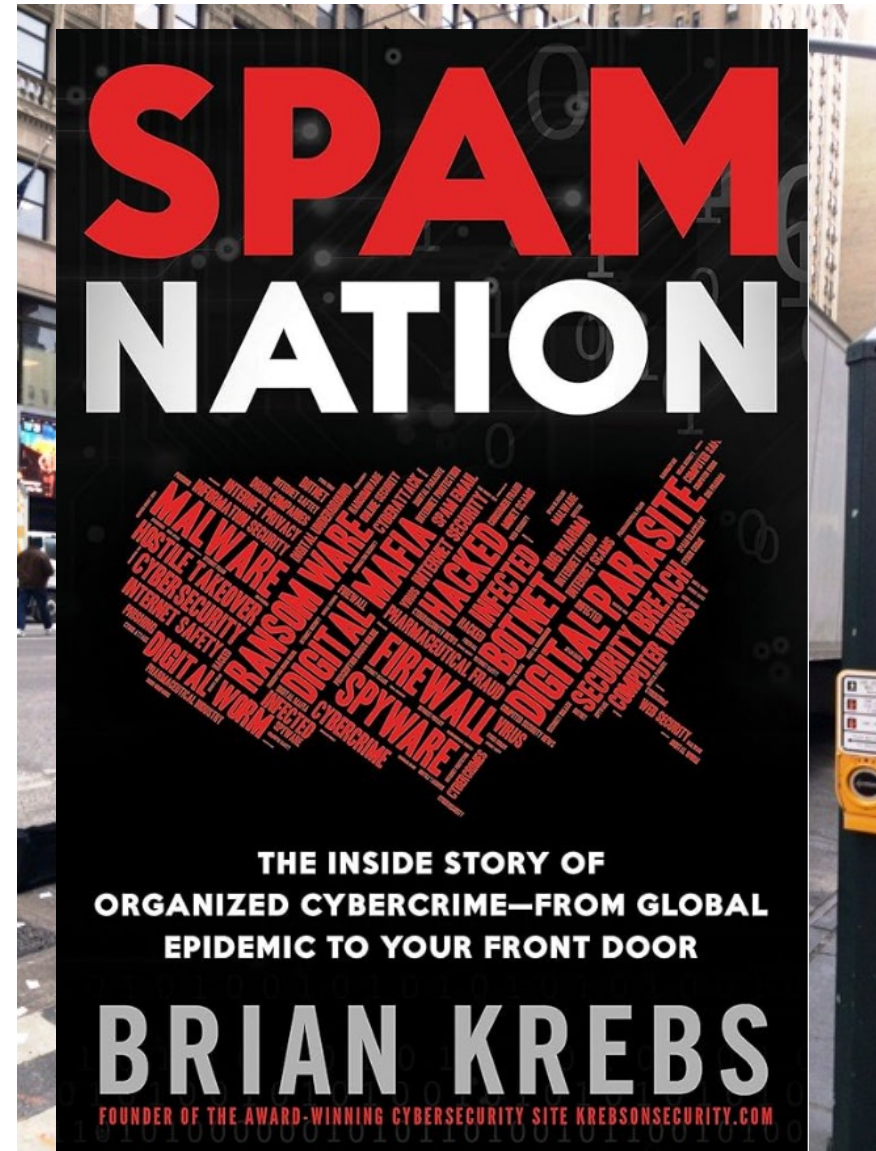
Online sales of controlled substances have become a robust underground advertising channel. These sales, often referred to as "dark web" sales, are typically conducted through encrypted channels and are often used to circumvent traditional regulatory barriers. This paper introduces a new artifact, we develop no new inference technique, we deploy no new measurement infrastructure. We do none of these things because we don't need to; we have the actual data sets that we would otherwise try to measure, infer or estimate. Thus, while there are significant challenges to the use of this data, we believe that it provides a unique and valuable perspective on the overall business model.

1 Introduction

Much like the legitimate Internet economy, advertising is a major driver for the "underground" online economy as well. For all their negative connotations, online ads, forum spam and social spam—as well as the bots, botnets, and other malicious infrastructure that enable these activities—are all simply low-cost, high-impact advertising channels. Consequently, an emerging research agenda has developed around understanding the economic structure of these businesses, both to understand the scope and drivers for the problem (8, 13, 15) and to help policy-makers and law enforcement understand the key questions, much remain unclear. This is because, as a rule, there is little "ground truth" data in the field for either validating such results or to provide focused analysis that can be obtained via inference.

This paper provides a rare counterpoint to this rule. Under a variety of circumstances (legally obtained via multiple sources – relationship with Brian Krebs (journalist), commercial investigations community, contacts adjacent to Russian underground)

ev's Glavmed/SpamIT pharmacy
Dmitri Stupin (Gusev's site – all chat logs
Publvisky's RXPromotions (pharma) and Chronopay (payment processor)



Pharmaleaks paper

Glavmed/Spamit, Rx-Promotion

- 185M in gross revenue, 1+ million customers, 1.5+ million purchases, 2600+ affiliates
- **For Glavmed/Spamit: Web logs for every transaction**

4

Program	Period	Affiliates	Customers	Billed orders	Revenue
GlavMed	Jan 2007 – Apr 2010	1,759	584,199	699,516	\$81M
SpamIt	Jun 2007 – Apr 2010	484	535,365	704,169	\$92M
RX-Promotion	Oct 2009 – Dec 2010	415	59,769 – 69,446	71,294	\$12M



Aside: what are the issues with leaked/found data?

How do you know its authentic?

- Internal consistency
- External consistency

Can you legally use illegally obtained data in research?

- Did you participate in the obtaining of data? Did you solicit its acquisition?
- What kind of data and how did you use it? (e.g., sales, redistribution, etc)?

Ethics of use

- No new harms argument (data availability)
- Harm minimization
 - Anonymization of customer names/credit cards
 - No true names of employees/affiliates
- But, there can be differences of opinion on this (remember Trolley Problem paper)
Possibility for revictimization, etc.

How to tell where traffic came from?

“Referer” field in HTTP GET requests to shop domain

method path version
GET /index.html HTTP/1.1

headers

Accept: image/gif, image/x-bitmap, image/jpeg, */*
Accept-Language: en
Connection: Keep-Alive
User-Agent: Mozilla/1.22 (compatible; MSIE 2.0; Windows 95)
Host: www.example.com
Referer: http://www.google.com?q=cheapgucci

body
(empty)

How to tell where traffic came from?

“Referer” field in HTTP GET requests to shop domain

- Available in 30% of spamit, 61% of glavmed (why not all?)
- What are these categories?

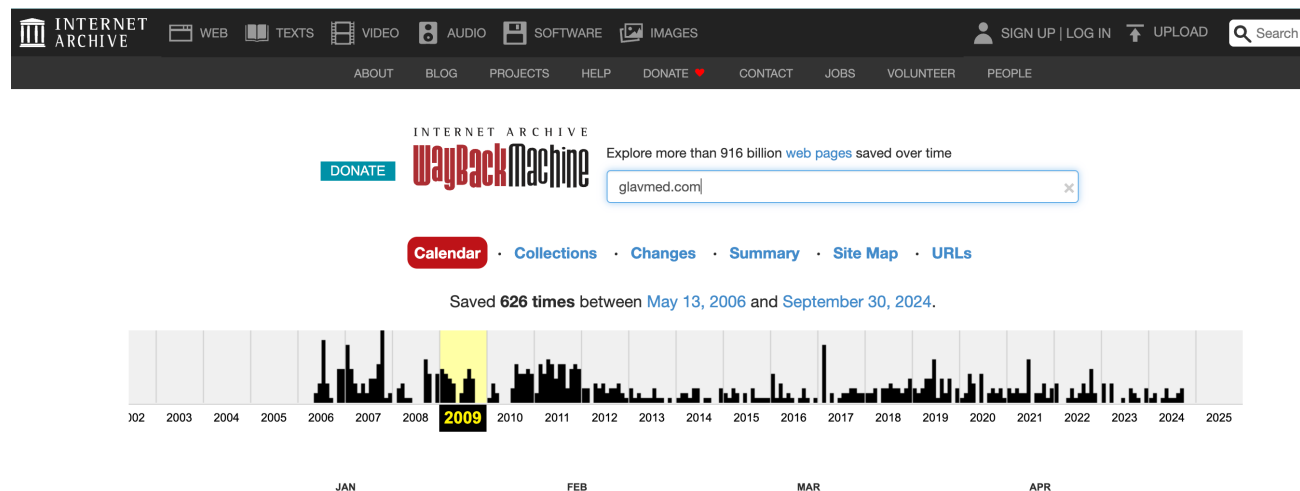
Category	Referrer
Email spam	http://mail.live.com/mail/readmessagelight.aspx?action=markasnotjunk&folderid=...
Web search	http://search.yahoo.com/search?p=canadian+viagra&ei=utf-8&fr=blie7
Free hosting	http://groups.google.com/group/... http://www.umbc.edu/ddm/wiki/user:cheap_cialis http://answers.yahoo.com/my/profile?show=...
Bulk purchased	http://accutanewithoutprescription.org
Compromised	http://library.newschool.edu/askal/request/.inc/c/clomid-without-prescription.html
Traffic	http://traffic-analytics.net/tds/in.cgi?3&seoref=http://search.comcast.net/?...q=lavitra&http_referer=http://www.plantright.org/?id=49&default_keyword=http://klikcentral.com/traffic/in.cgi?11&parameter=buy%20viagra&seoref=http://www.google.com/search?q=buy+viagra&...&http_referer=www.vfcc.edu

Table 3: Example referrers for advertising vectors (email and Web search), infrastructure domains (free hosting, bulk, and compromised), and purchased traffic.

Pros/cons to using referrer for categorization?

Validating historic URLs

How do you check what “kind” of site a URL used to point to *in the past*?





HOME HOW IT WORKS SITES SIGN UP MEMBERS FAQ CONTACT US

ENG
RUS

MEMBERS LOGIN

OK

SSL login form

Forgot your password?
Sign up now

SUPPORT

E-Mail: support@glavmed.com

ICQ: 383835814

397061228

WELCOME TO GLAVMED



GlavMed is a BEST way to convert your pharmacy traffic into real money. Forget about miserable sums you're getting sending your visitors to PPC pharmacy results.

You're loosing at least half of YOUR money converting traffic like this. GlavMed offers you a possibility to eliminate any agents and sell most popular pharmacy products directly. It means 30-40% revenue share.

FEATURES & BENEFITS



HIGHEST INDUSTRY
COMMISSIONS



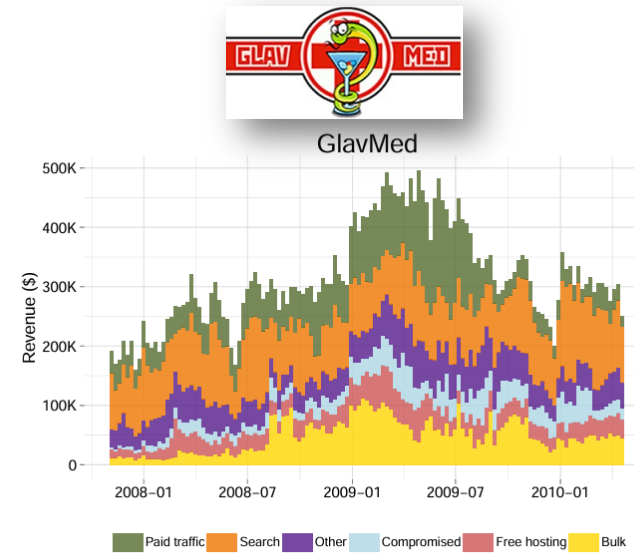
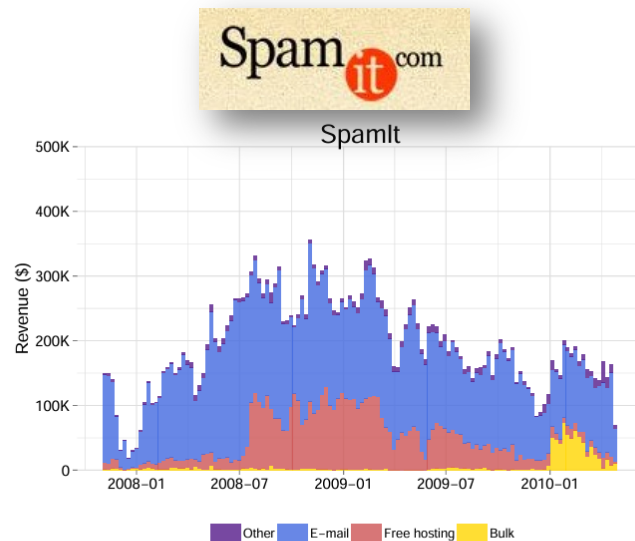
MOST POPULAR
PHARMACY PRODUCTS



BIWEEKLY PAYMENTS AND
PAYOUT-ON-DEMAND



Characterizations of traffic source vs revenue



How were these graphs made? What was the methodology?

Why such a big difference between the traffic makeup?

- Different marketing to affiliates: spamit for spam, glavmed for !spam
- What is up with the "free hosting" and "bulk" traffic in spamit?

Why didn't they look at source of people who visit but don't purchase?

More characterization

	Shop Sites	Sales	Revenue	Revenue/Sale	Affiliates
Advertisement vectors	11957	147582	\$18.05M (73.2%)	\$122.36	330
<i>Email spam</i>	11898	145041	\$17.83M (98.7%)	\$122.94	326
<i>Web search</i>	173	2541	\$0.23M (1.25%)	\$89.21	71
Infrastructure domains	1402	54351	\$6.58M (26.7%)	\$121.17	174
<i>Free hosting</i>	1282	45941	\$5.67M (86.1%)	\$123.37	154
<i>Bulk purchased domains</i>	120	7781	\$0.84M (12.8%)	\$108.36	50
<i>Compromised sites</i>	64	629	\$0.07M (1.13%)	\$119.33	27
Purchased traffic	11	199	\$0.02M (0.08%)	\$104.15	4
Uncategorized	863	4610	\$0.54M (2.20%)	\$117.91	165

Table 1: Classification of referrers used by SpamIt affiliates.

	Shop Sites	Sales	Revenue	Revenue/Sale	Affiliates
Advertisement vectors	1433	134977	\$13.8M (38.3%)	\$102.25	787
<i>Email spam</i>	615	10855	\$1.35M (9.81%)	\$124.79	578
<i>Web search</i>	1182	124122	\$12.4M (90.2%)	\$100.27	537
Infrastructure domains	1017	134832	\$13.71M (38.0%)	\$101.68	898
<i>Free hosting</i>	684	38094	\$3.91M (28.5%)	\$102.65	654
<i>Bulk purchased domains</i>	374	63639	\$6.27M (45.7%)	\$98.55	356
<i>Compromised sites</i>	456	33099	\$3.53M (25.7%)	\$106.59	393
Purchased traffic	458	86657	\$8.55M (23.7%)	\$98.68	366
Uncategorized	1047	45337	\$4.72M (13.1%)	\$104.21	890

Table 2: Classification of referrers used by GlavMed affiliates.

Aside: narrative charts – anecdotes as graphs

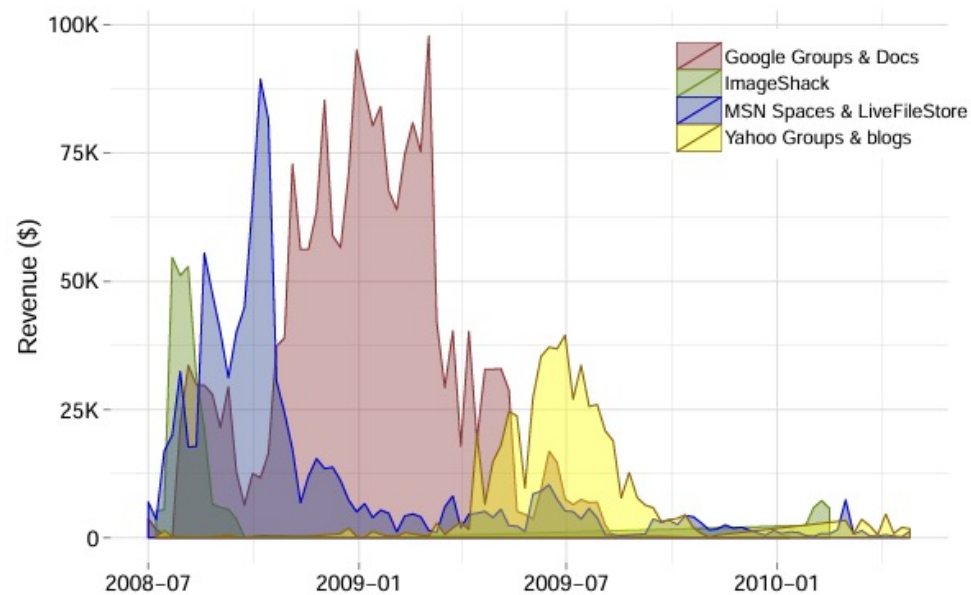


Figure 2: Spammers seamlessly switch from one free hosting site to another in the face of takedowns.

Domain blocklisting

One of oldest anti-spam interventions around (20th century!)

- Basic idea: Identify known bad domains and share list so e-mail containing them can be filtered; core part of anti-spam to this day
- What are choices for an e-mail with a blocklisted domain?
(i.e., what might Gmail do with such a message?)

Lots of analyses existed about how well it protects

- How much spam filtered?
- How quickly do domains end up on blocklists?
- Etc

This paper focuses on a different question: **does it disrupt?**

- How much does blocklisting undermine revenue for spammers?

Blocklisting part of the analysis

Only focuses only on SpamIt domains.
Why?

What other data do they need?

- When data appears in blocklist?

URIBL

Blocklist has predictive features, how do they address that issue?

- When was data advertised in spam?

Why do they need this?

Two contemporaneous spam feeds
(same source as Click Trajectories)

- MX honeypot
- Human identified

	Blacklisted	Non-blacklisted
Shop-sites	34959	4751
Sales	56K	80K
Revenue	\$5.9M	\$9.7M
Sales/Site	1.6	16.9
Affiliates	119	144
Sites seen in feeds	34771 (99%)	1193 (25%)
<i>MX</i>	30647	27
<i>HI</i>	33701	1185
Sales with referrers	6076	28576
<i>Email Spam</i>	4798 (78.9%)	18206 (63.7%)
<i>Purchased Traffic</i>	-	168 (0.59%)
<i>Free hosting</i>	284 (6.31%)	4507 (15.8%)
<i>Compromised sites</i>	8 (0.13%)	124 (0.44%)
<i>Web search</i>	7 (0.11%)	22 (0.07%)
<i>Bulk Purchased Domains</i>	709 (11.7%)	4375 (15.3%)
<i>Uncategorized</i>	170 (2.79%)	1172 (4.10%)

Table 4: Statistics differences between blacklisted and non-blacklisted domains.

Some takeaways

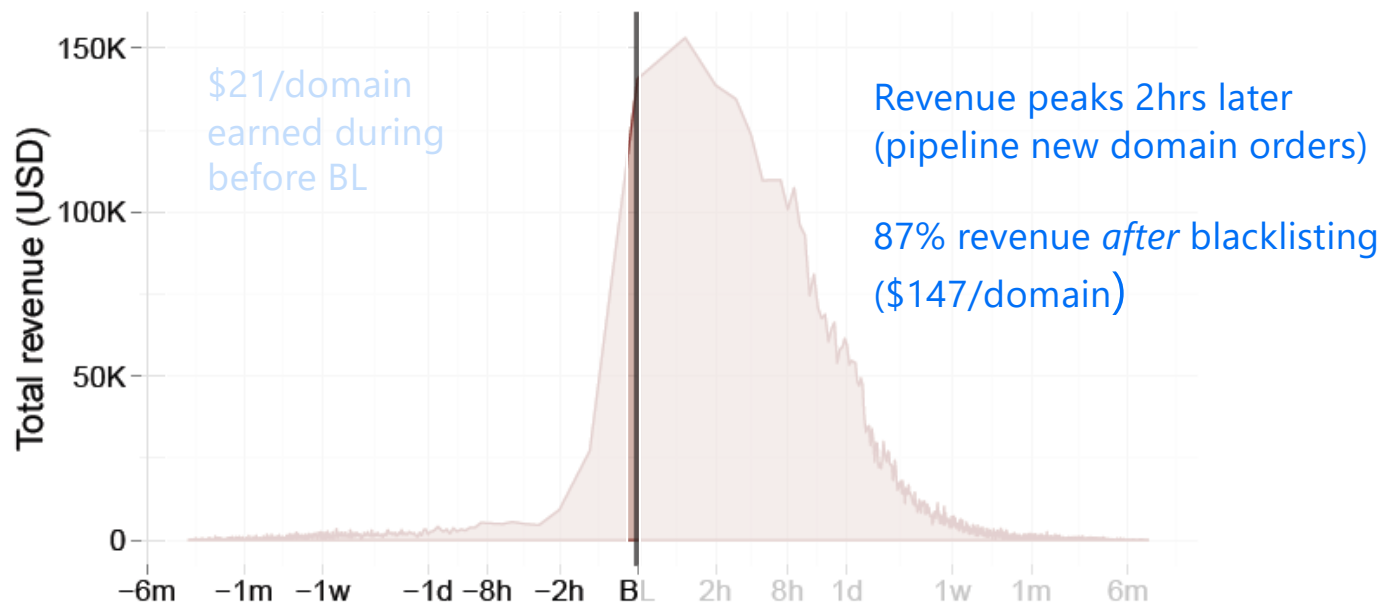
Blocklisting is fast (88% of Spamit domains listed < 48hrs of use)

Most overall *revenue* (62%) is earned via domains *never* blocklisted

Explanation? Implications?

Unintuitive reality

Most revenue from domains that *are* blocklisted occurs *after* blocklisting (~90%)



Other unintuitive reality

Web mail referrers frequently give the “folder” the click came from

– <http://mail.yahoo.com/mc/showFolder?fid=foldername...>

39% of Yahoo, >20% of Hotmail sales are from users who went into spam folder ad clicked on link

Implications?

What-if questions: extrapolated defense

- Simplified revenue model
 - Revenue per domain
 - Cost to register
- Idealized interventions
 - 100% detection
 - Post-block revenue regime
 - Filtering: empirical data
 - Takedown: \$0

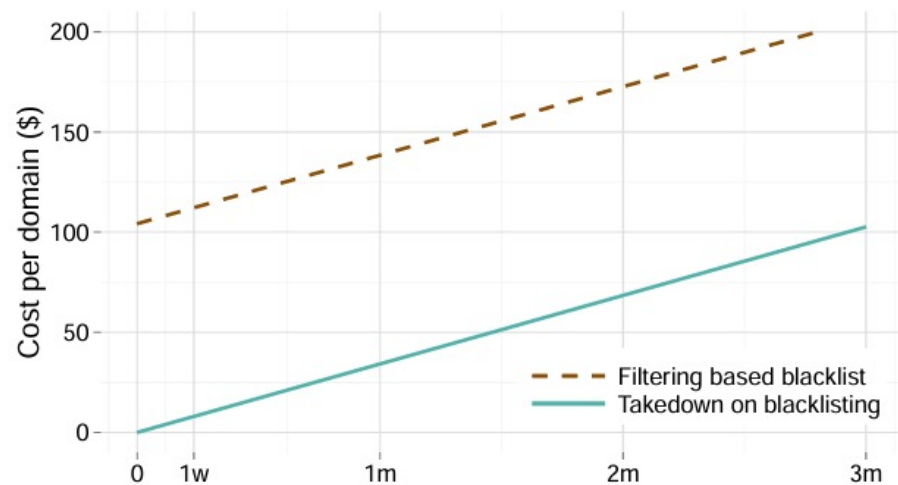


Figure 4: The highest cost of domain a spammer can afford (*y*-axis) against the time delay (*x*-axis) in blacklisting.

Is “takedown” realistic? How might that happen?

Ok, what about traffic from Search?

“Search + Seizure: The Effectiveness of Interventions on SEO Campaigns,”

Search is a different kind of vector (pull not push)

- Demand goods (user searches for them)
- Malware and click fraud
- Largely **untouched by blocklists** (<5% for GM)

How to get high-ranking search results

- Malware
- **Search Engine Optimization (SEO)**
 - Create linked content sites associated with keywords
 - Poisoned Search Results (PSRs)



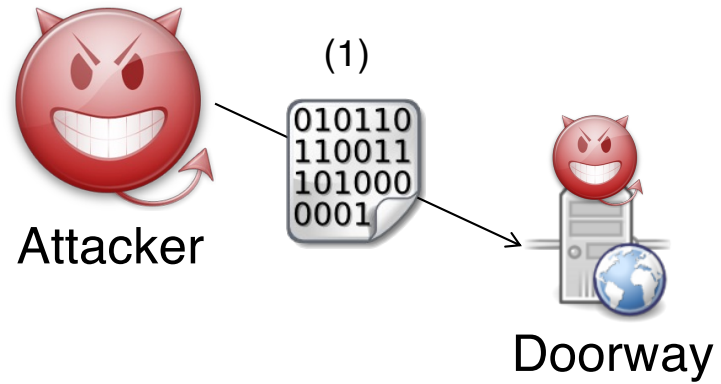
Attacker



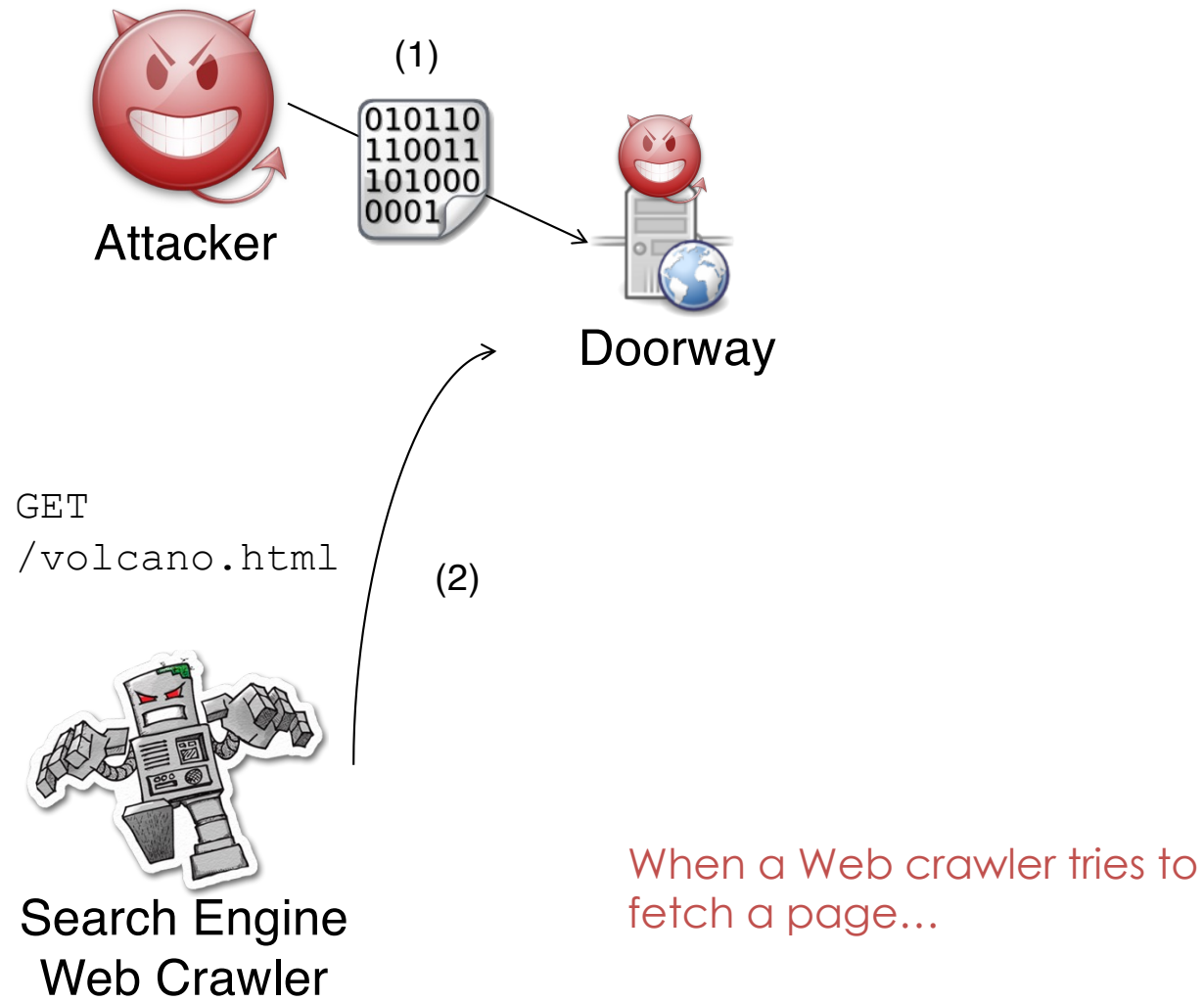
Target

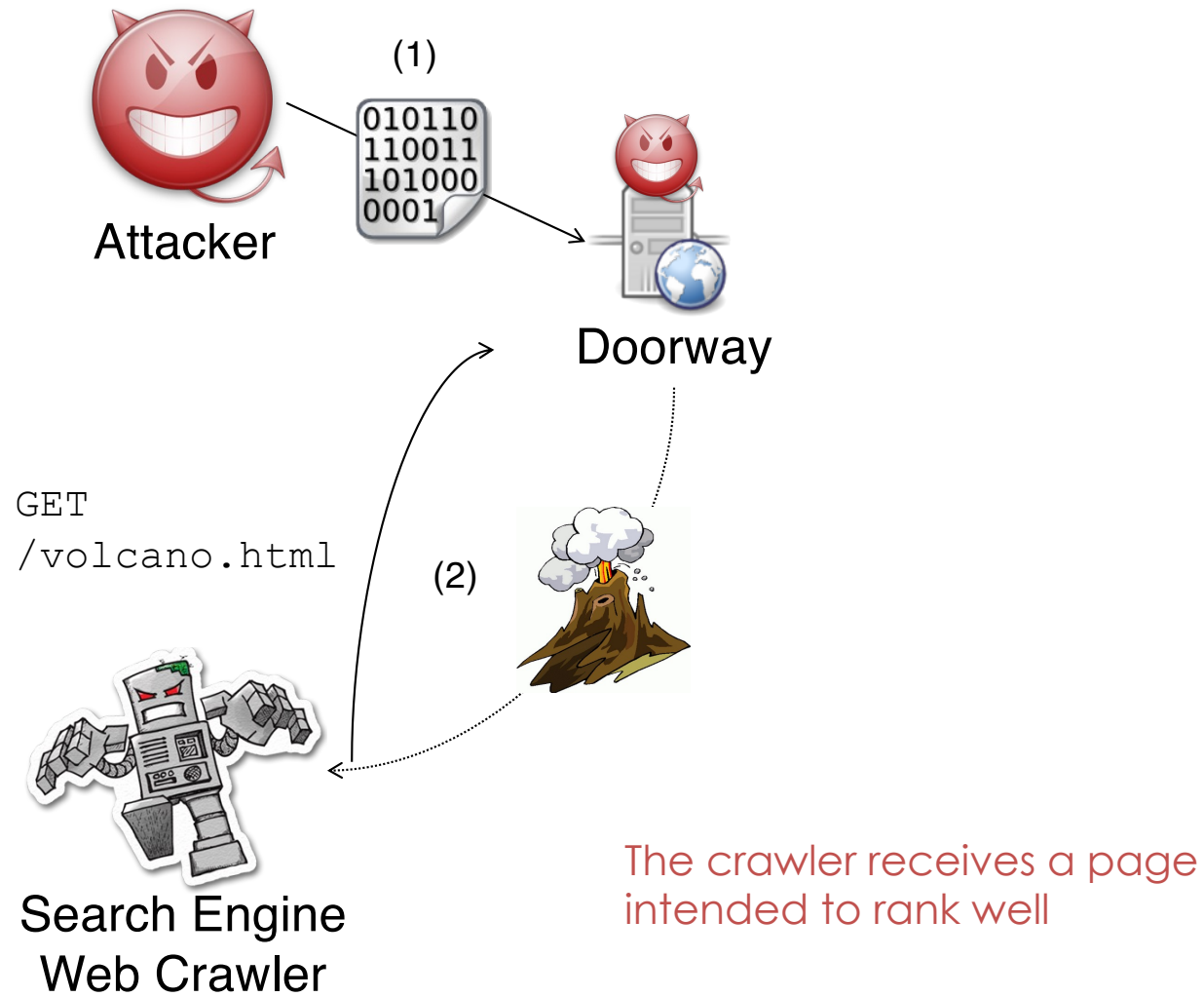
Quick walkthrough on
one kind of SEO
poisoning

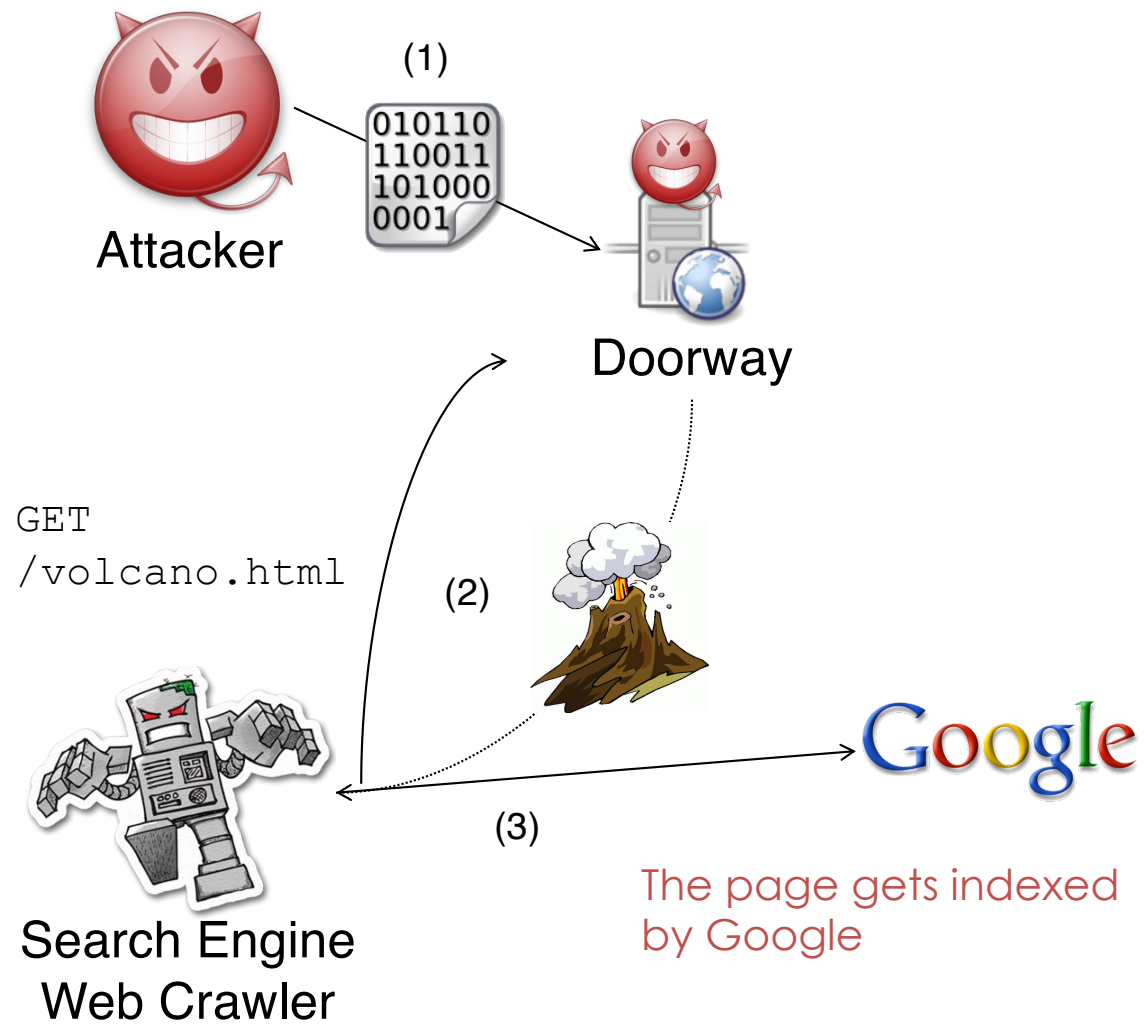
We begin with an attacker
+ a targeted Website

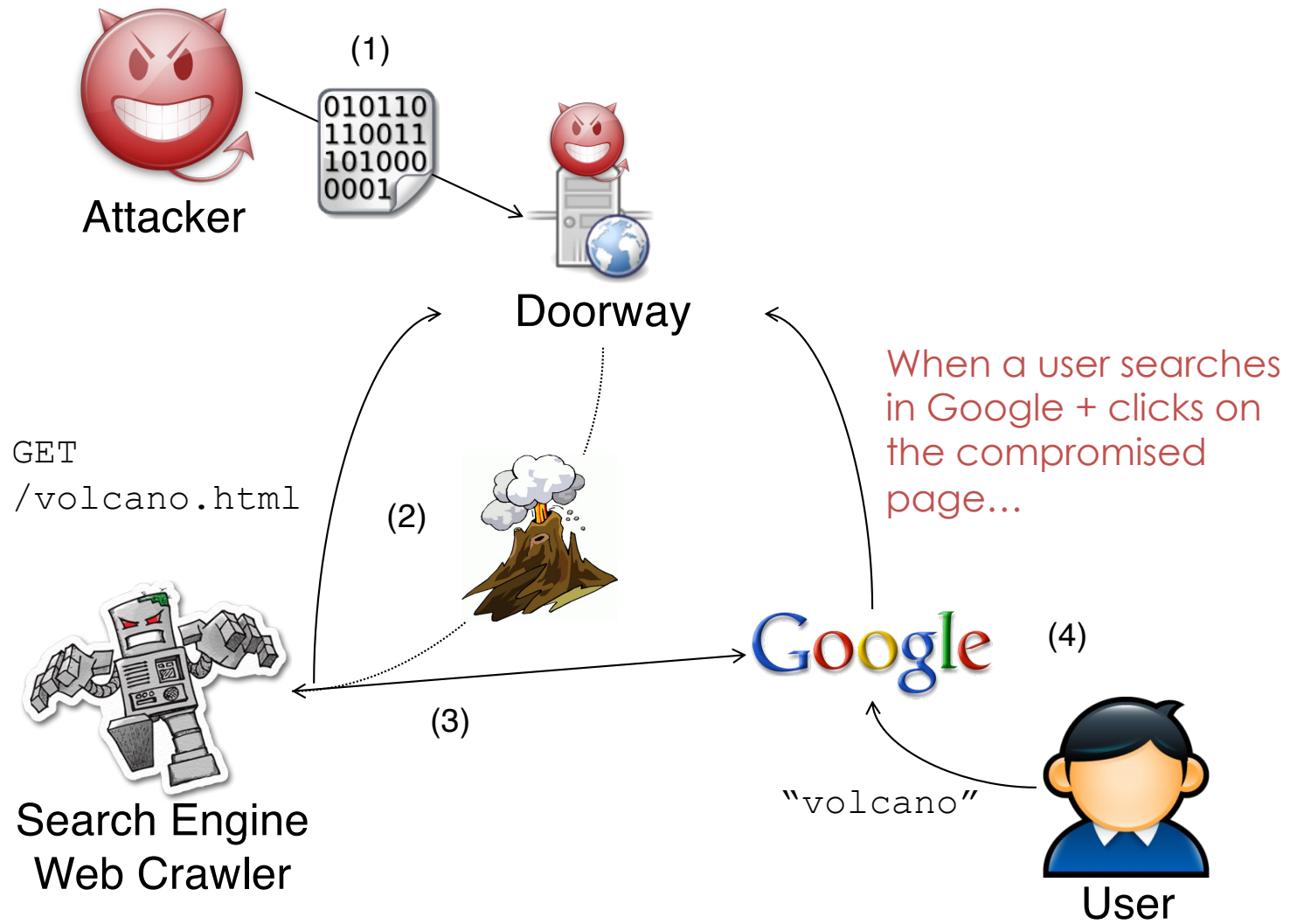


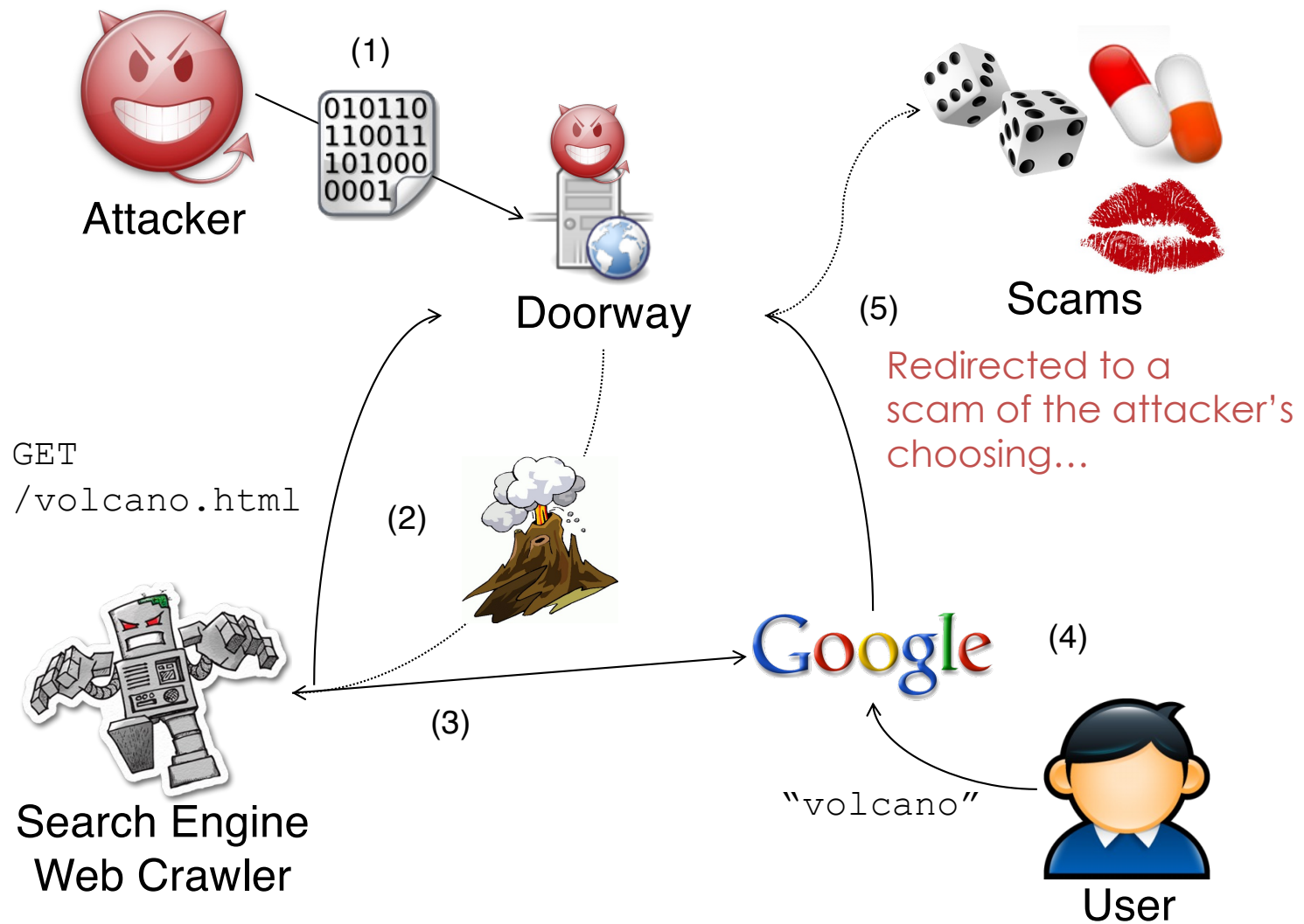
The attacker compromises the Website using an open vulnerability + installs an SEO kit











Aside: Cloaking

One of the key complexities here is “cloaking”, Web servers giving different answers to different visitors

- Shove keyword filled page to Google crawler to get it to rank well (1)
- Normal visitor sees their normal page (i.e., no reason to believe they are hacked) (2)
- Visitor who comes via Search Engine Result Page (SERP) is redirected (3) to scam/counterfeit site

Other kinds of cloaking exist

- Geographic cloaking: different answers based on geolocation of visitor
- Machine type: e.g., serve different content to Mac vs PC users
- Cookie cloaking: different content if you have a particular cookie
- History cloaking: different content if you’ve visited before

Very tricky to suss out – frequently need to enumerate different options and compare

How to find Poisoned Search Results?

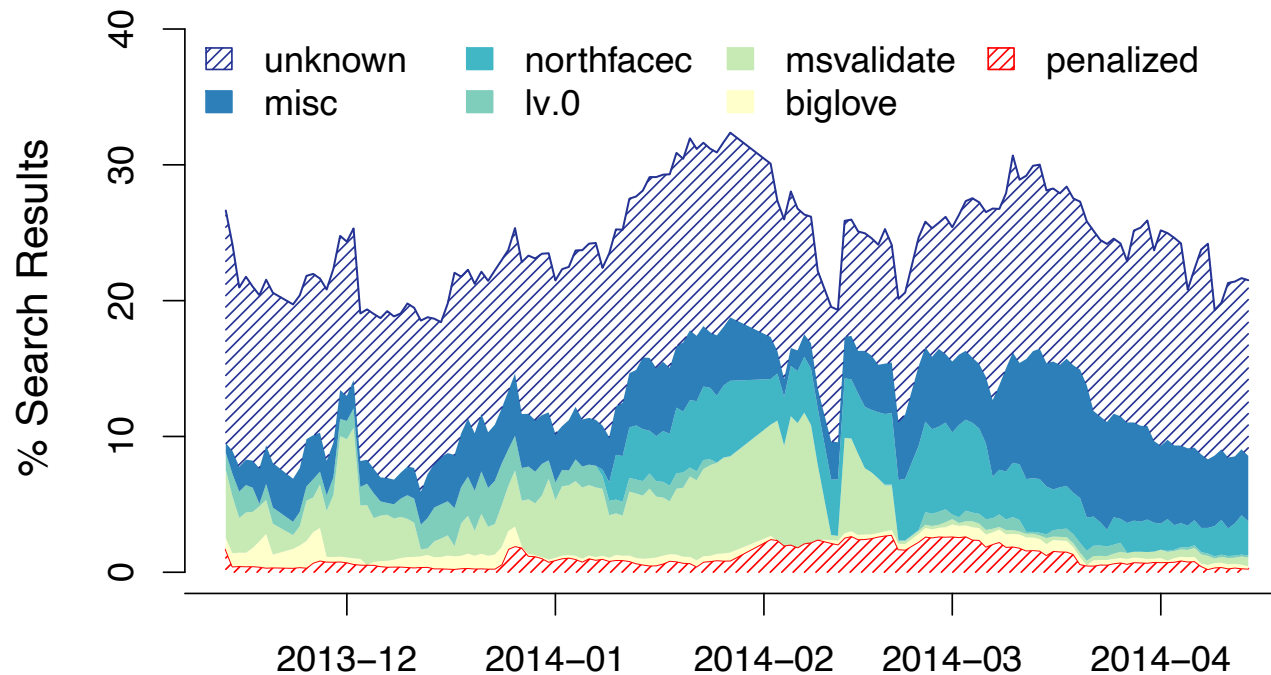
Search for likely targeted keywords

- Repeatedly over time
- Issues with this approach?

In the paper, they identify 28 distinct campaigns with PSRs in the first 100 responses

	Total	Classified
# PSRs	2M	1.1M (55%)
# Doorways	19K	8K (42.1%)
# Stores	5.5K	513 (9.3%)
# Campaigns	28+	28

Counterfeit brand Web sites (as seen via Google search)



30% poisoned! Top 10 also (20%)

Only small fraction labeled by Google as bad (2.7%)

Type of Interventions

Two groups with most incentive and greatest means to intervene

Search engine

- Demote do
- Label doorn

Brand holder

- Domain sei

This paper j

Cheap Beats By Dre Studio Outlet On Sale

www.nwtrappers.com/ ▼

Authentic **beats by dre studio** 2014 free shipping worldwide. ... low-cost beats by just dre headphones is to be found. **beats by dre** on sale,**beats by dre** sale.

New Arrival Beats By Dre Pro Studio Outlet Cheap Sale

www.sebcoax.com/ ▼

This site may be **hacked**.

The best headphones are **just** the thing for iPods, i-phones, and compact audio participants. **beats by dre studio**. wireless **beats by dre**. If you are willing to ...

Beats by Dre Studio High-Definition Headphones JustBeats...

www.sainsmart.com/vanilla/.../beats-by-dre-studio...justbeats-by-dre.../p1 ▼

Like Monster Beats, the products made by them are **Beats By Dre** Headphone Solo Wireless White Beats Solo Wireless White so good that the **Studio** takes ...

engine result page)

feit storefronts

Quick aside: revenue inference

This Web paper didn't have the luxury of having "ground truth"

Need to estimate revenue and/or change in economic activity?

- What are ways you might try to do that? (that weren't in the paper)

The paper uses a technique called "**purchase pair**"

Purchase pair technique

support@drugs-support-24.com to gheart show details Jan 11 Reply

support@drugs-support-24.com to ck2171 show details Jan 11 Reply

Dear Customer, your order was processed. The status of your order is "Sent to Warehouse".

support@rx-drugs-support.com to ghasrxp show details Feb 8 Reply

Dear Customer, your order was processed. The status of your order is "Sent to Warehouse".

We have successfully checked all your data and prepared your order for the shipment.

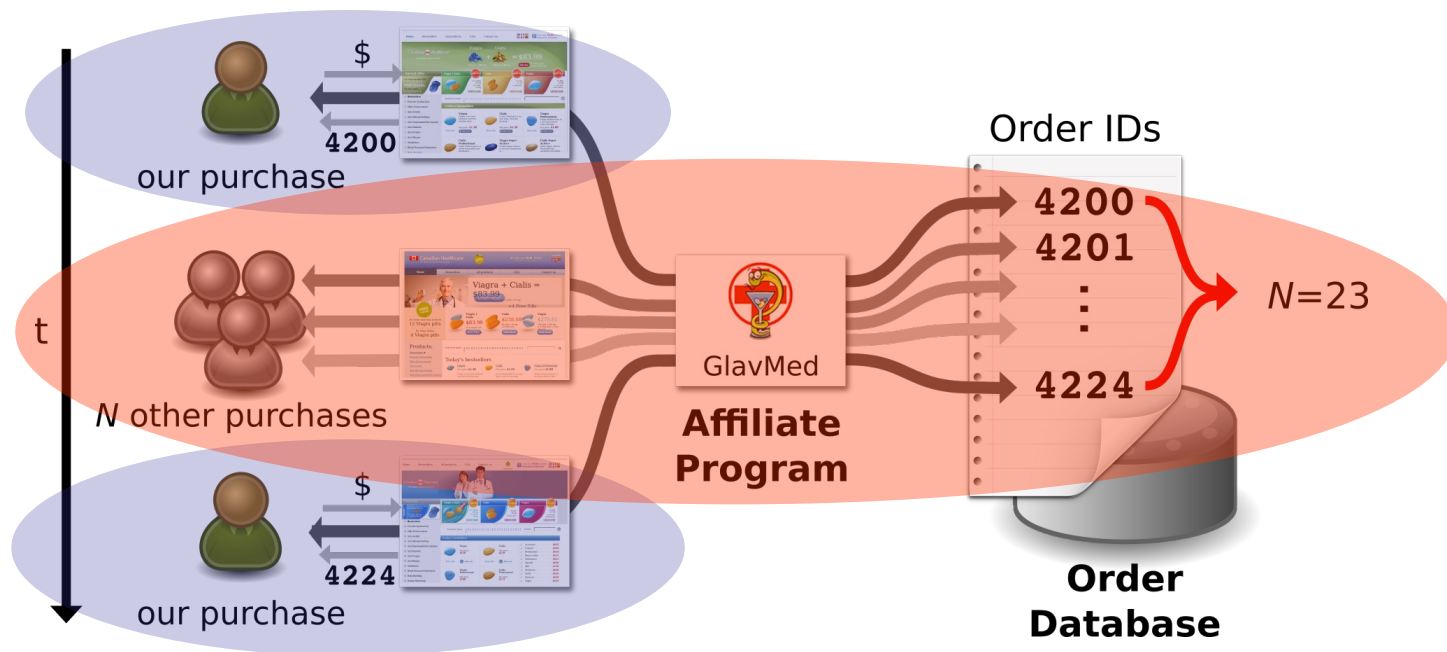
The transaction ID is 553809
Your Order ID is 496427
You have purchased the following products at <http://the-us-drugs.com/> :

VPXL 1 bottle x 1 pc , qty: 1, amount: 37.99 USD

Total amount of your order is: 52.99 USD (with shipping)

Kanich, Weaver, McCoy, Halvorson, Kreibich, Levchenko, Paxson, Voelker, Savage, *Show Me the Money: Characterizing Spam-advertised Revenue*, USENIX Sec 2011

Order Throughput Inference



Valida
seque

Stand

Cons

Cons

Time

average order size to gauge overall revenue [9]. Across four years, we find that a significant number of order numbers never appear in the database due to either filtering for customer fraud or shopping cart abandonment (between 13–28% for SpamIt and 7–17% for GlavMed). The lower number of absent orders for GlavMed is likely because the search engine vector used by its affiliates generates less antipathy among consumers. In both cases, 8–12% of the orders that do appear in the database are ultimately declined and do not ship. Consequently, true turnover is between 8% (low of GlavMed) and 35% (high of SpamIt) less than predicted by the “purchase pair” technique. However, since the average successful order size is between \$115 (GlavMed) and \$135 (SpamIt), revenue estimates based on an average sale of \$100 are roughly in-line with true revenue (within 6% overall for GlavMed and 13% overall for SpamIt).

Year.Month

Domain Seizure

CHANEL,INC. v. CHANEL-BESTSELL.COM, et al.

On June 12, 2014, in the United States District Court for the Southern District of Florida, Chanel, Inc. filed a Complaint for Injunctive Relief. The filings and orders in this matter are available below.

NOTICE TO DEFENDANTS

If you fail to respond to this complaint, judgment by default will be entered against you for the relief demanded in the complaint.

To prevent this from happening you must file a response with the court clerk or administrator within 21 days of the date this message was posted. Your response must be in proper form and have proof of service on the plaintiffs' attorney, Stephen M. Gaffigan, P.A, 401 East Las Olas Boulevard, Suite 130-453, Ft. Lauderdale, Florida 33301.

06/12/2014 COMPLAINT FOR INJUNCTIVE RELIEF

[Case 0:14-cv-61366-WJZ](#) [Document 1](#) Entered on FLSD Docket 06/12/

**UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF FLORIDA
CASE NO.**

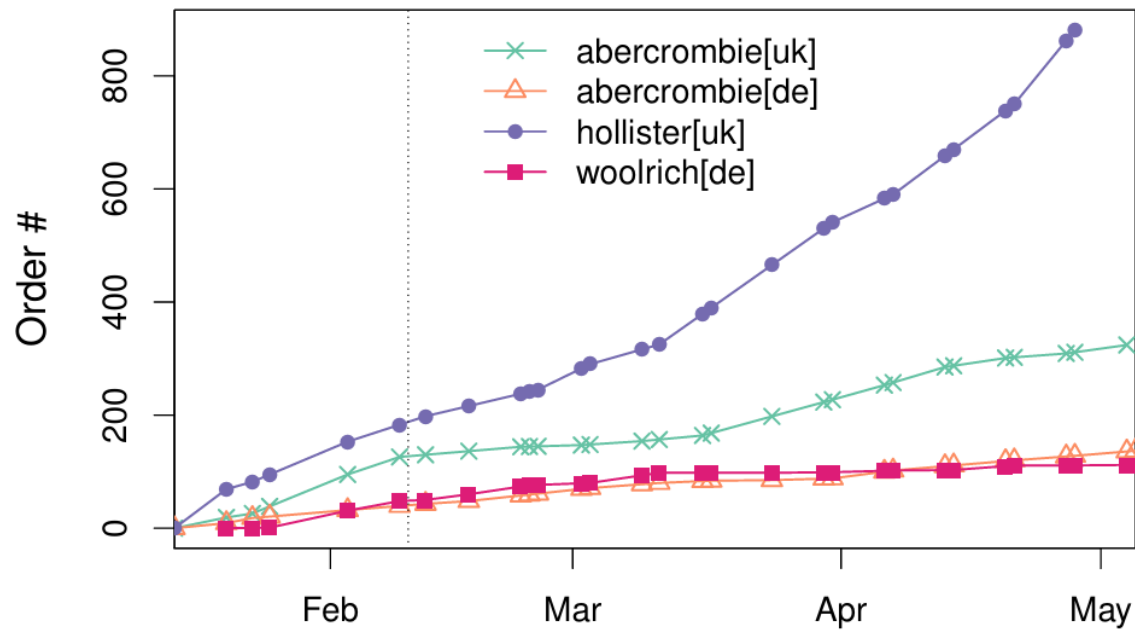
CHANEL, INC.,

Plaintiff,

vs.

IS
7
2

Takedown can reduce order volume (Abercrombie.uk – more anecdotal)



But ineffective overall

<4% of domains found in PSRs were taken down via seizures

- But didn't do per brand analysis... why might that have mattered?

Average lifetime of a domain

- From when it first showed up in PSRs (i.e., via searches) to court order:
~60 days

Redirection after seizure

- 64% of doorway pages just redirect to new domain
- Why can't domains for doorway pages be seized?

Aside: more found data – why interdiction isn't the answer

Got inside **one** of the wholesale suppliers of counterfeit luxury goods

2013-11-29 16:50:30	Malaysia MY8930, customers sign: BATIAR ✓
2013-11-29 16:50:00	Malaysia MY8930, delivery and receipt
2013-11-28 15:04:00	Malaysia MY8800, not Tuotou
2013-11-28 14:55:00	Kuala Lumpur, Malaysia, customs clearance
2013-11-26 23:25:00	Kuala Lumpur, Malaysia sent to Customs
2013-11-26 23:24:00	Kuala Lumpur, Malaysia, to reach the processing center from Shanghai, China
2013-11-23 23:32:02	Shanghai, leave the treatment center, sent to Kuala Lumpur
2013-11-23 20:10:11	Shanghai to reach treatment centers, bulk mail acceptance from the Shanghai Office
2013-11-23 19:48:00	Shanghai bulk mail acceptance office, leaving the treatment center, sent to Shanghai Company
2013-11-23 19:24:00	Shanghai Office bulk mail acceptance, acceptance
2013-11-23 15:49:05	Operations Center sent to the next stop
2013-11-23 15:48:20	Arrival in sorting operations center

80100184962 detailed tracking information

Date and Time	State
---------------	-------

Questions?

For Tuesday

Read at least one of:

Zesen Zhang, Alexander Marder, Ricky Mok, Bradley Huffaker, Matthew Luckie, kc Claffy, Aaron Schulman. [Inferring Regional Access Network Topologies: Methods and Applications](#), In *Proceedings of the ACM Internet Measurement Conference (IMC'21)*, November 2021, pp. 720–738.

Nishant Bhaskar, Maxwell Bland, Kirill Levchenko, Aaron Schulman. [Please Pay Inside: Evaluating Bluetooth-based Detection of Gas Pump Skimmers](#), In *Proceedings of the 28th USENIX Security Symposium*, August 2019, pp. 373–388.