Juice: A Longitudinal Study of an SEO Campaign

David Y. Wang, Stefan Savage, and Geoffrey M. Voelker

University of California, San Diego

UCSDCSE
Computer Science and Engineering

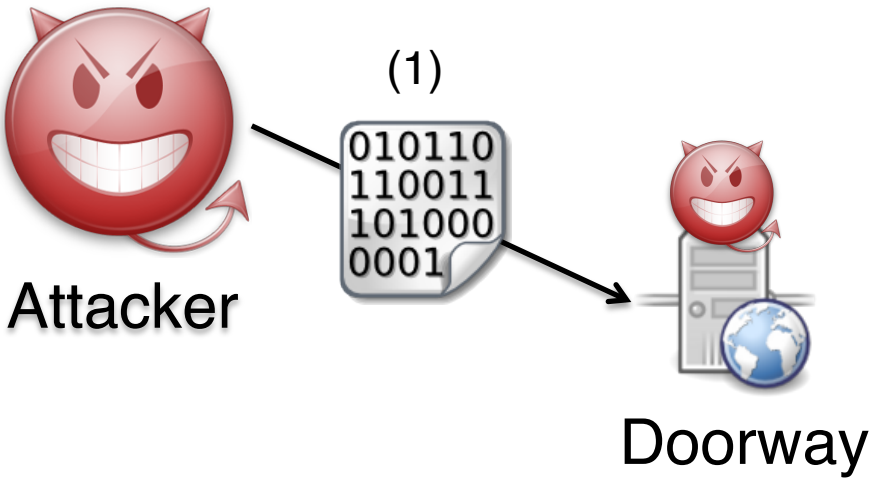
Background

- A Black Hat **S**earch **E**ngine **O**ptimization **(SEO)** campaign is a coordinated effort to obtain user traffic through abusive means
  - Supported by botnet of compromised Web Sites
  - Poison search results
  - Feed traffic to scams (e.g. Fake Anti-Virus)
- **Link Juice** refers to the backlinks (references) a site receives
  - Believed to influence search result ranking

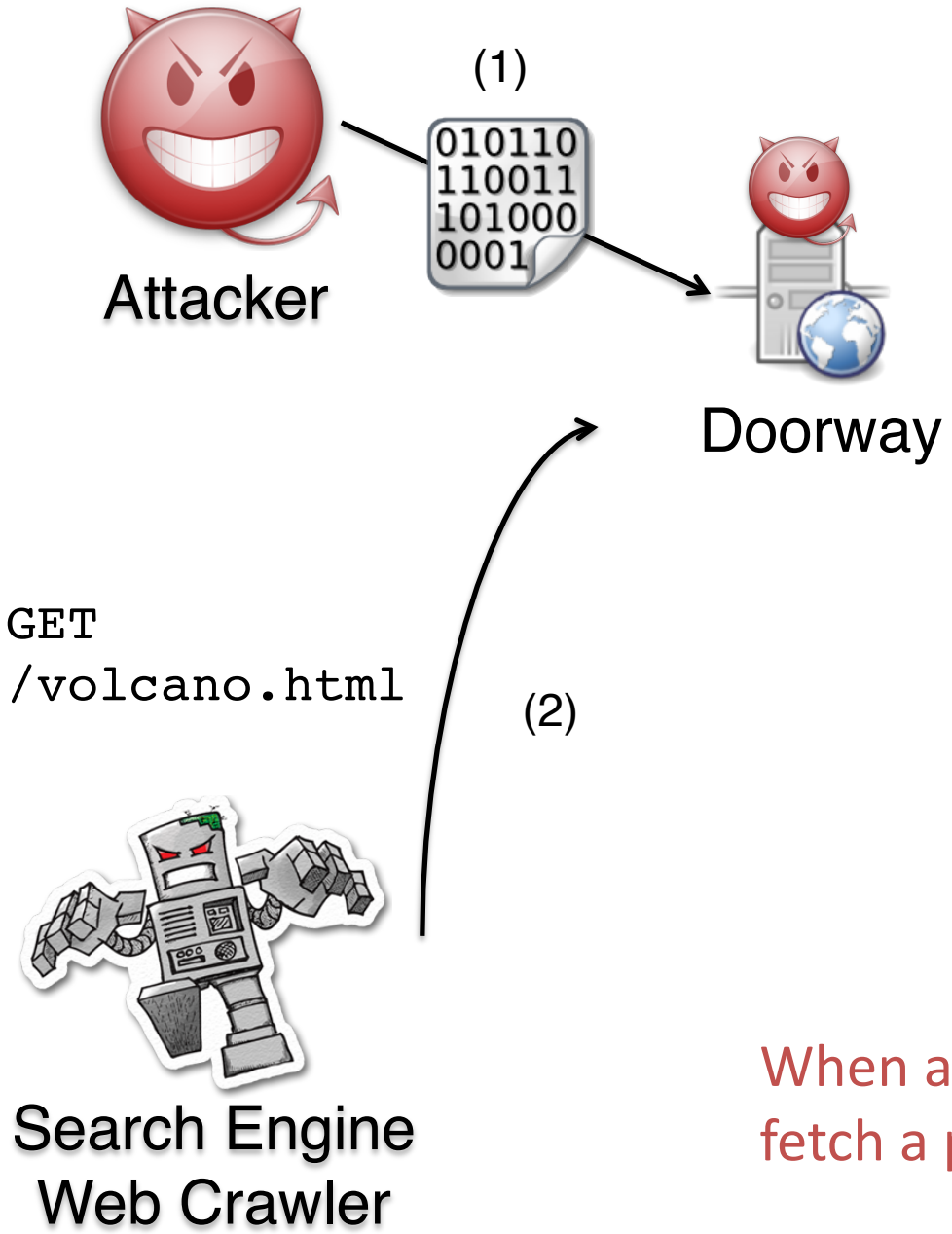
Attacker

Doorway

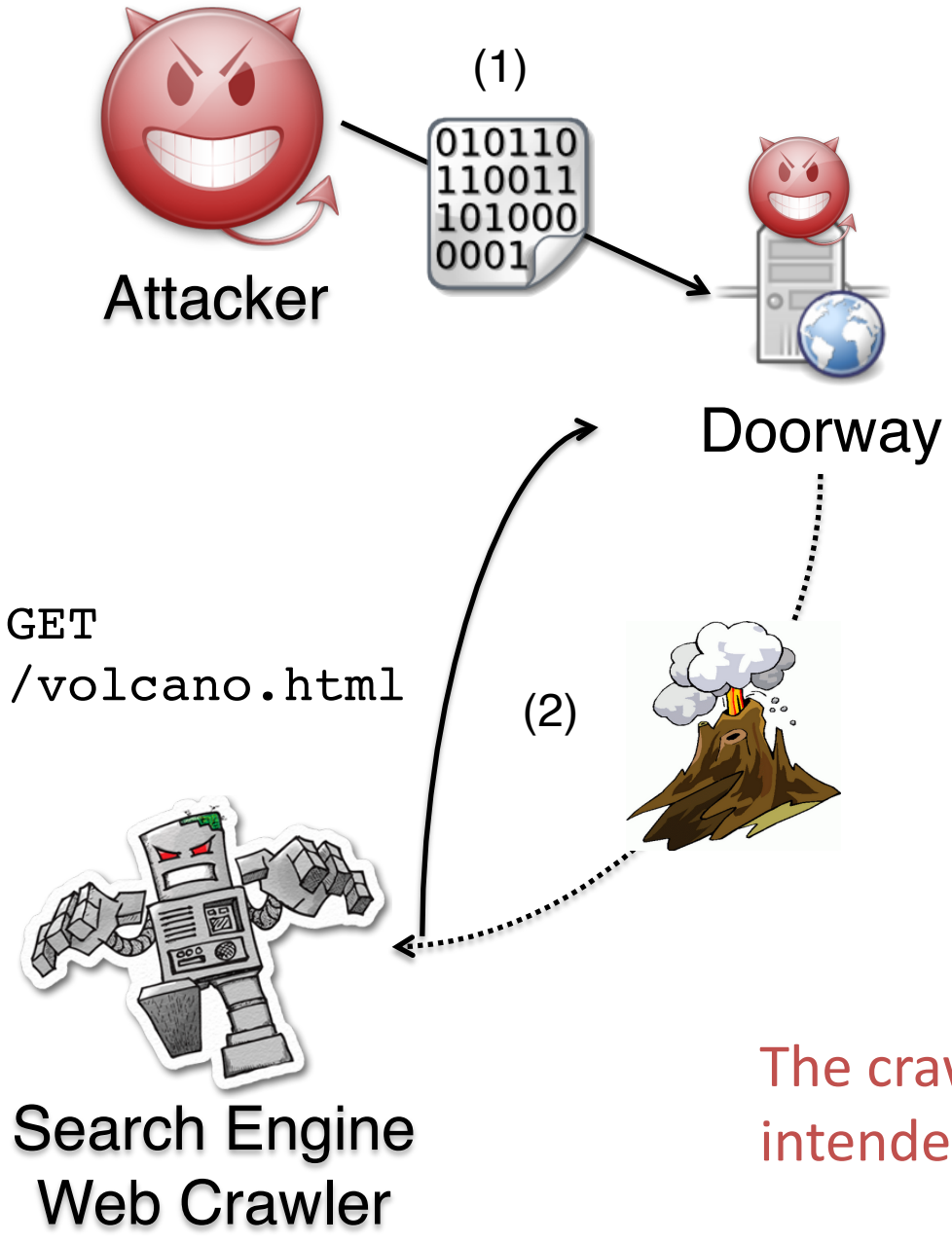
We begin with an attacker
+ a targeted Website



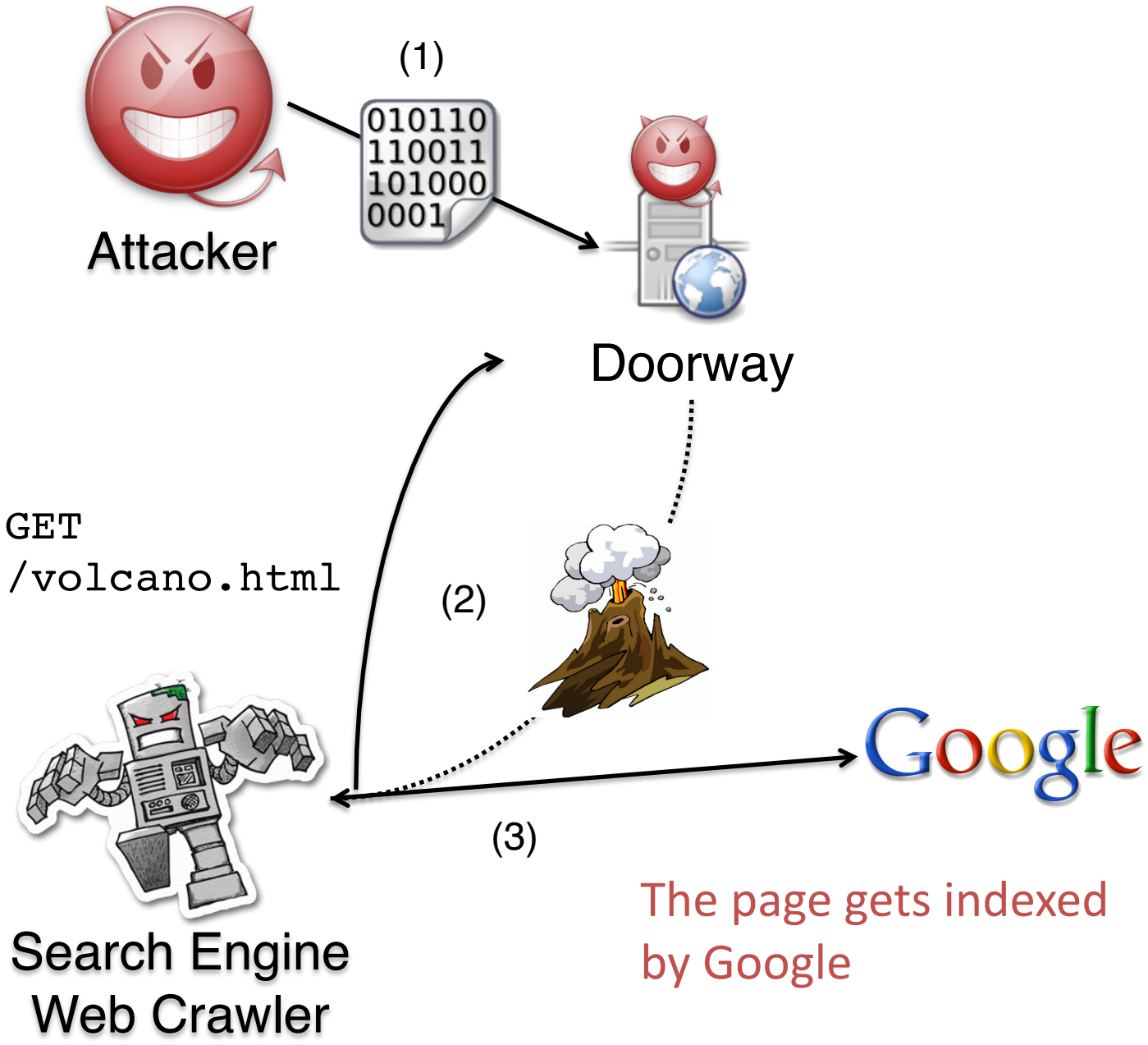
The attacker compromises the Website using an open vulnerability + installs an SEO kit

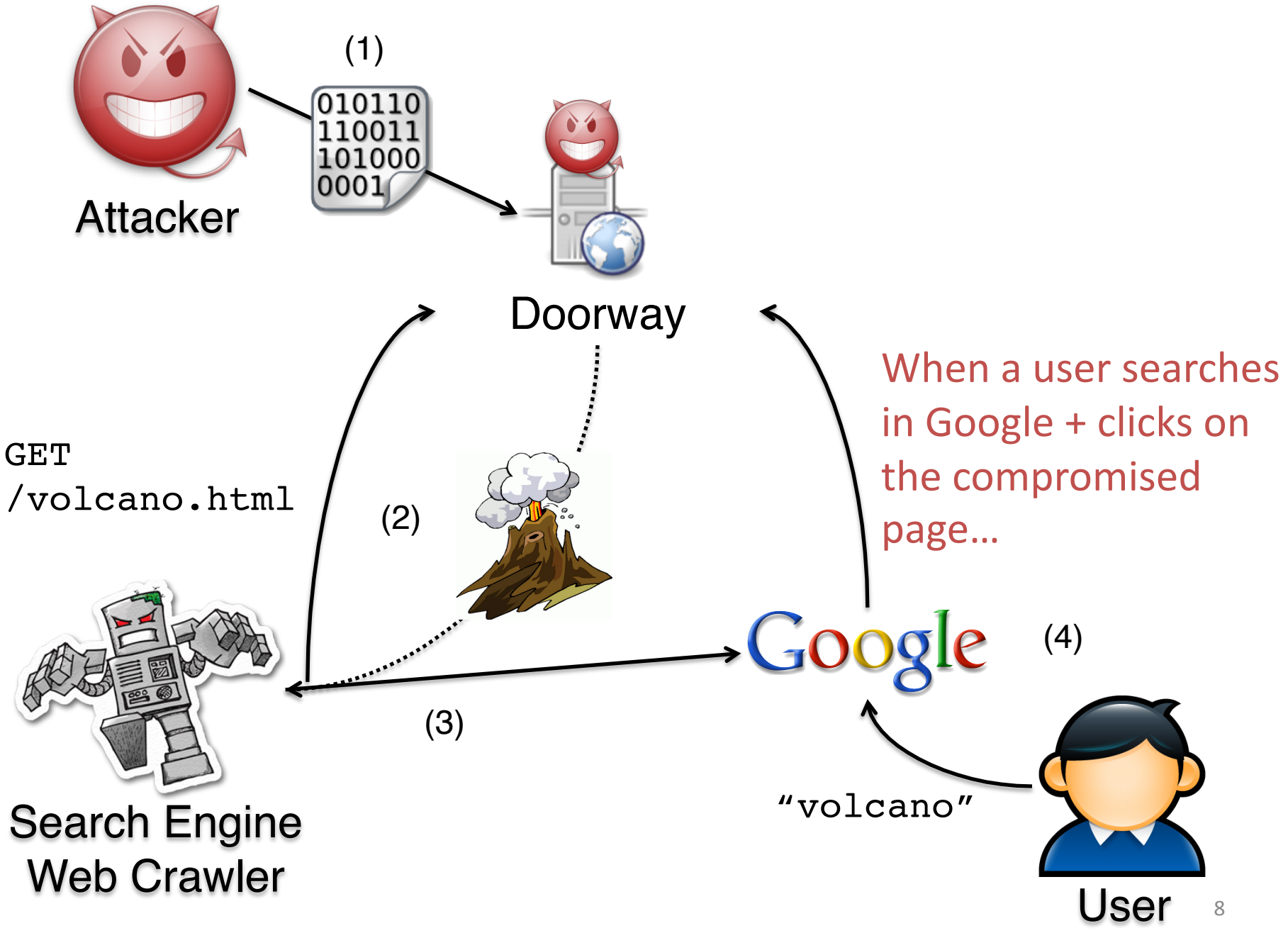


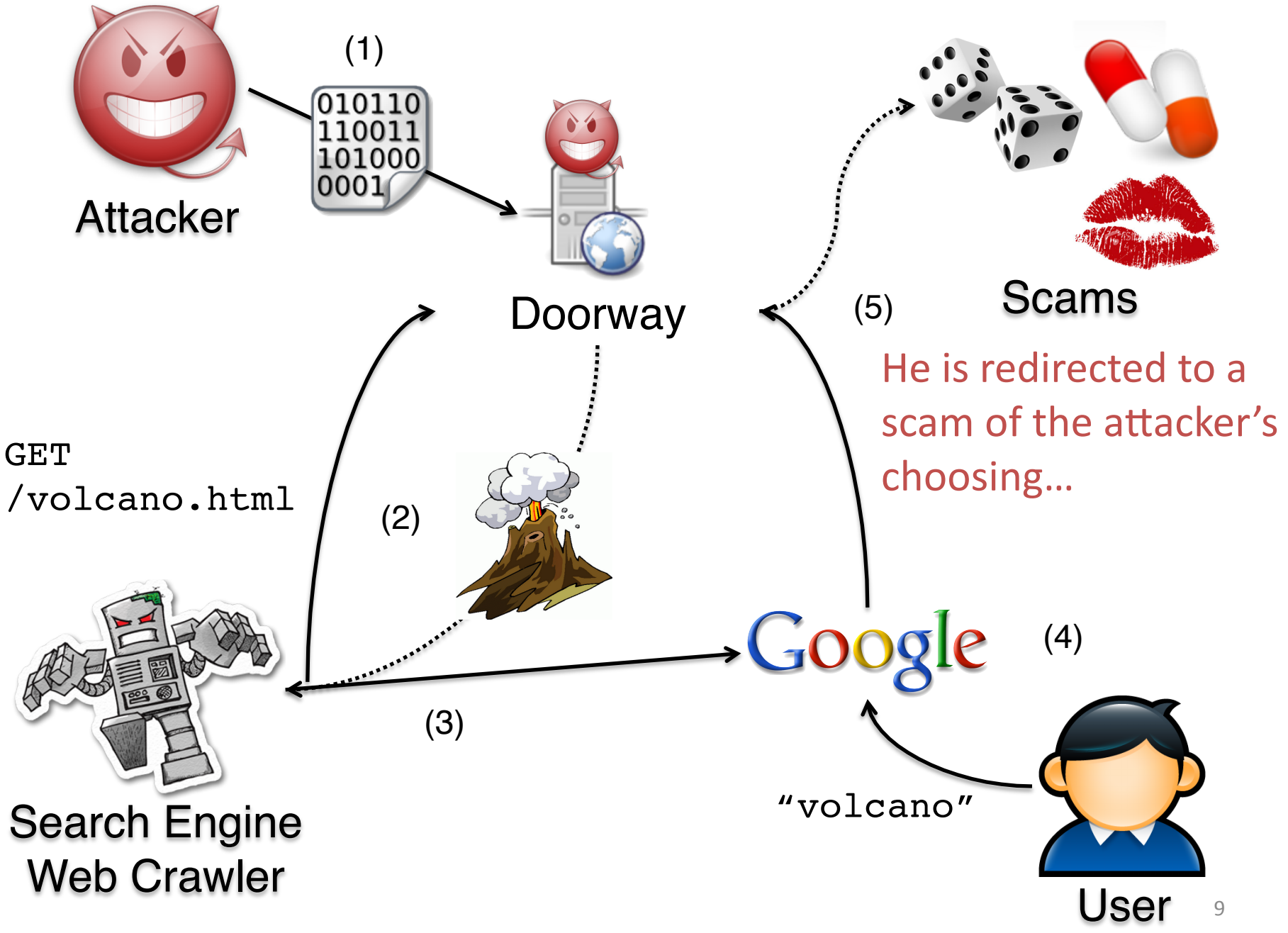
When a Web crawler tries to fetch a page…



The crawler receives a page intended to rank well


(1)
010110
110011
101000
0001
Attacker
Doorway
GET
/volcano.html
(2)
Google
(3)
Search Engine
Web Crawler
The page gets indexed
by Google


(1)
010110
110011
101000
0001
Attacker
Doorway
When a user searches in Google + clicks on the compromised page…
GET
/volcano.html
(2)
Google
(4)
(3)
"volcano"
Search Engine
Web Crawler
User


Attacker
(1)
010110
110011
101000
0001
Doorway
Scams
(5)
He is redirected to a scam of the attacker's choosing…
GET
/volcano.html
(2)
Google
(4)
(3)
"volcano"
Search Engine
Web Crawler
User

Our Contributions

- Infiltrate an influential SEO botnet **(GR)**
  - In depth characterization of GR's operation
    - One time leader in poisoned search results on Google
  - Our work builds on previous work studying search result poisoning [John11, Lu11, Moore11]
- Draw insights from *combining data from three separate data sources (crawlers)*:
  - Estimate GR's effectiveness
  - Examine impact of scams funding GR

SEO Kit

- An **SEO kit** is software installed on compromised sites
  - Allows backdoor access for botmaster
  - Performs Black Hat SEO (i.e. cloaking, content generation, user redirection)
  - Typically they are obfuscated code snippets injected into pages

```
<?php
if(!function_exists('cm4y2wui5w153')) {
    function cm4y2wui5w153($smcx)
    {$dix5xk='x);';…}
?>
```

```
<?php
// Общее
define("GR_CACHE_ID", "v8_cache");
define("GR_SCRIPT_VERSION", "v8.0
(28.02.2012)");
?>
```

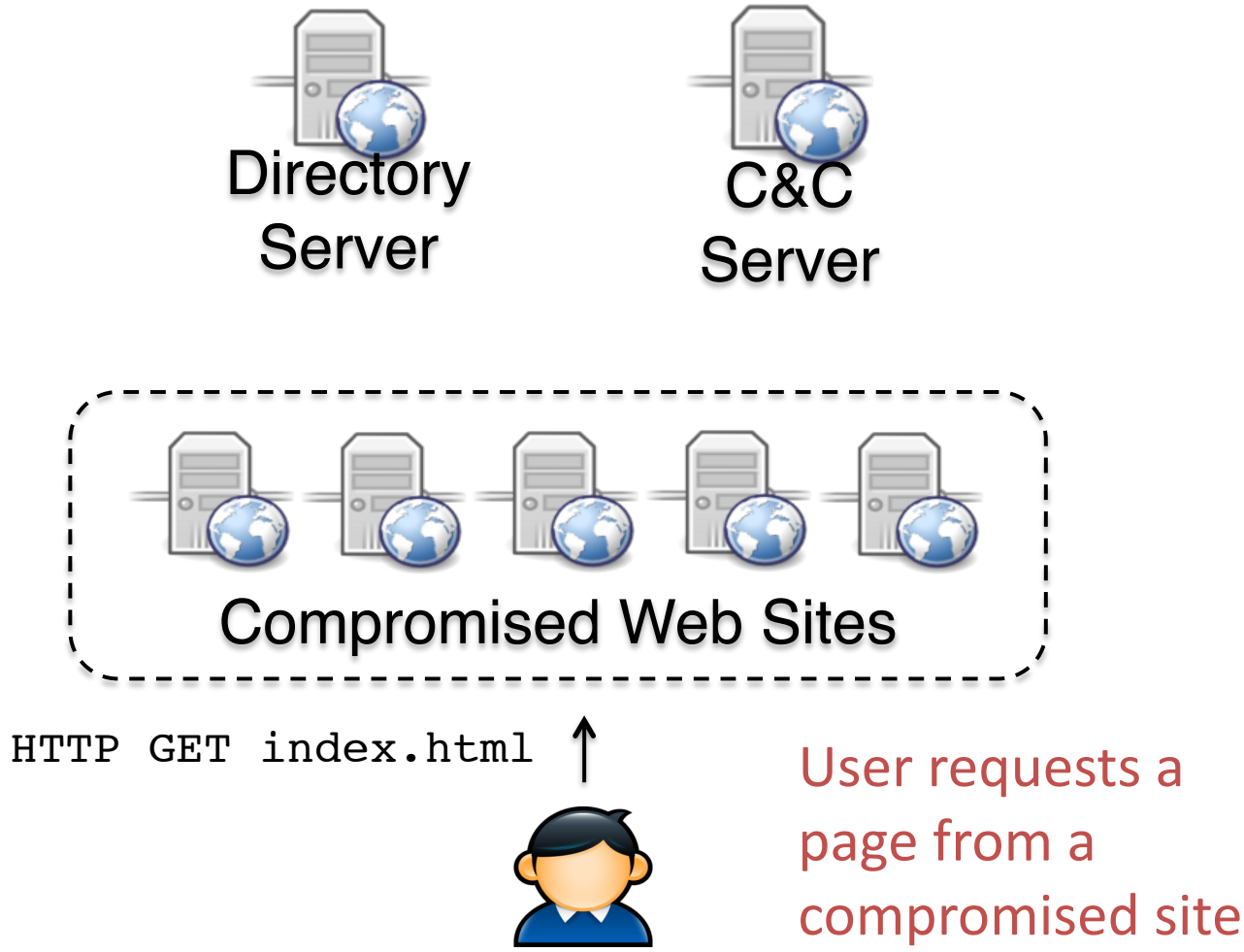
Anecdote

- Obtained a copy of the GR SEO kit by contacting owners of compromised sites
  - Roughly *40 attempts*
  - A *handful* were willing to help
  - But, only *1 person was able* to disinfect their site and send us the kit
- The SEO kit allows us to infiltrate the botnet and understand how the campaign works

GR Botnet Architecture

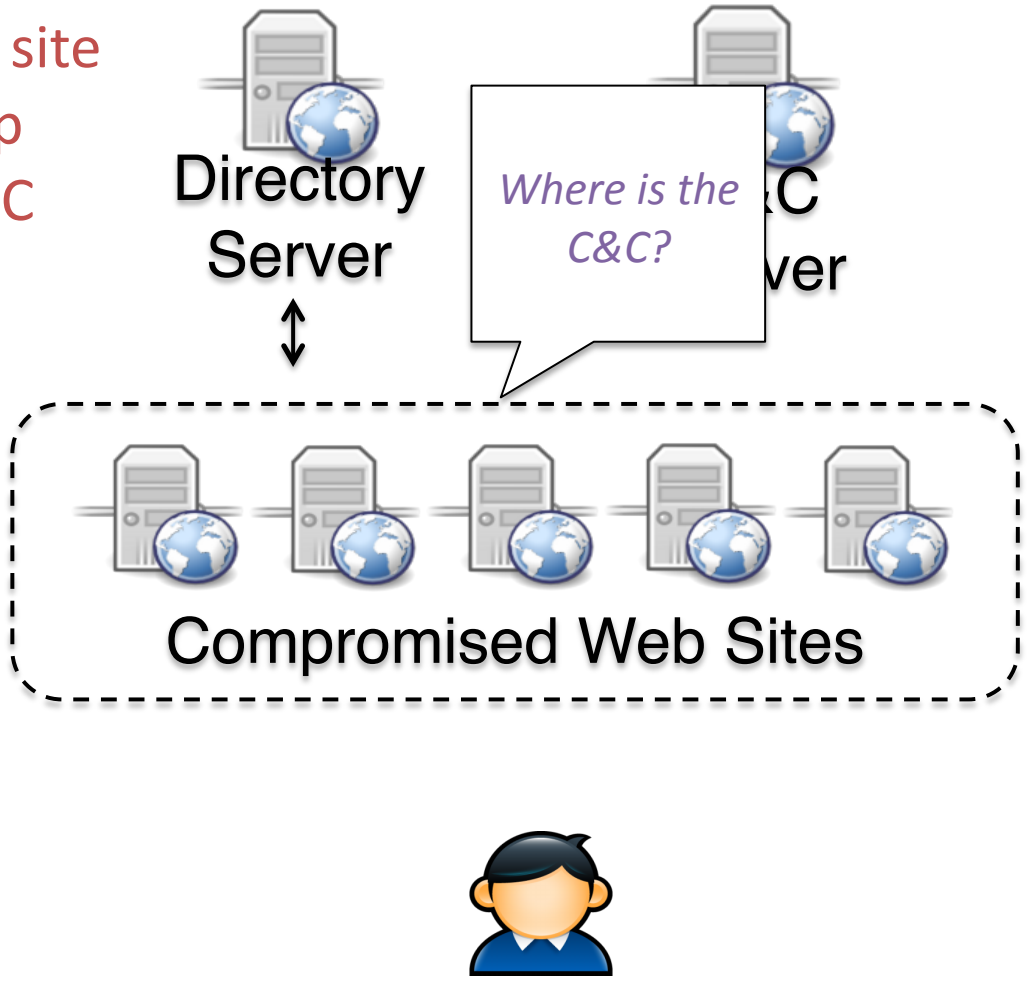
- The GR Botnet is built using **pull mechanisms** and is comprised of **3 types of hosts**:
  - Compromised Web Sites act as *doorways* for visitors and control which content is returned
  - The Directory Server's only role is to *return the location of the C&C Server*
  - The C&C Server acts as a *centralized content server* for the GR Botmaster

Example of User Visit



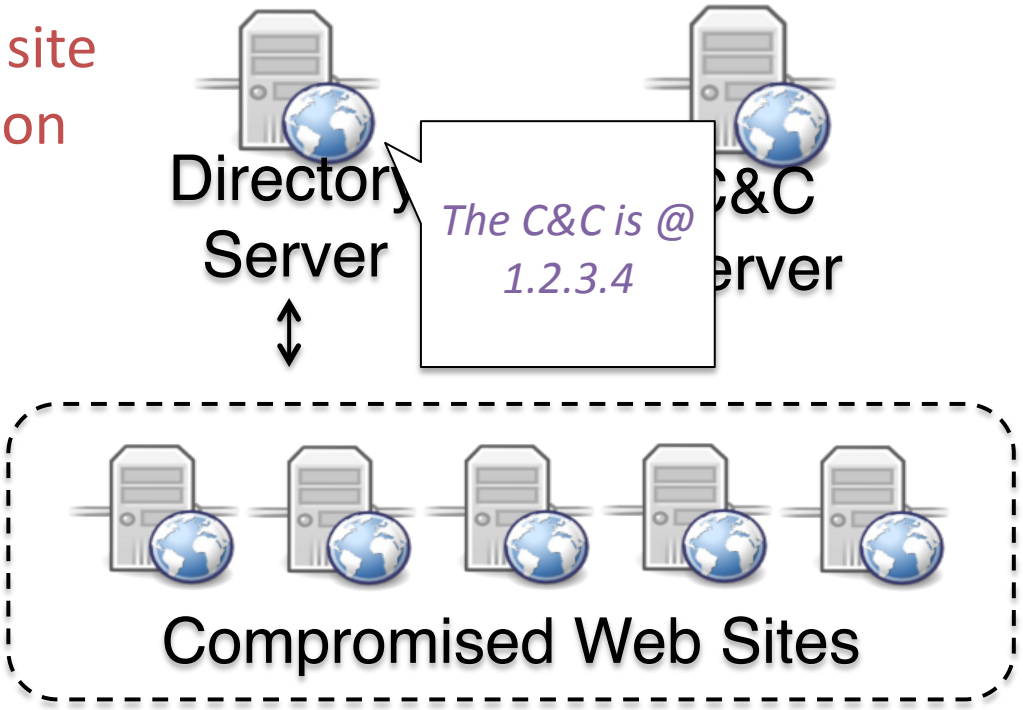
Example of User Visit

Compromised site tries to look up location of C&C

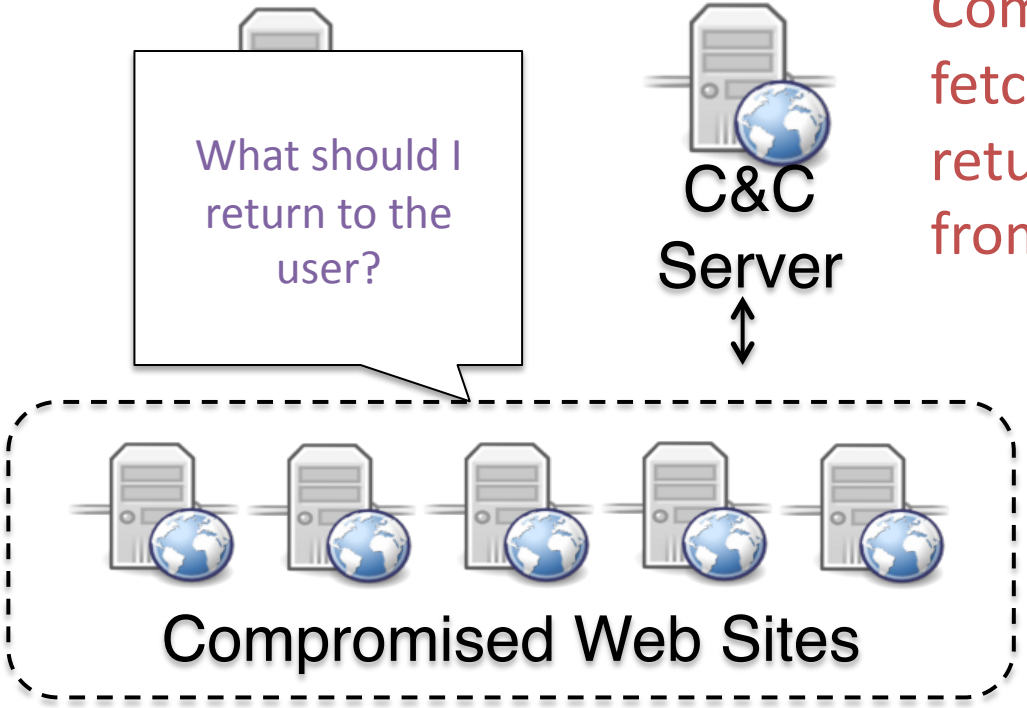


Example of User Visit

Compromised site looks up location of C&C Server

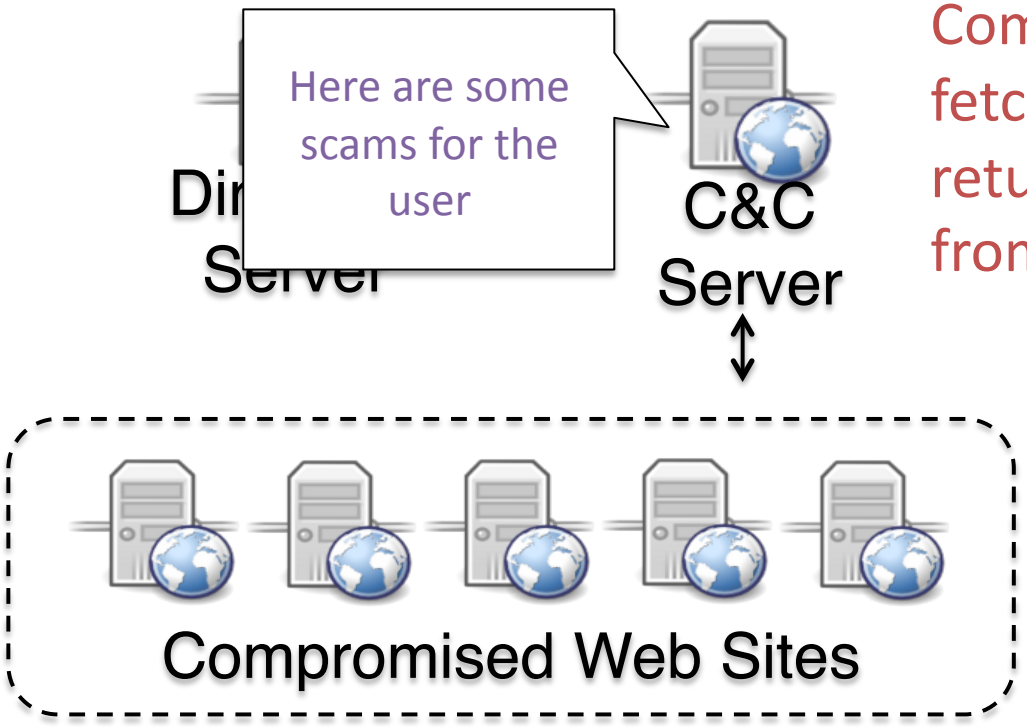


Example of User Visit



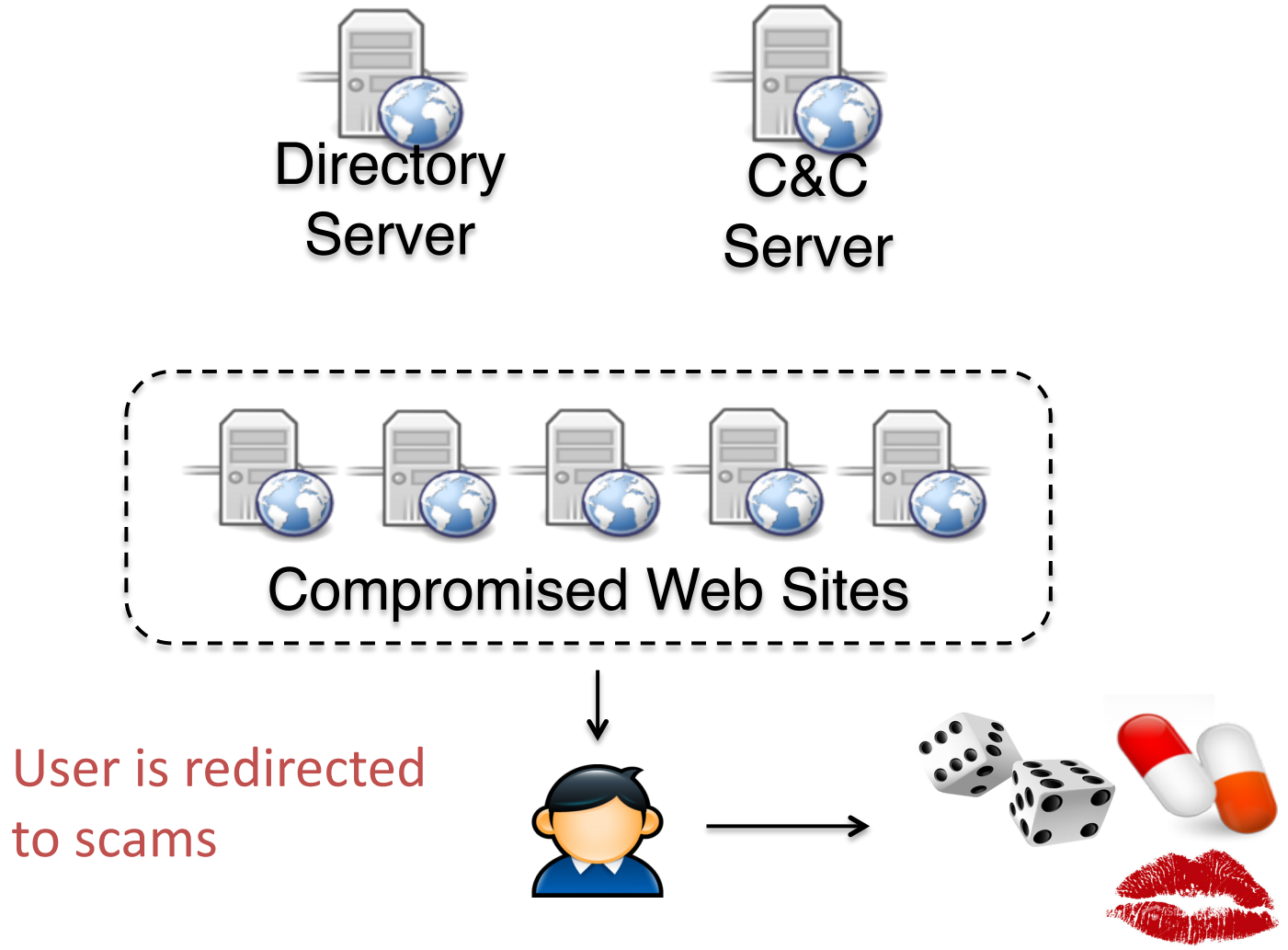
Compromised site fetches content to return to user from C&C Server

Example of User Visit



Compromised site fetches content to return to user from C&C Server

Example of User Visit



Data Collection

- We collect data using 3 distinct crawlers
  - **Odwalla** *crawls and monitors compromised sites in the GR botnet* (October 2011 – June 2012)
  - **Dagger** *measures poisoned search results for trending searches* (April 2011 – August 2011)
  - **Trajectory** *crawls pages using a Web browser to follow redirects* (April 2011 – August 2011)
- Although timeframes do not overlap cleanly, we can still draw insights

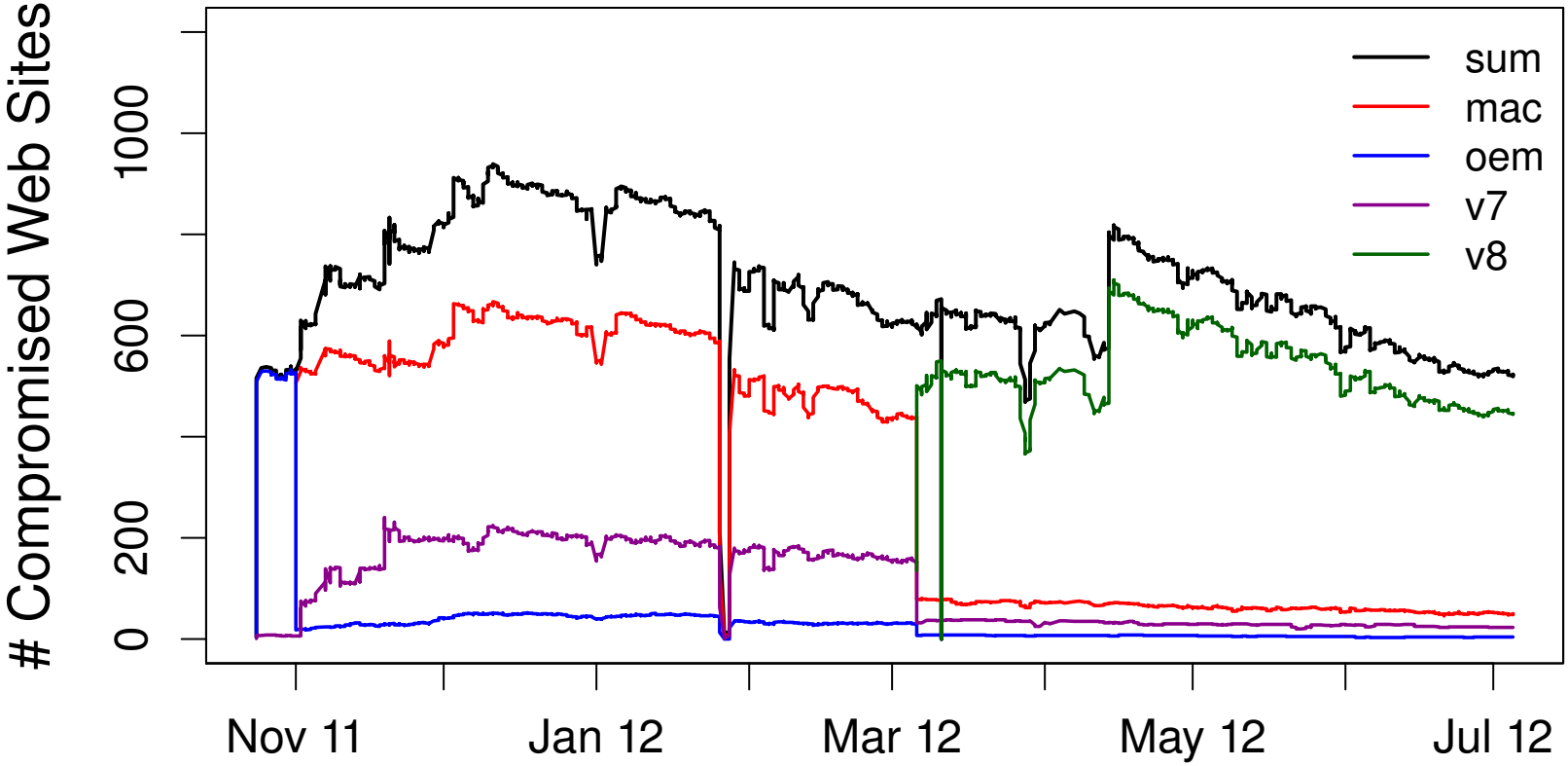
Odwalla

- **Odwalla** crawls GR's topology
- Begin w/ poisoned search results [Dagger]
- Takes advantage of two characteristics of the compromised sites in GR:
  - Sites respond to the C&C protocol by returning diagnostic information (easy confirmation)
  - Sites are cross linked with other compromised sites in order to manipulate search rankings (find more compromised sites)

Results

- What are the characteristics of GR?
  - Size, Churn, Lifetime
- How effective is GR in poisoning Google?
  - We focus on *how many poisoned search results are exposed to the user*
- Longitudinal data allows us to identify long term trends
  - Monetization through scams

GR Size + Churn

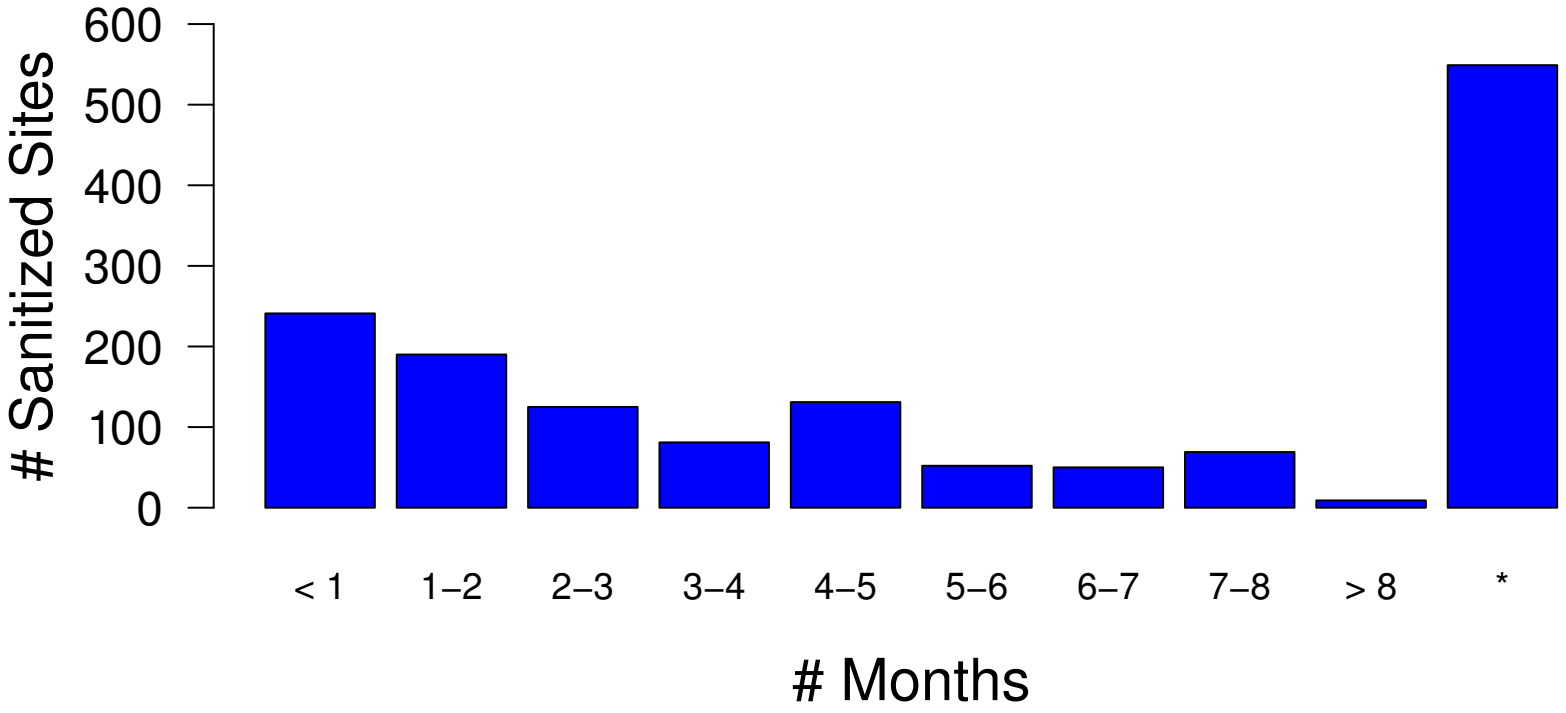


- GR is **modest in size**
- There is **little churn** amongst nodes

GR Lifetime

- We define **lifetime** as the *time between the first and last time Odwalla observed the SEO kit running on a site*
- A site is **sanitized** when it *no longer responds to the C&C protocol for 8 consecutive days*

GR Lifetime

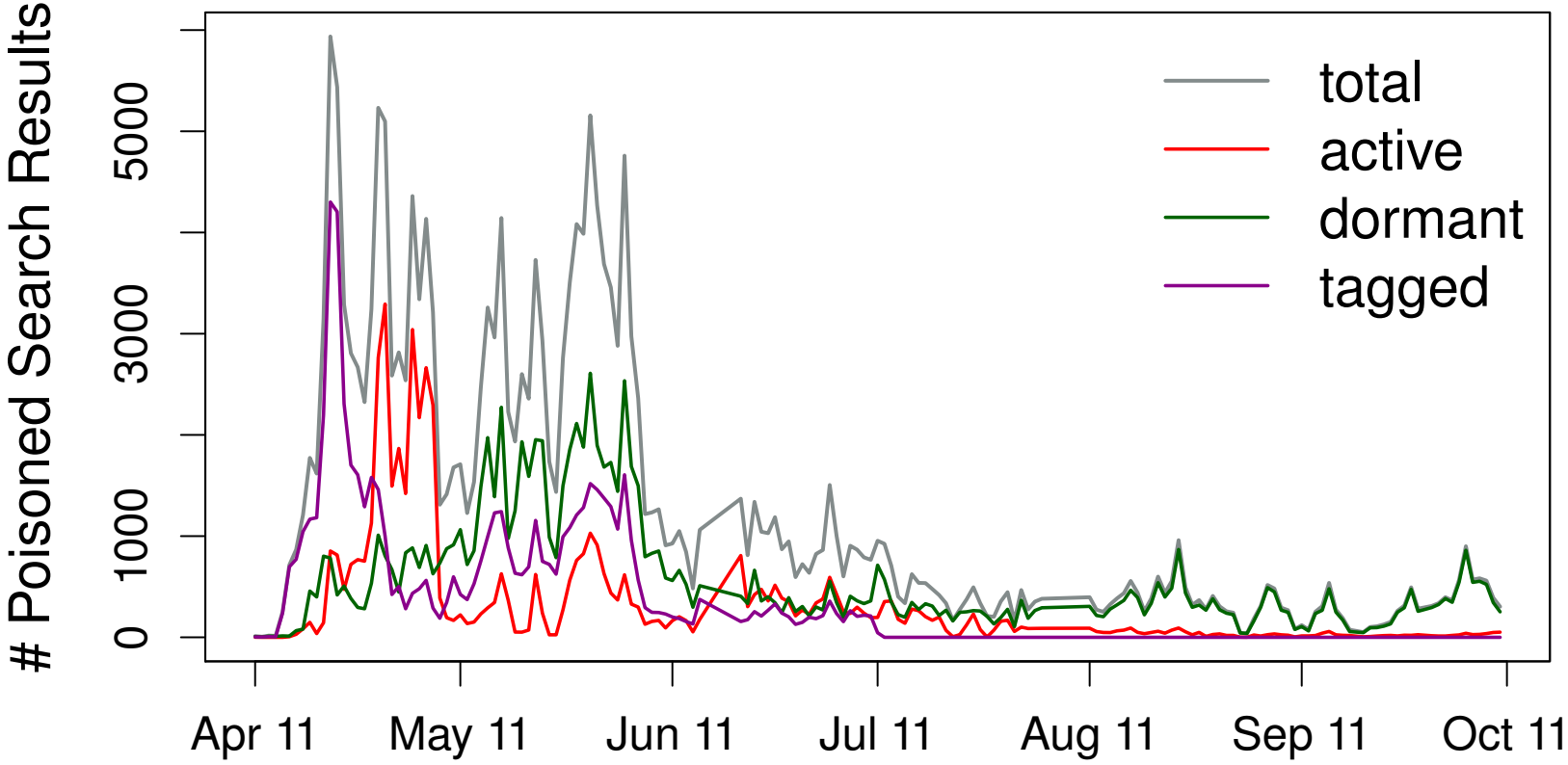


- Compromised sites are **long lived** (months at a time) and able to support GR w/ **high availability**
- SEO kits want to hide their presence from site owners

Effectiveness

- Measure **effectiveness** of GR by the *volume of poisoned search results*
- Intersect known compromised sites [Odwalla] with poisoned search results on Google [Dagger]
- Label each poisoned search result as:
  - **Active:** cloaking + redirecting users
  - **Tagged:** neutralized via Google Safe Browsing
  - **Dormant:** cloaking, but not redirecting users

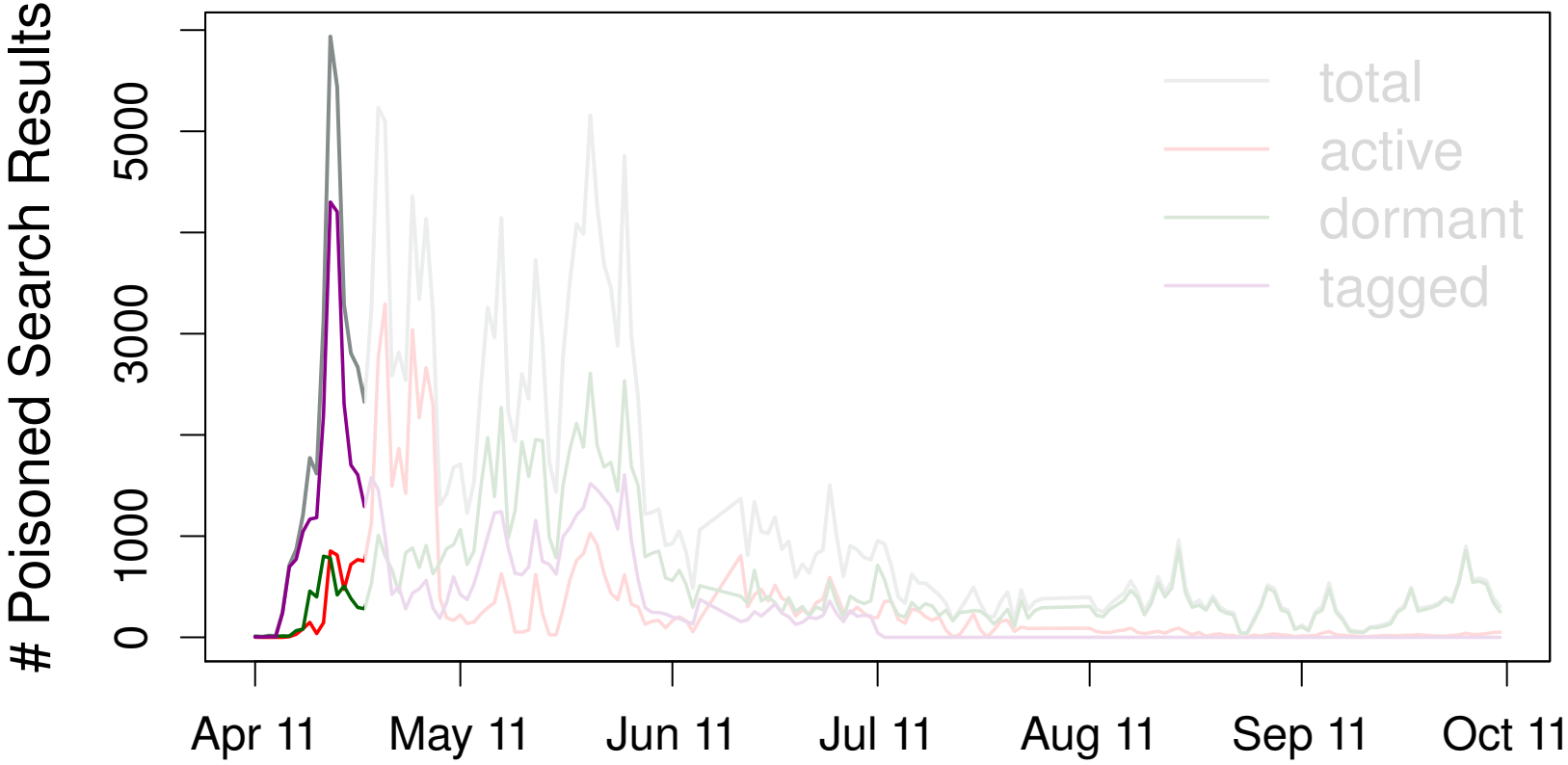
Effectiveness



- Multiple periods of activity:

  Start → Surge → Steady → Idle

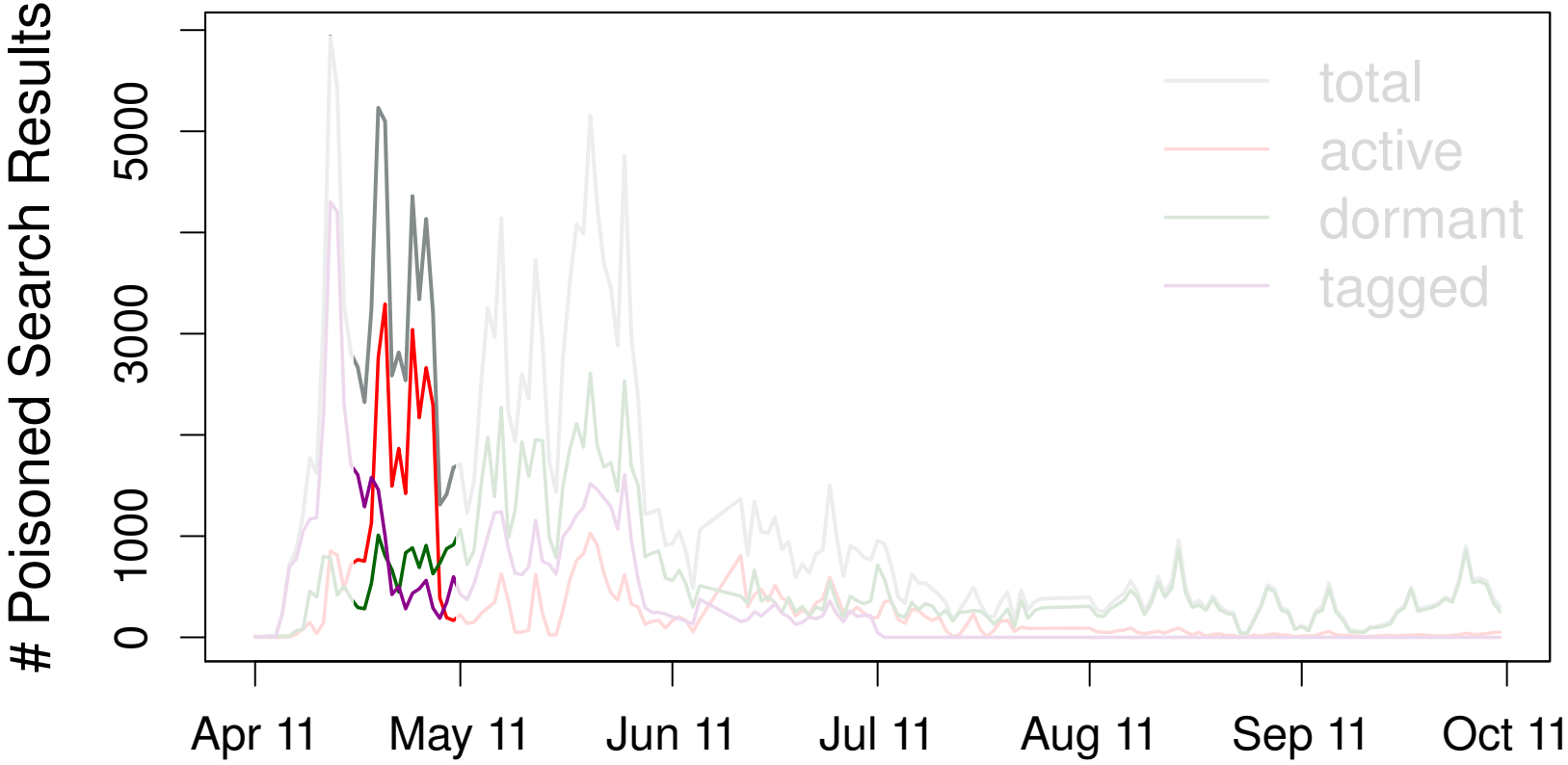
Effectiveness



Start → Surge → Steady → Idle

Mostly tagged, active ramping up

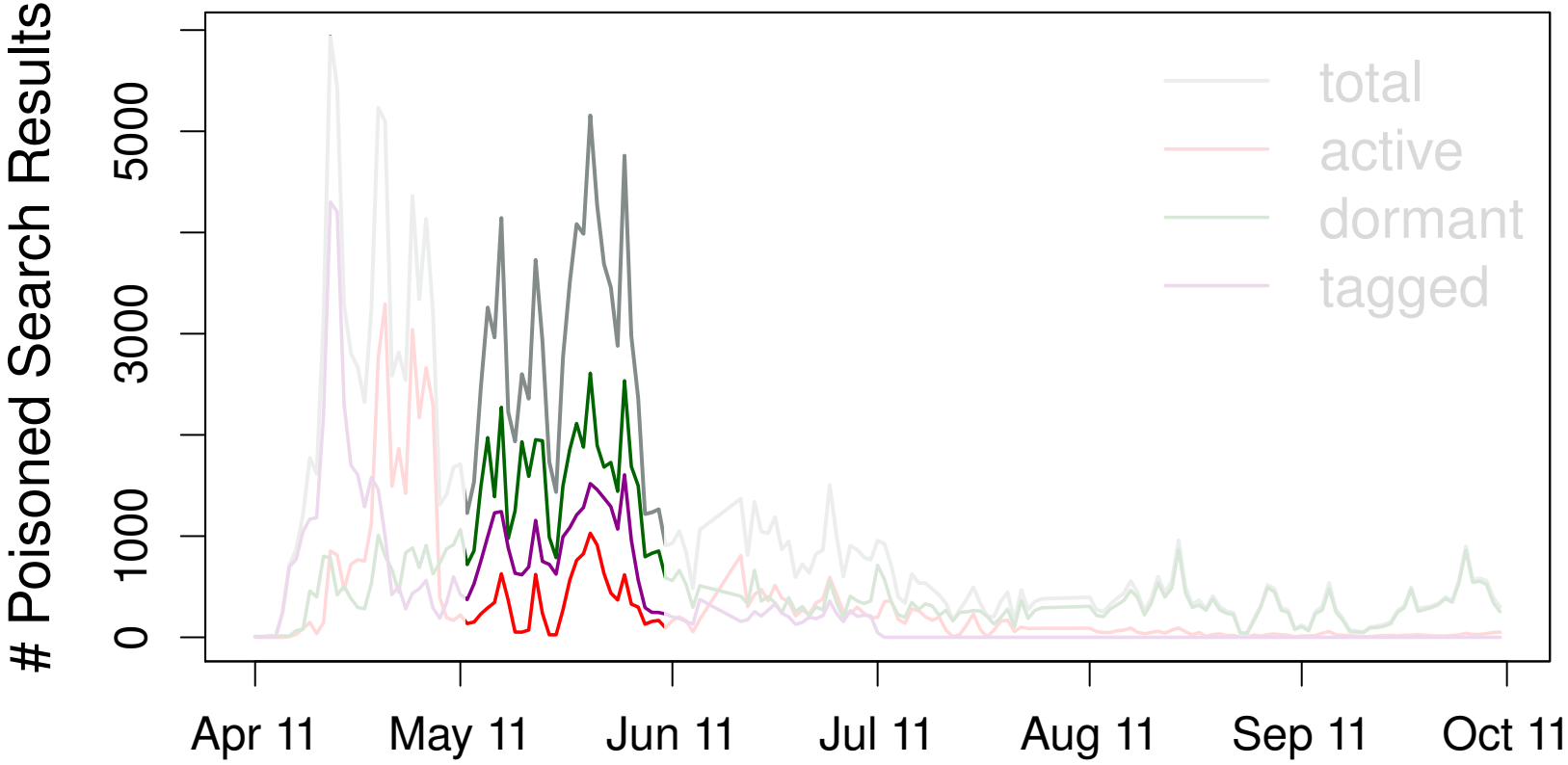
Effectiveness



Start → **Surge** → Steady → Idle

Active surges with little pressure from GSB

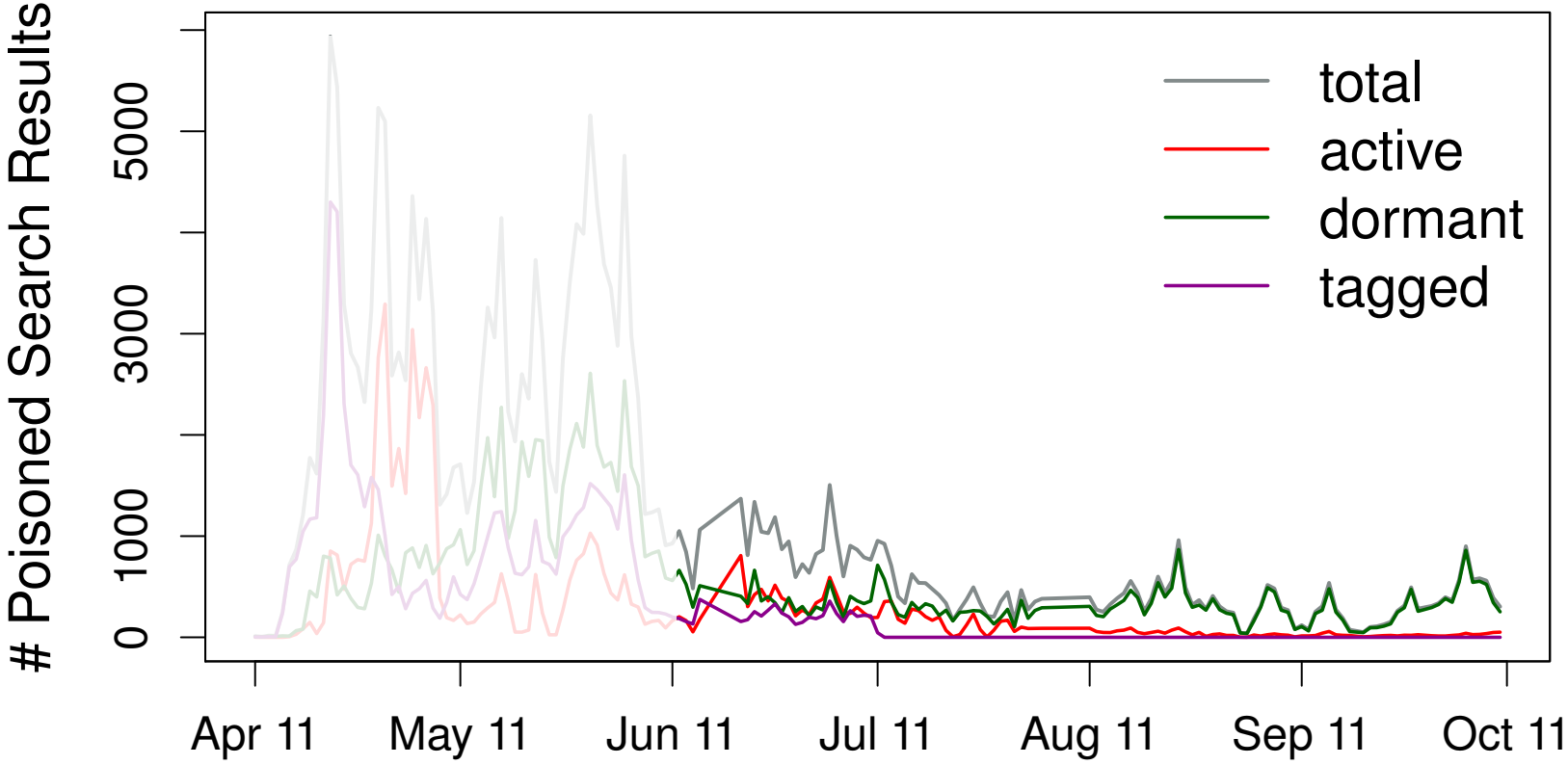
Effectiveness



Start → Surge → **Steady** → Idle

Tagged increases, but many active still present

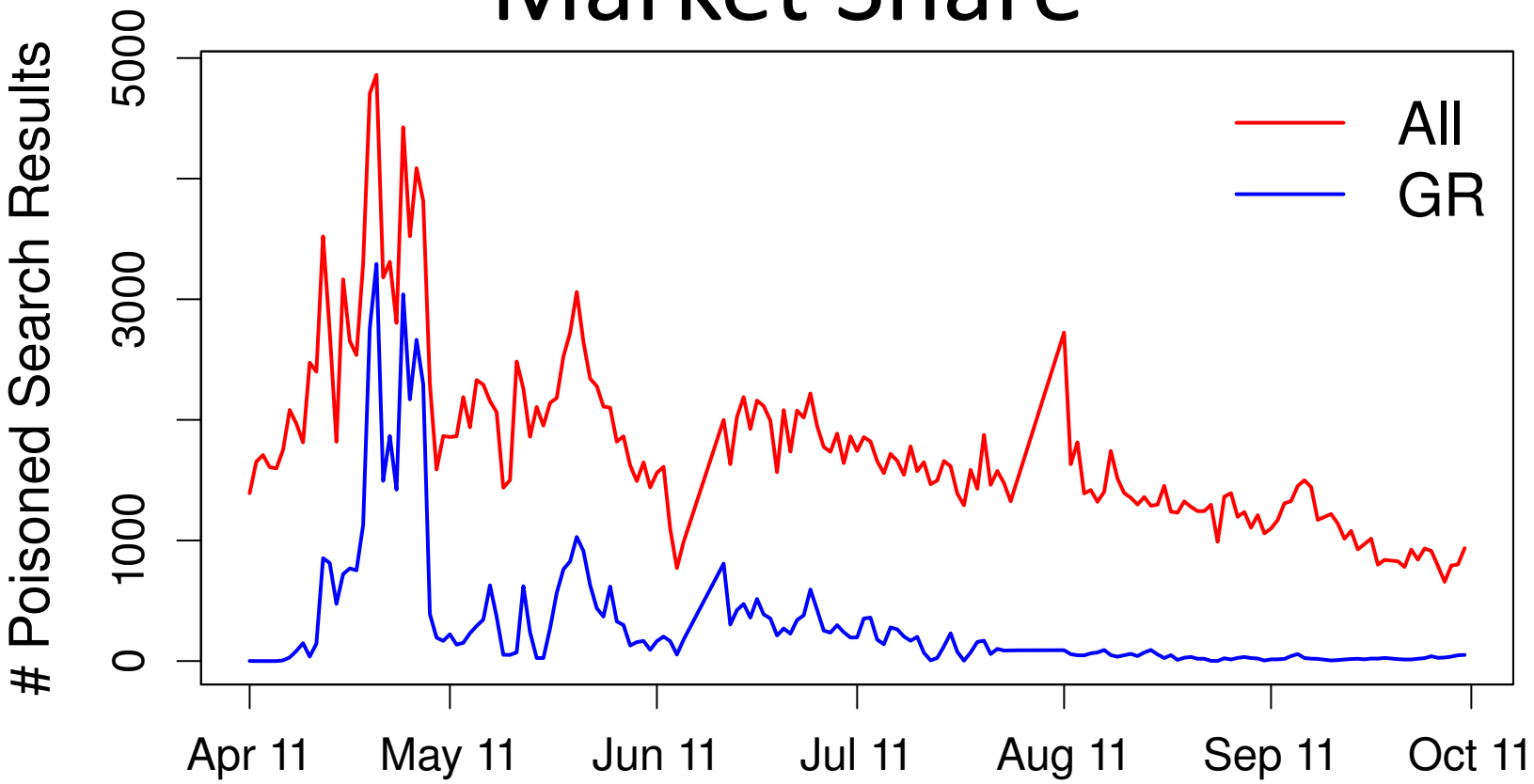
Effectiveness



Start → Surge → Steady → **Idle**

Total volume drops, lack of monetization

Market Share

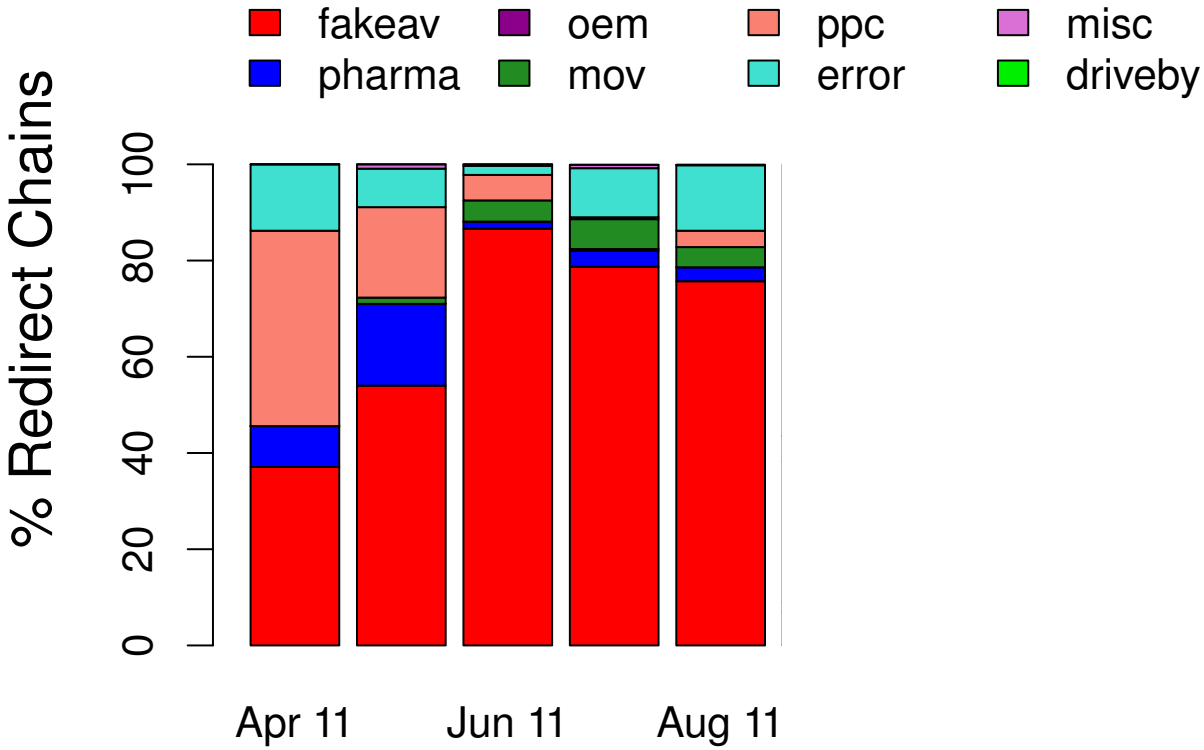


- Compare GR against all poisoned search results
- GR accounts for the **majority** of poisoned search results during the surge period (58%)

Monetization

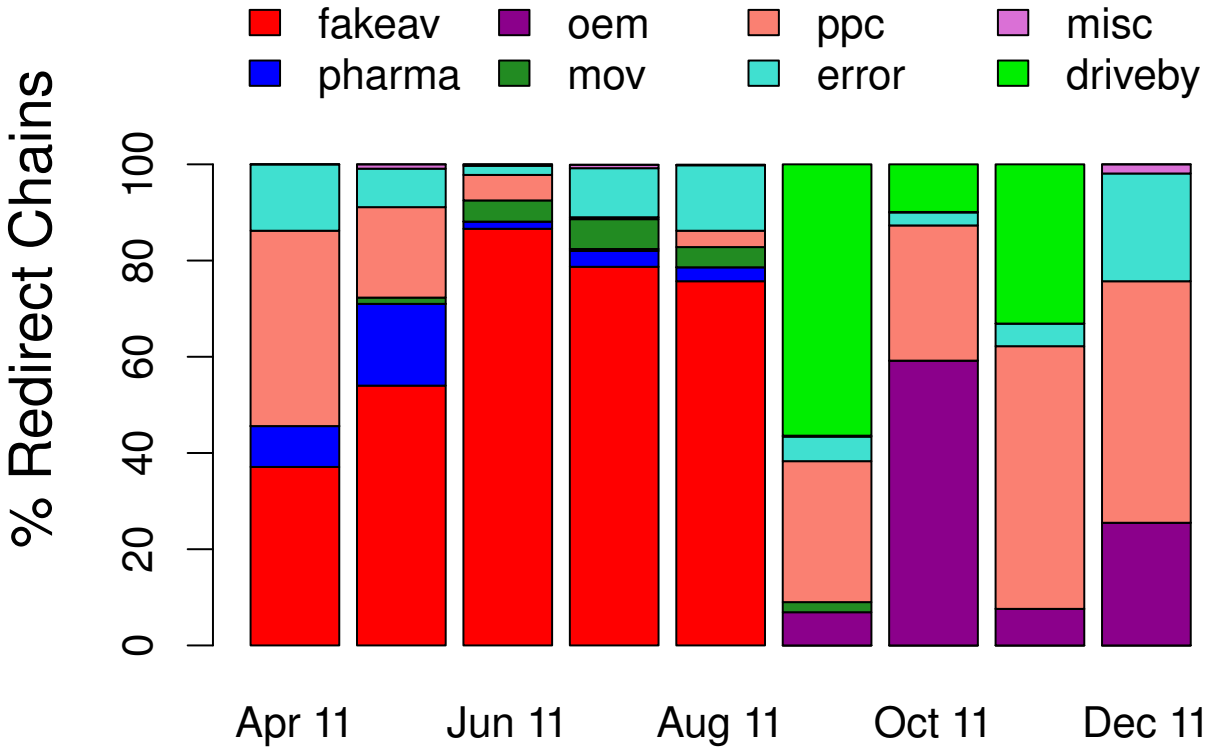
- To identify final scam from redirection data [Trajectory], we select chains:
  - Originate from GR doorway
  - Contain 1+ cross site redirect
  - Occur while mimicking MSIE
- Manually cluster + classify scams

Monetization



- Experimentation w/ affiliate programs
- Early on *Fake AV is the scam of choice*

Monetization



- FBI crackdown on Fake AV industry sent GR into flux

Conclusion

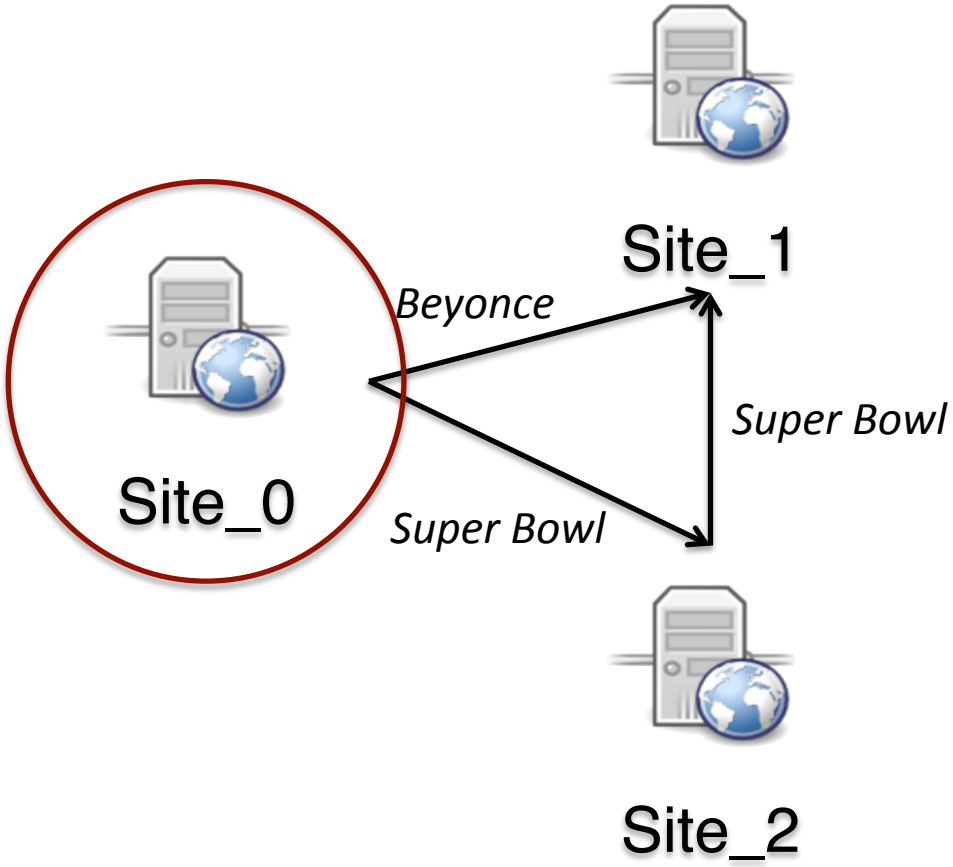
- GR is very effective at poisoning search results even with modest resources
- *Fake AV was the financial motivation* that drove innovation in GR **(the killer scam)**
- Pure technical interventions had some effect, but it was the *financial intervention that forced GR into retirement*

Thank You!

- Questions?

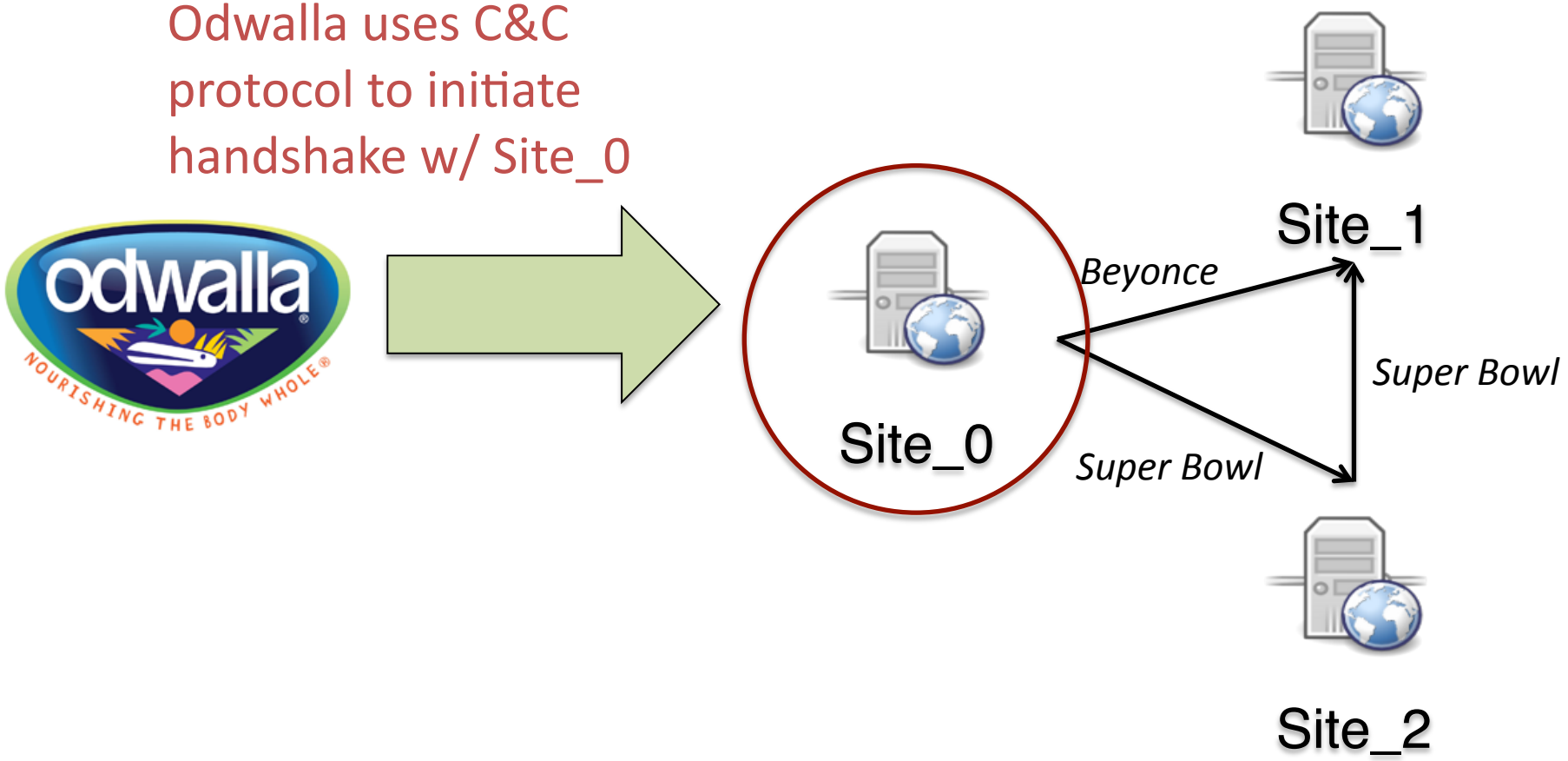
Odwalla Example

Odwalla wants to test whether Site_0 is part of GR



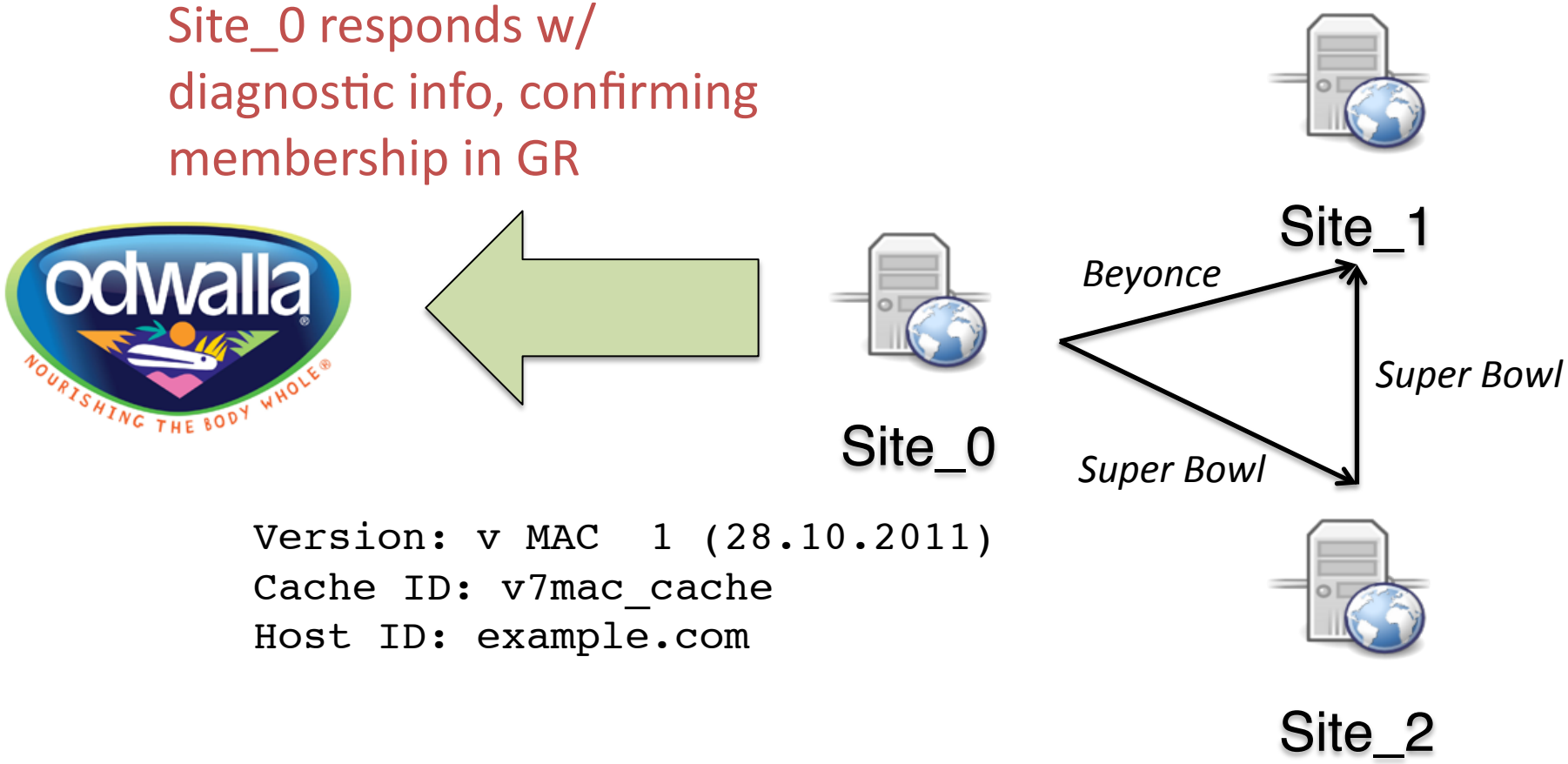
Odwalla Example

Odwalla uses C&C protocol to initiate handshake w/ Site_0



Odwalla Example

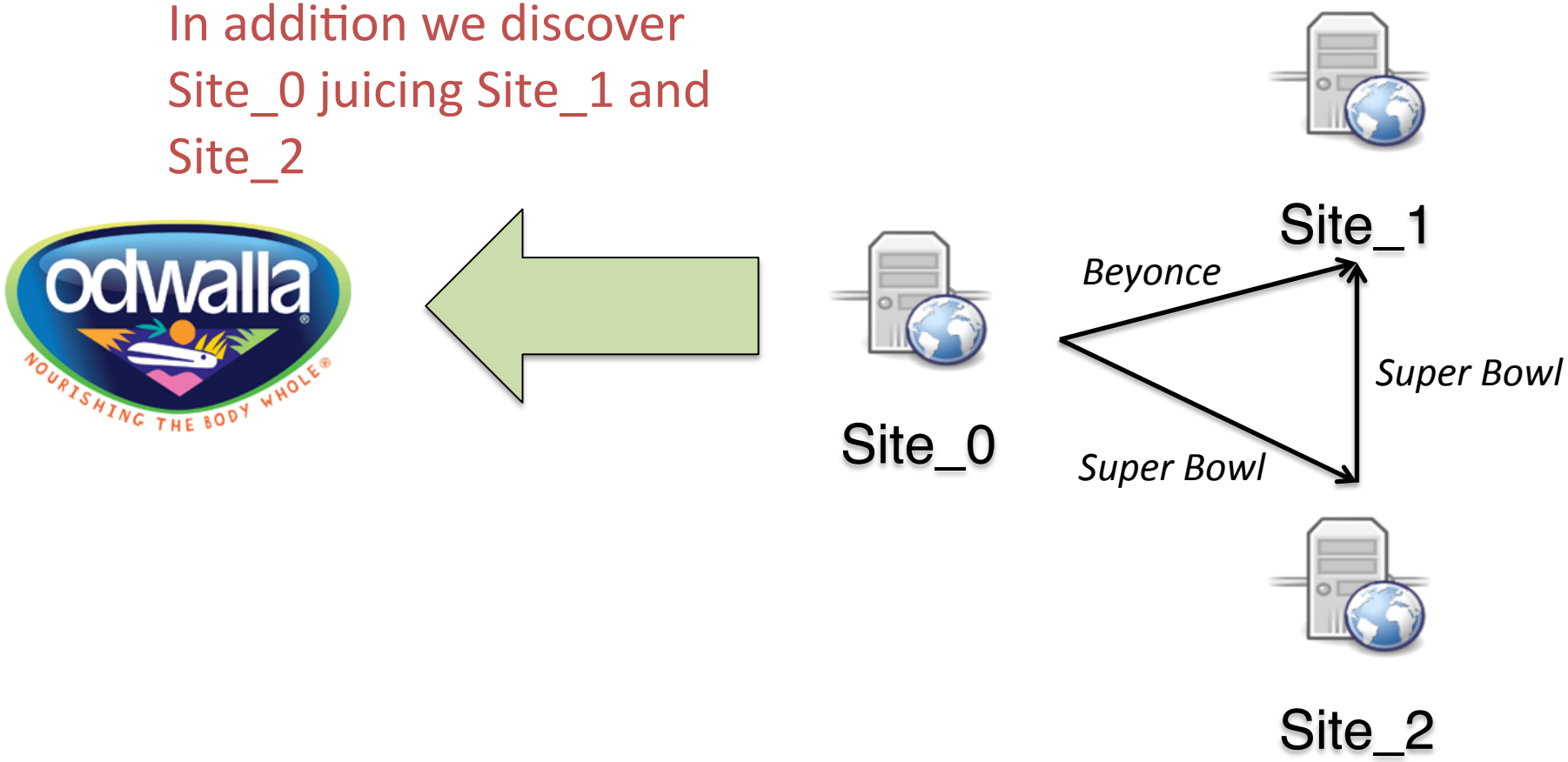
Site_0 responds w/ diagnostic info, confirming membership in GR




```
Version: v MAC  1 (28.10.2011)
Cache ID: v7mac_cache
Host ID: example.com
```

Odwalla Example

In addition we discover Site_0 juicing Site_1 and Site_2



Odwalla Example

Odwalla tests Site_1 and Site_2