POLITECNICO
DI MILANO

Modern Botnets

and the Rise of Automatically Generated Domains

Joint work with

Stefano Schiavoni (POLIMI & Google, MSc),

Edoardo Colombo (POLIMI)

Lorenzo Cavallaro (RHUL, PhD),

Stefano Zanero (POLIMI, PhD)

Federico Maggi

federico@maggi.cc

Politecnico di Milano

Who I am

POLITECNICO
DI MILANO

Federico Maggi, PhD

Post-doctoral Researcher

Topics

Android malware, malware analysis, web measurements

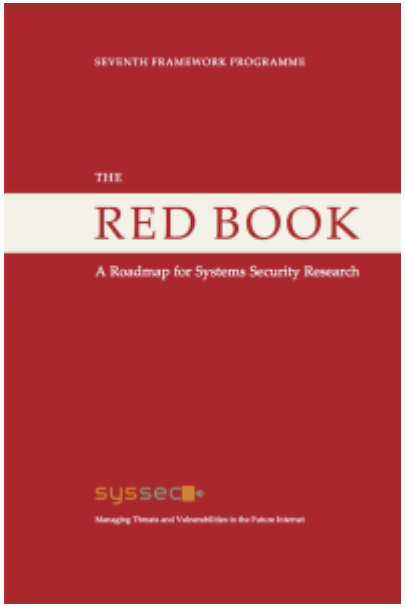
Background

Intrusion detection, anomaly detection

www.red-book.eu

The RED BOOK

A Roadmap for Systems Security Research



Free PDF

Audience

Policy makers

Researchers

Journalists

Content

Vulnerabilities

Social Networks

Critical Infrastructure

Mobile Devices

Malware

Roadmap

1. Botnets
2. Communication channels
3. Domain generation algorithms (DGAs)
4. Detecting DGA-based botnets
5. Results

Roadmap

1. **Botnets**
2. Communication channels
3. Domain generation algorithms (DGAs)
4. Detecting DGA-based botnets
5. Results

Botnets: from malware to service

Botnet

- Network of (malware infected) computers
- Controlled by an external entity (e.g., cybercriminal)

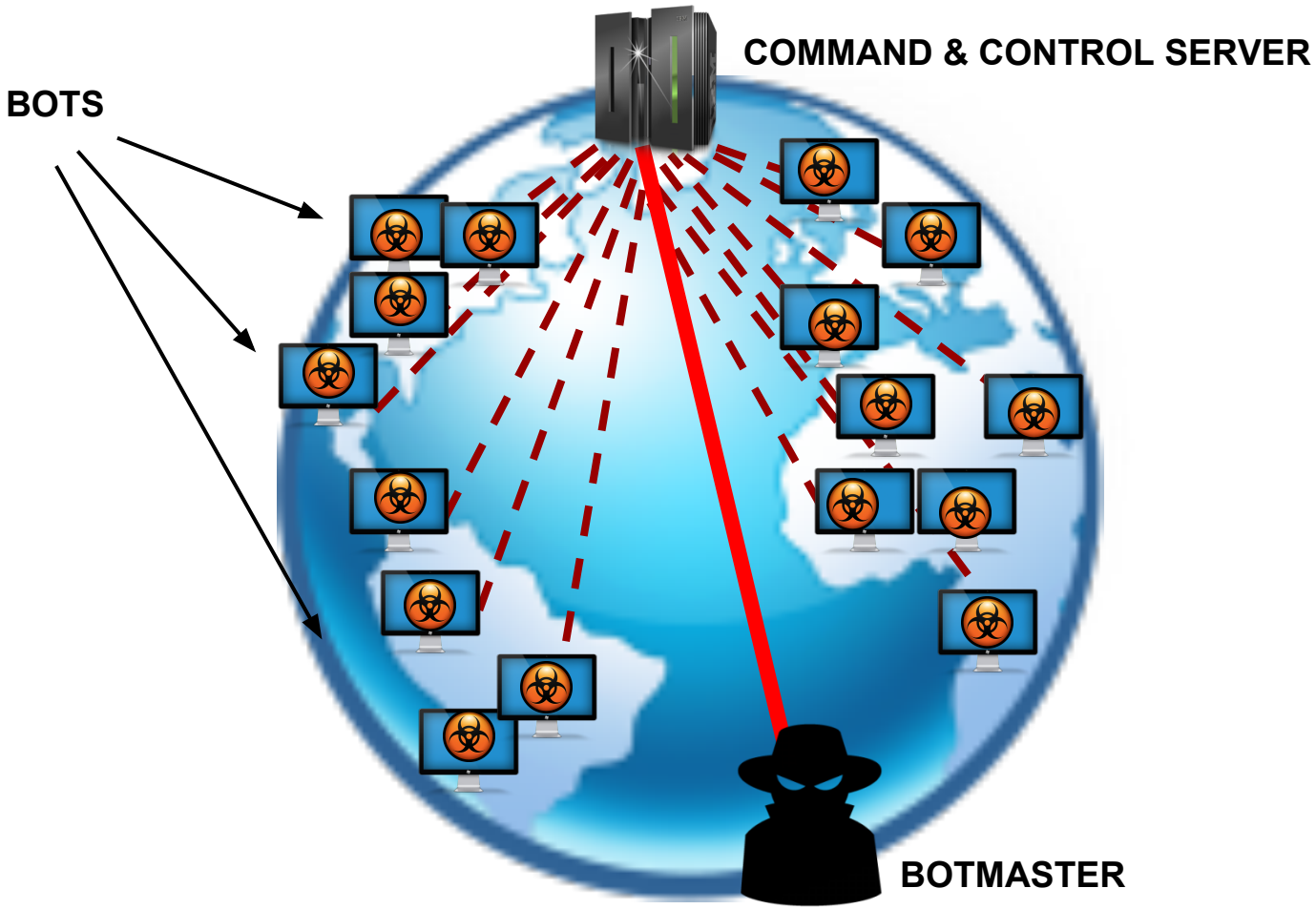
Bot

- Computer member of a botnet
- Infected with malicious software

Botmaster

- Person or group managing the botnet

Centralized topology example



Infected machines = $$$

Steal sensitive information

- harvest contacts
- online banking credentials

Run malicious activities

- send spam, phishing emails, click fraud
- denial of service

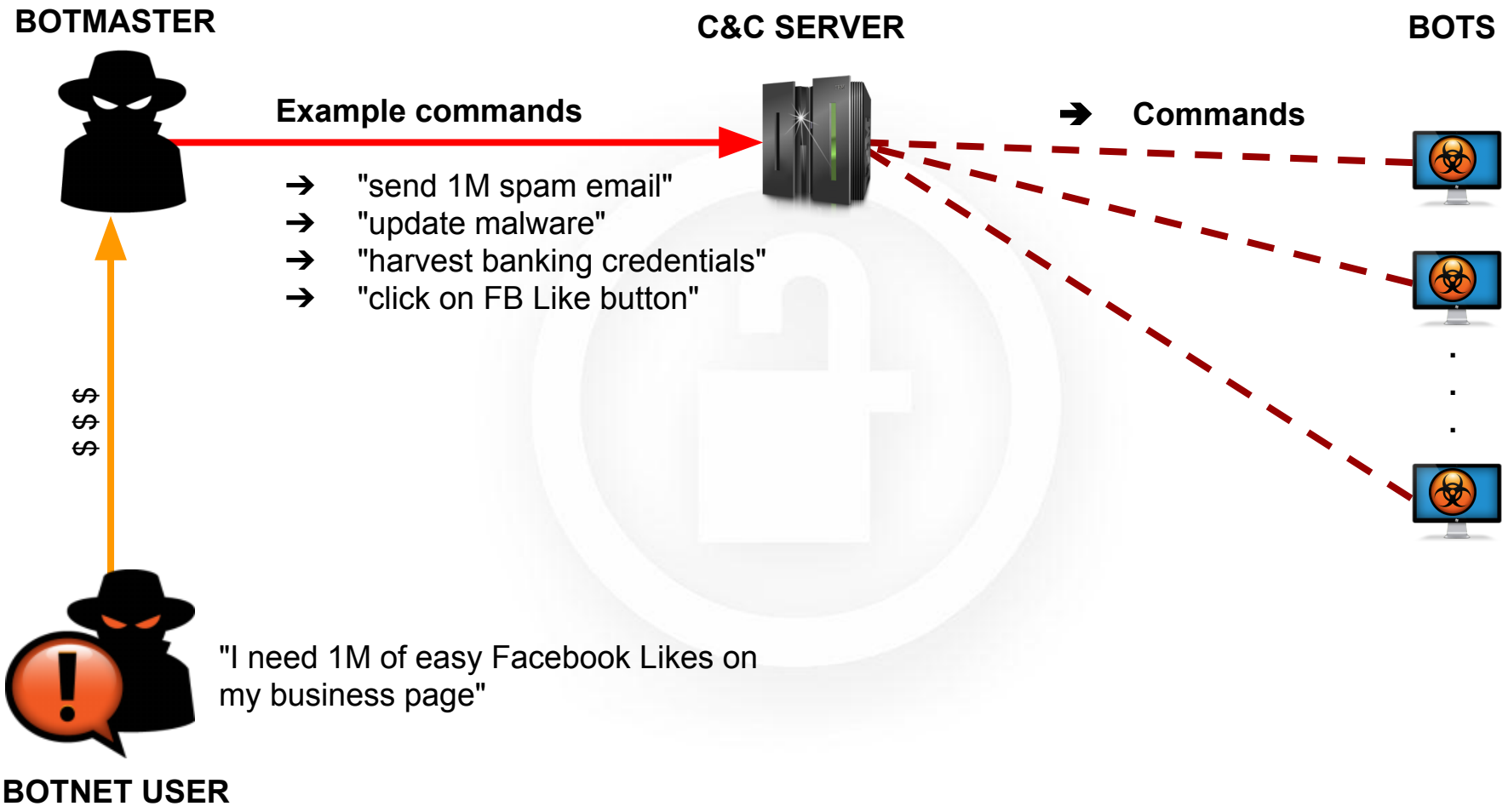
Make money

- rent the infrastructure as a service

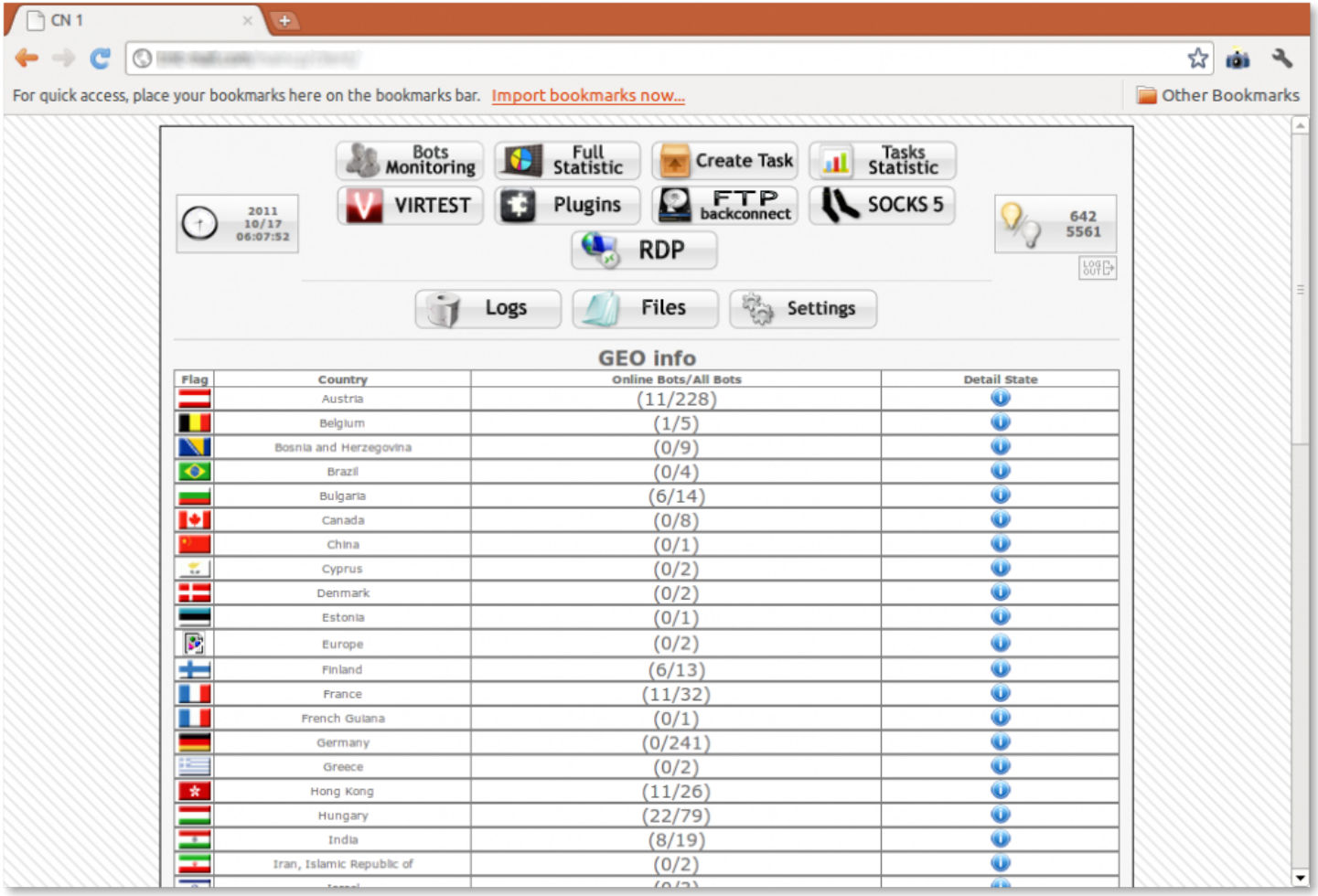
Maintenance

- update the malware

Command & control flow



Administration dashboard (spyeye)



Source (webroot.com)

Some notable examples

Flashback (2012–today)

- 600K compromised Macs (so, it's not just Windows)
- credentials stealing

Grum (2008–2012)

- 840K compromised devices,
- 40bln/mo spam emails

TDL-4 (2011–today)

- 4,5M compromised machines (first 3 months)
- known as "indestructible".

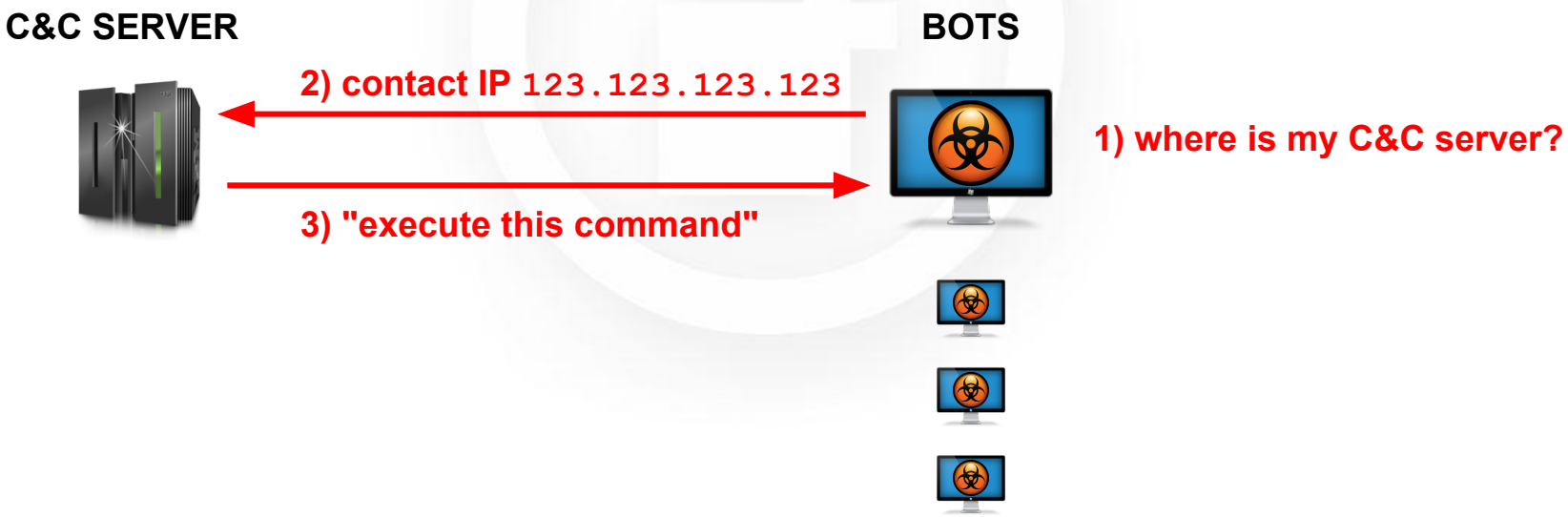
Cryptolocker (October 2013–today) NEW

Roadmap

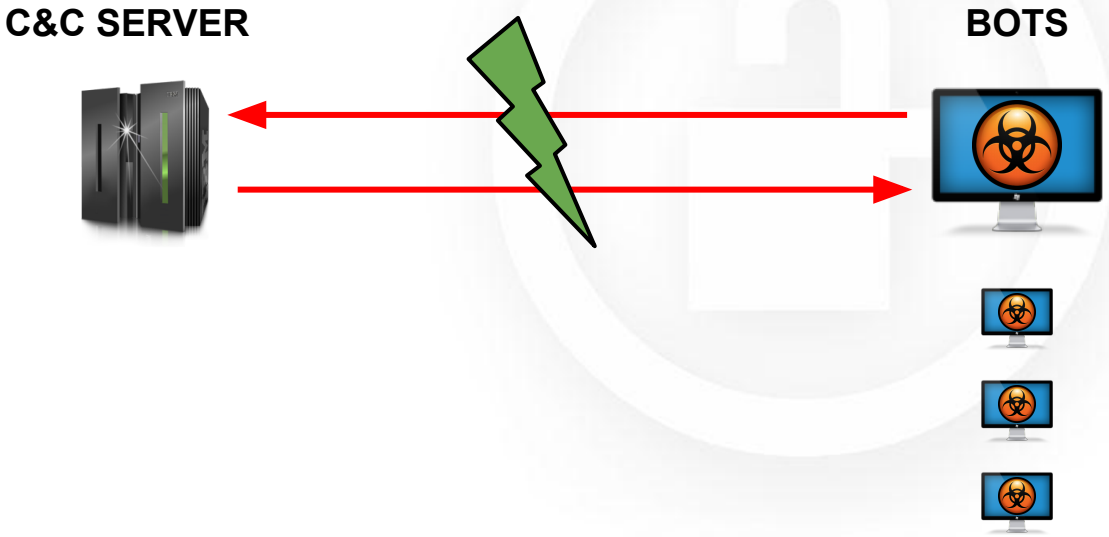
1. Botnets
2. **Communication channels**
3. Domain generation algorithms (DGAs)
4. Detecting DGA-based botnets
5. Results

Where is the my C&C server?

1. Where is my C&C server located?
2. Contact the C&C server
3. Receive command



C&C channel: single point of failure



P2P is the natural answer.

We focus on **centralized botnets** because they're still a **majority.**

Centralized C&C mechanisms

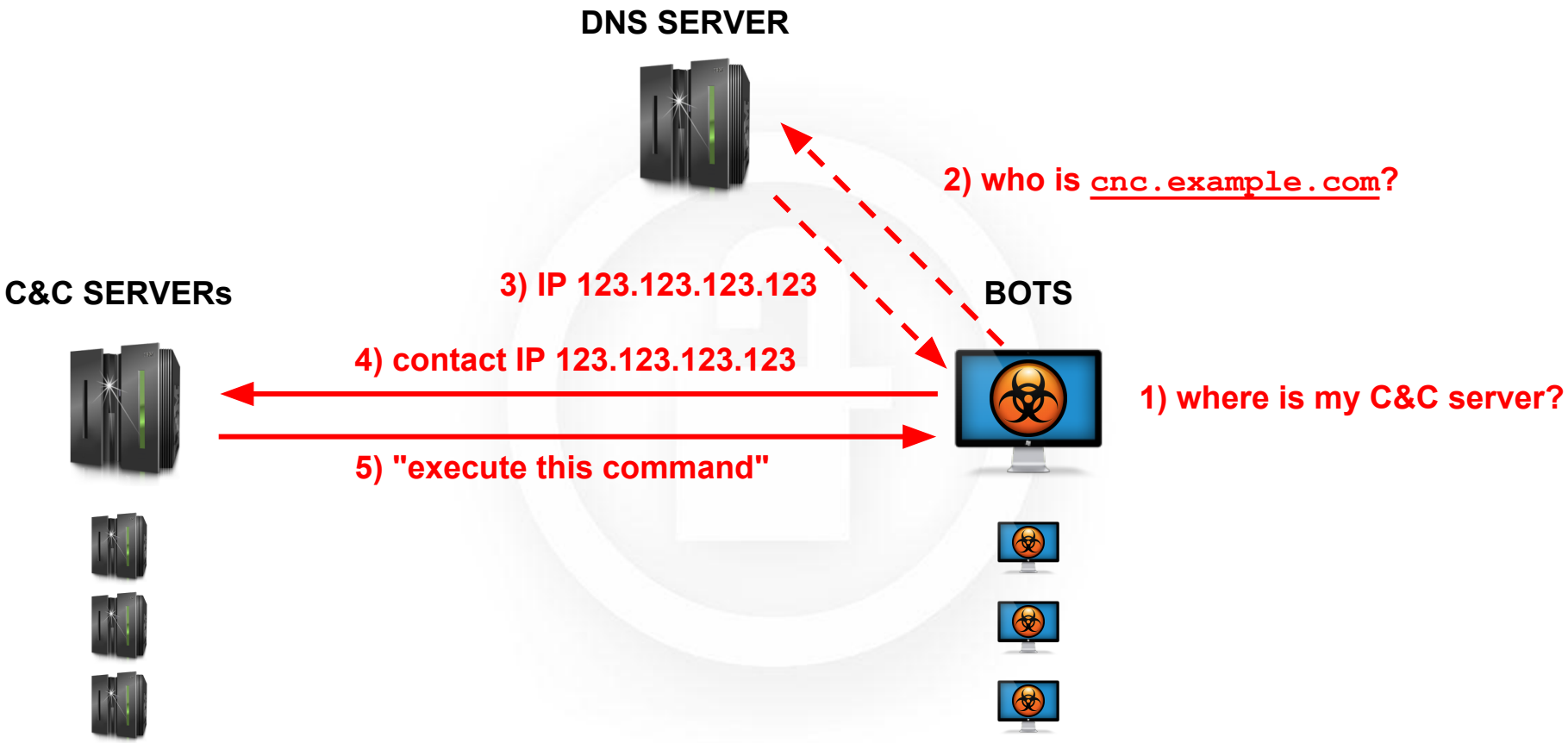
Hardcoded IPs (e.g., 123.123.123.123)

- Bot software (malware) ships with the IPs
- Botmaster can update IPs regularly
- Knowing the IP makes takedown easy

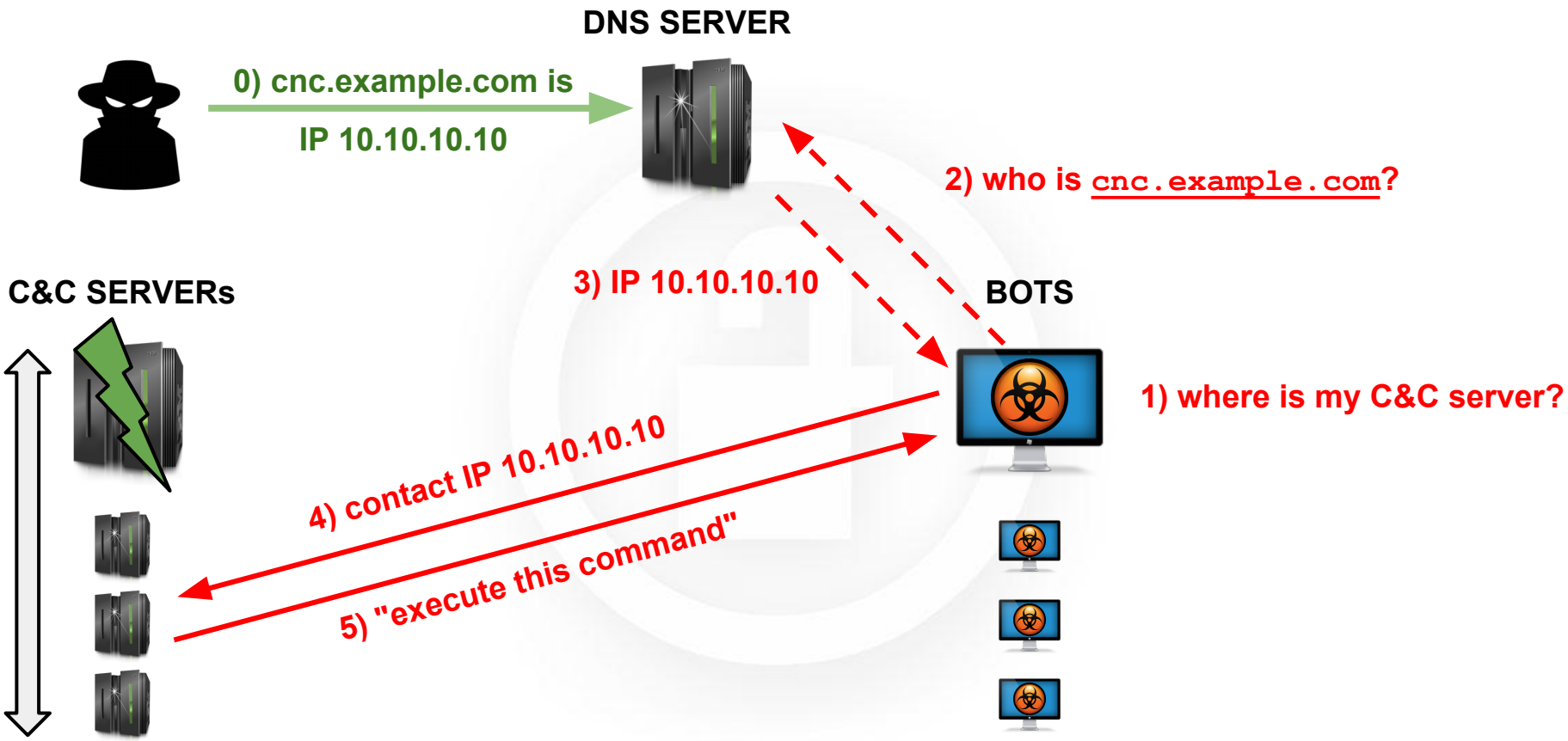
Hardcoded domain names (e.g., cnc.example.com)

- Decouple IP from domain
- Botmaster free to change domain names and IPs
- Frequently changing IPs make takedown harder
- Botmaster must own many IPs

Hardcoded domain names (2)



Hardcoded domain names (1)



Roadmap

1. Botnets
2. Communication channels
3. **Domain generation algorithms (DGAs)**
4. Detecting DGA-based botnets
5. Results

Game-changing approach

Goals of the botmaster

- Make the C&C server **harder to locate**
- Make the C&C channel **resilient to hijacking**

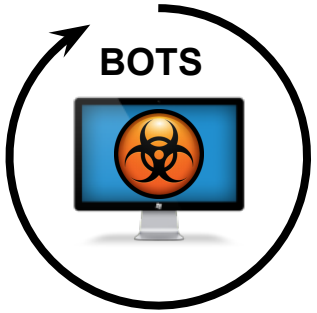
Reversing the malware binary should not reveal the location of the C&C nor any useful information toward that.

Single domain vs. Domain flux

cnc.example.com

vljiic.org
f0938772fb.co.cc
jyzirvf.info
hughfgh142.tk
fyivbrl3b0dyf.cn
vitgyyizzz.biz
nlgie.org
aawrqv.biz

yxipat.cn
rboed.info
79ec8f57ef.cc
gkeqr.org
xtknjczaafo.biz
yxzje.info
ukujhjg11.tk
...



SINGLE DOMAIN
predictable
easy to leak

THOUSANDS OF DOMAINS PER DAY
unpredictable
impossible to leak

Domain of the day

BOTMASTER

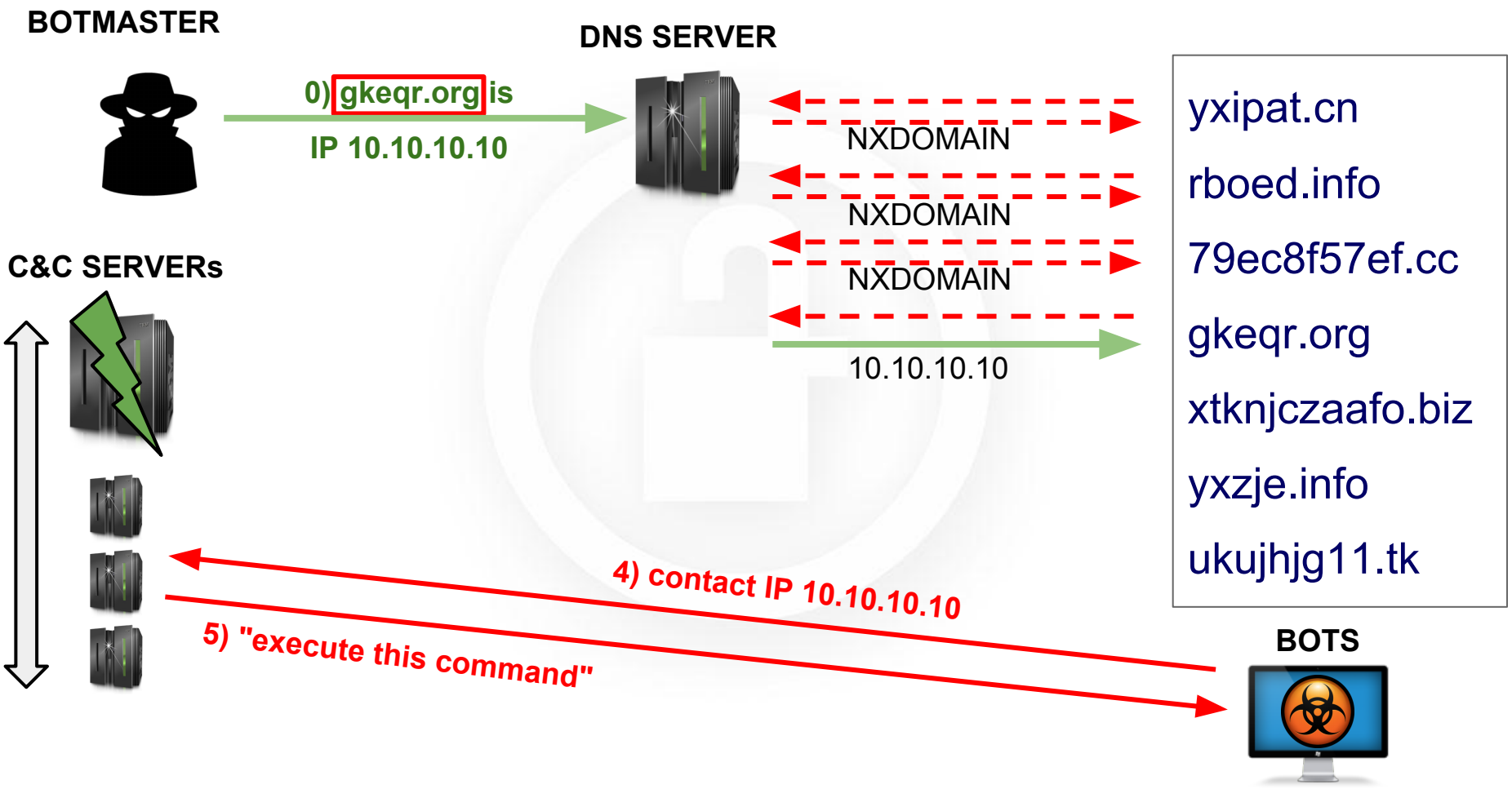
Domain of the day

Register only one domain every day (week) that resolve to the true IP of the C&C

vljiic.org	yxipat.cn
f0938772fb.co.cc	rboed.info
jyzirvf.info	79ec8f57ef.cc
hughfgh142.tk	gkeqr.org
fyivbrl3b0dyf.cn	xtknjczaafo.biz
vitgyyizzz.biz	yxzje.info
nlgie.org	ukujhjg11.tk
aawrqv.biz	...

THOUSANDS OF DOMAINS PER DAY
unpredictable
impossible to leak

Where is my C&C server?



Leveraging DNS

- Only the botmaster knows the **active domain**
- The **DNS** protocol does the rest
- The **DGA** can be made more **unpredictable** (e.g., Twitter trending topic)

Reversing the malware binary

only reveals the **generation algorithm**

not the active domain of the day!

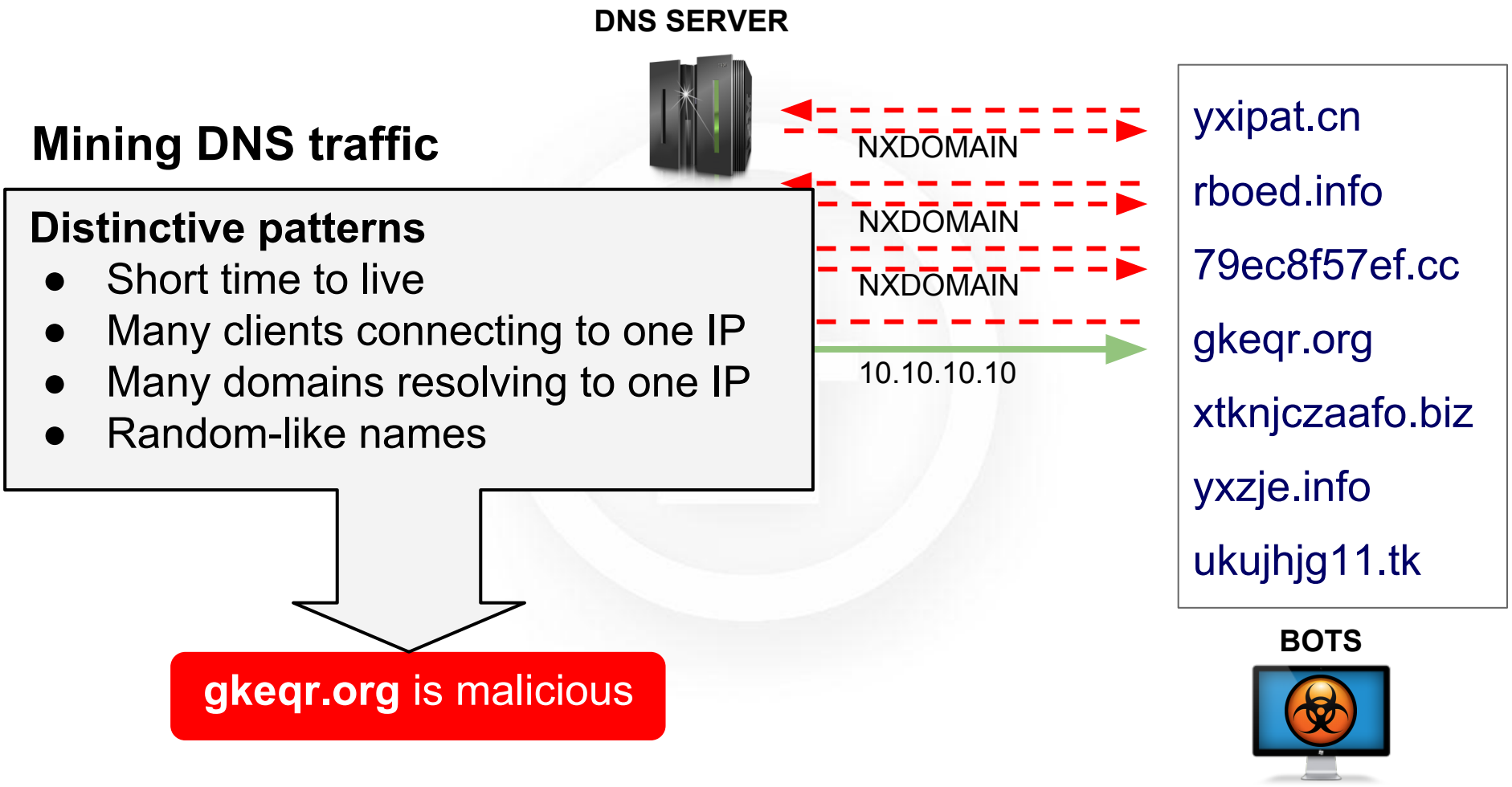
Message in a bottle

(Source)

Roadmap

1. Botnets
2. Communication channels
3. Domain generation algorithms (DGAs)
4. **Detecting DGA-based botnets**
5. Results

Natural observation point: DNS



Domain reputation systems

Notos

- [Antonakakis et al., 2010]

KOPIS

- [Antonakakis et al., 2011]

EXPOSURE

- [Bilge et al., 2011]
- http://exposure.iseclab.org

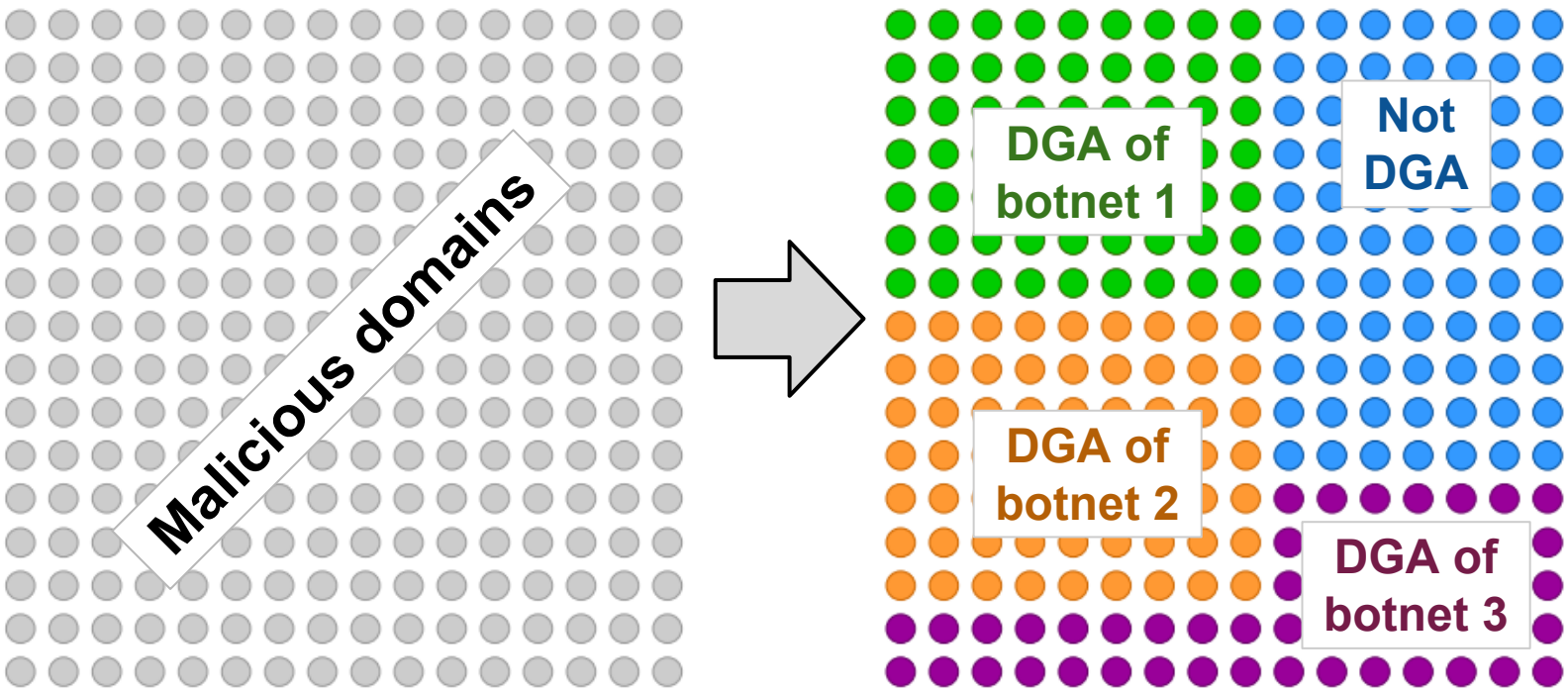
Drawbacks

They tell malicious vs. benign domains apart

No insights on what is the purpose of the domain

- C&C of what botnet?
- Could the same C&C be used for multiple botnets?
- Is the domain malicious for other reasons?
  - Phishing
  - Spam
  - Drive-by download

More insights needed

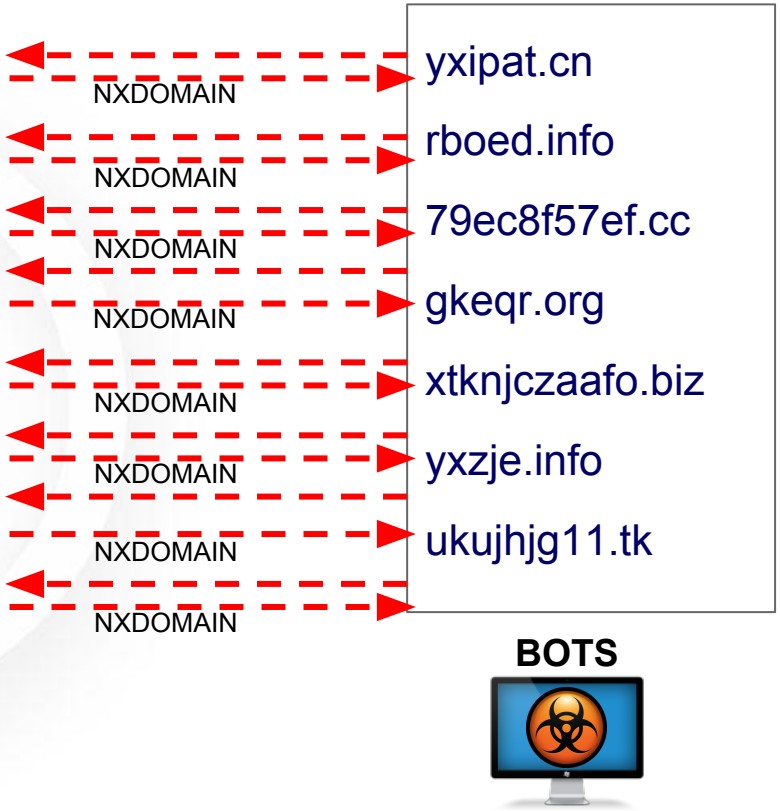


NXDOMAINs

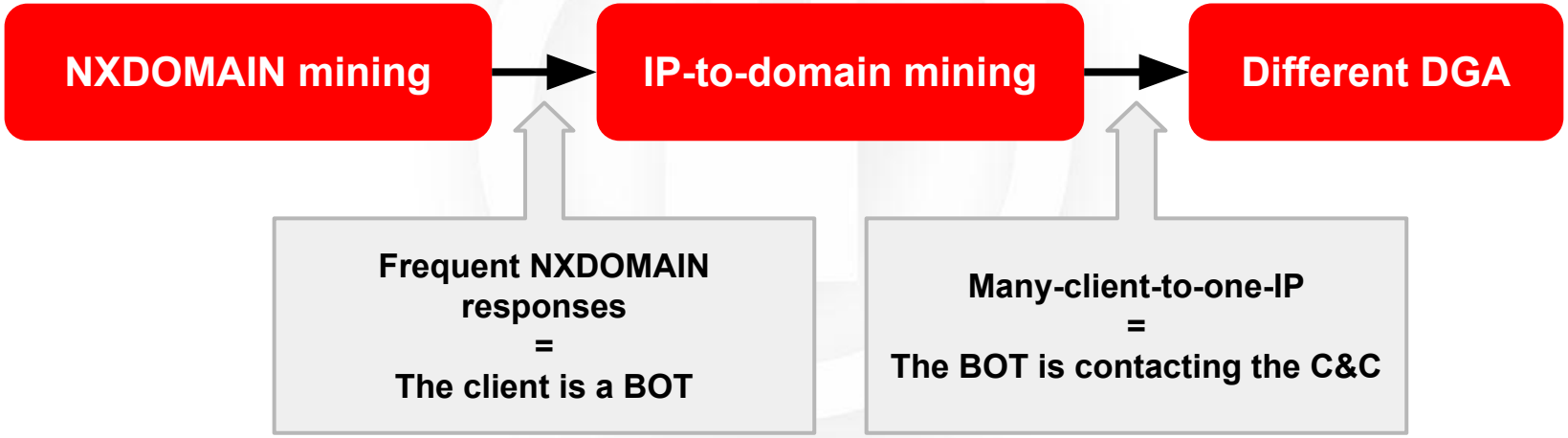
Infected clients try many domains

Many NXDOMAIN responses

Distinctive pattern of DGA



Finding distinct DGAs



Drawbacks

Needs an unpractical observation point

- No global view
- Hard to deploy

Needs the IP of the clients

- Privacy of the clients is not enforced

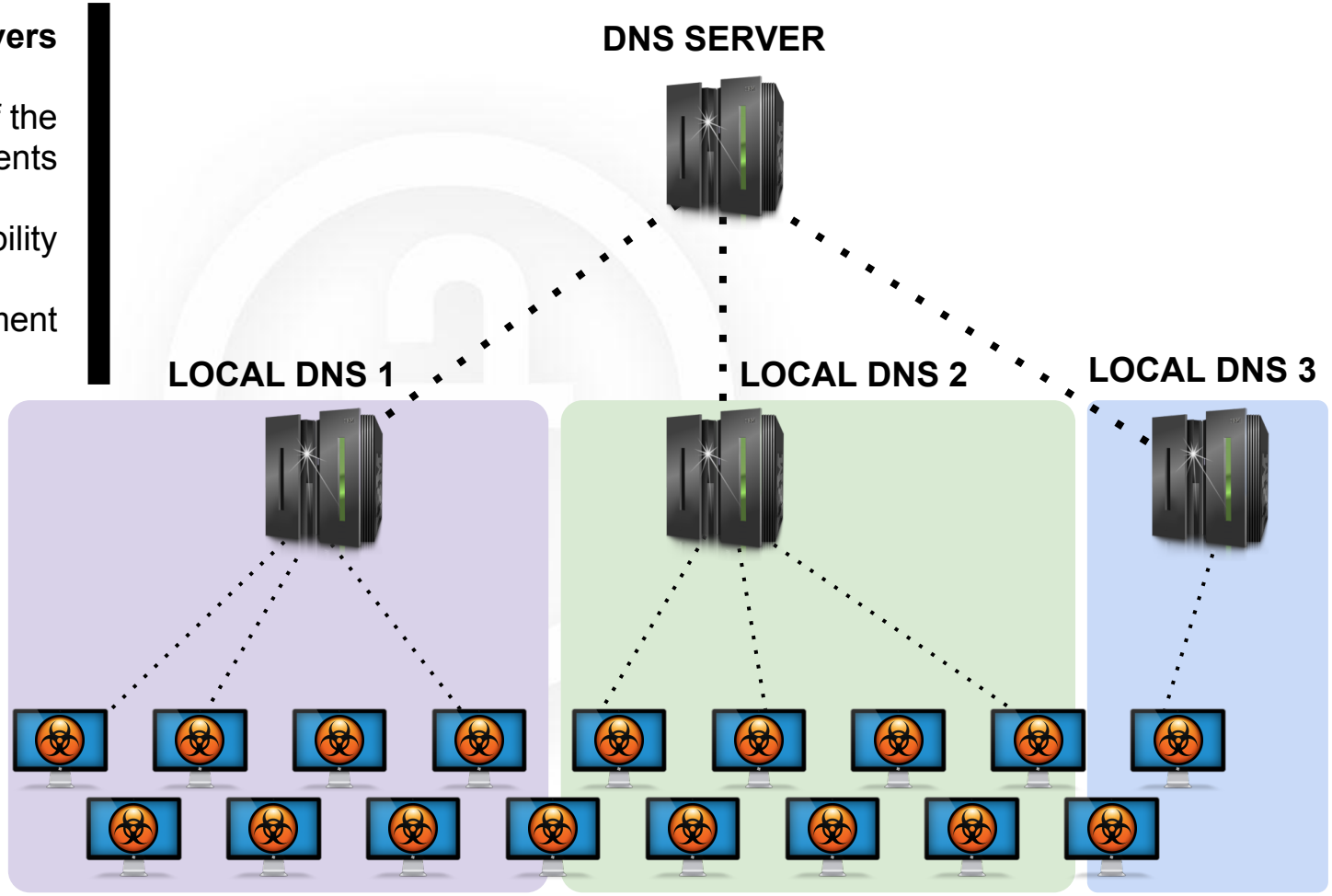
Lower level DNS servers

Middle-level resolvers

- No visibility of the querying clients
- Global visibility
- Ease of deployment

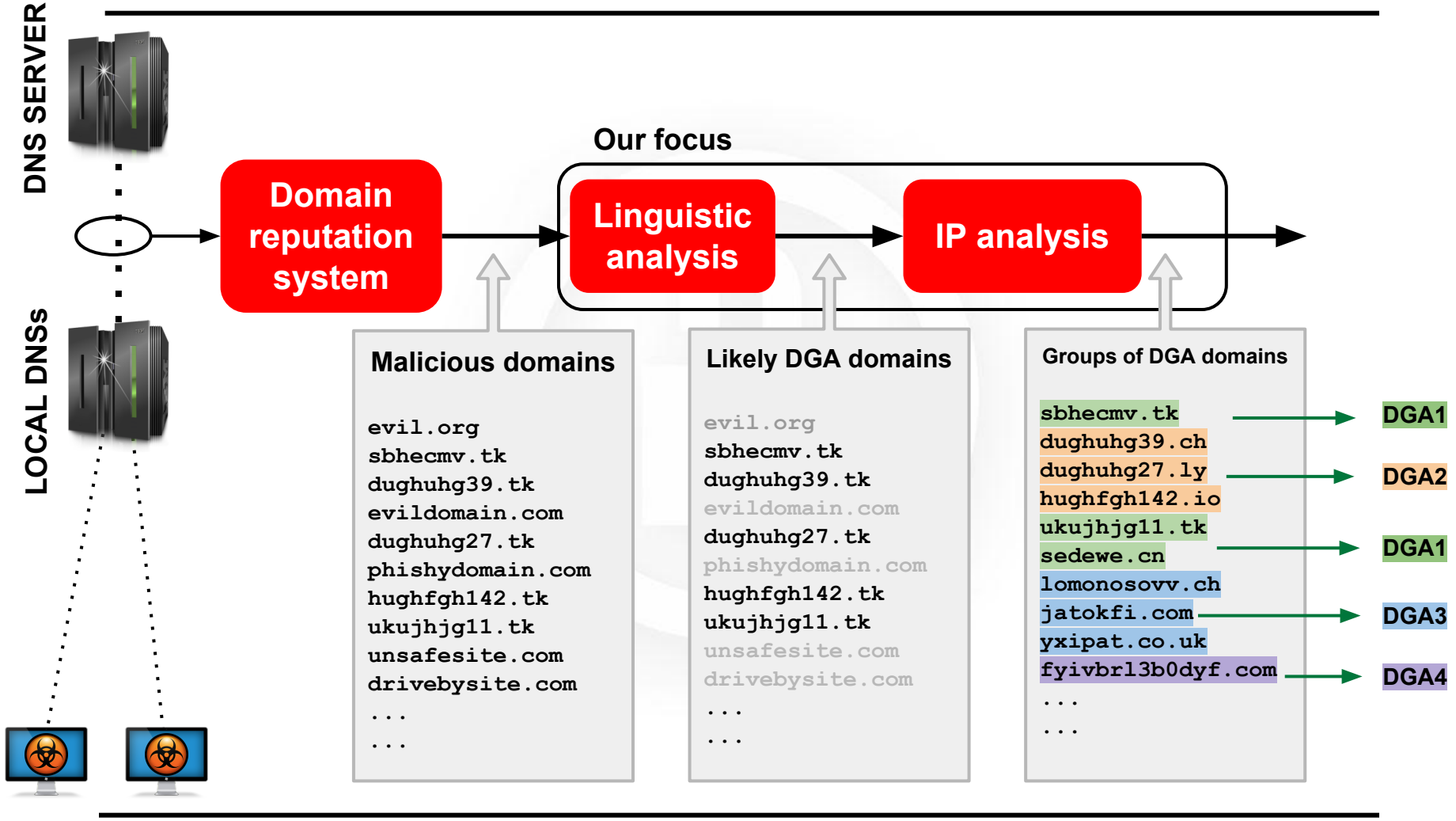
Low-level resolvers

- Visibility of the querying clients
- Local visibility
- Not easy to deploy



OUR SOLUTION

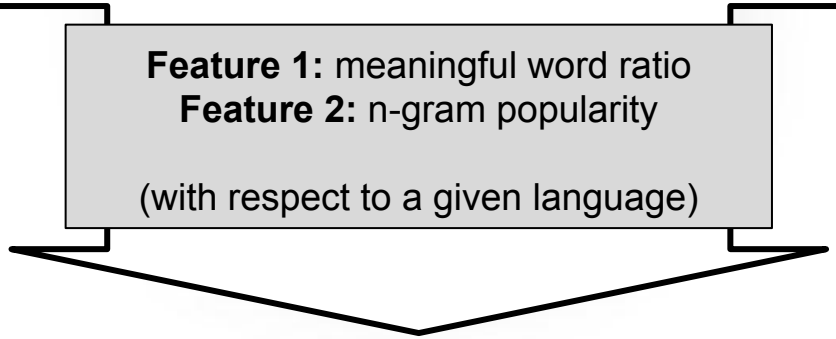
Overview of our solution



Step 1: Linguistic analysis

We measure the **"randomness"** of the strings with respect to non-DGA-generated domains

`malicious.cn`	`fyivbrl3b0dyf.cn`	`yxipat.cn`
`f0938772fb.co.cc`	`evildomain.com`	`evilrot.org`
`jyzirvf.info`	`nlgie.org`	`gkeqr.org`
`hughfgh142.tk`	`aawrqv.biz`	`xtknjczaafo.biz`



Likely non-DGA-generated

Likely DGA-generated

jyzirvf.info	nlgie.org	gkeqr.org
hughfgh142.tk	aawrqv.biz	xtknjczaafo.biz

Feature 1: meaningful word ratio
Feature 2: n-gram popularity

(with respect to a given language)

Likely non-DGA-generated

Likely DGA-generated

Feature 1

HIGH $1 = \frac{4+6}{10} = \frac{|\text{`evil`}| + |\text{`domain`}|}{|\text{`evildomain`}|} = LF1 = \frac{|word_1| + \cdots + |word_N|}{|\text{domainname}|} = \frac{|\text{`pat`}|}{|\text{`yxipat`}|} = \frac{3}{6} = 0.5$ LOW

Feature 2 (n = 2)

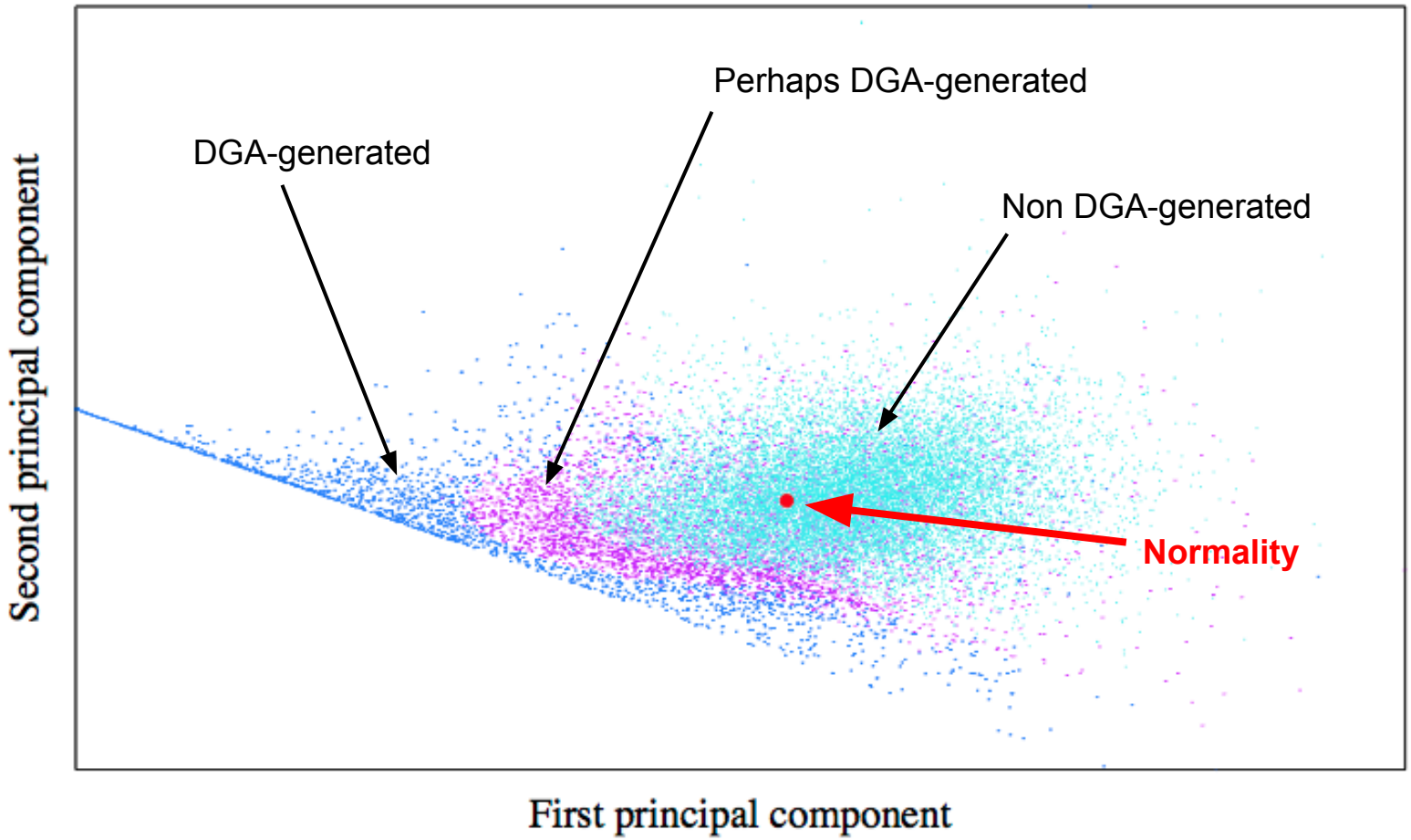
HIGH $= \text{`ev`} + \text{`vi`} + \cdots + \text{`ai`} + \text{`in`} = LF2 = \sum_i \text{popularity}(\text{n-gram}_i) = \text{`yx`} + \cdots + \text{`at`} =$ LOW

HIGH — **Feature 3 (n = 3)** — LOW

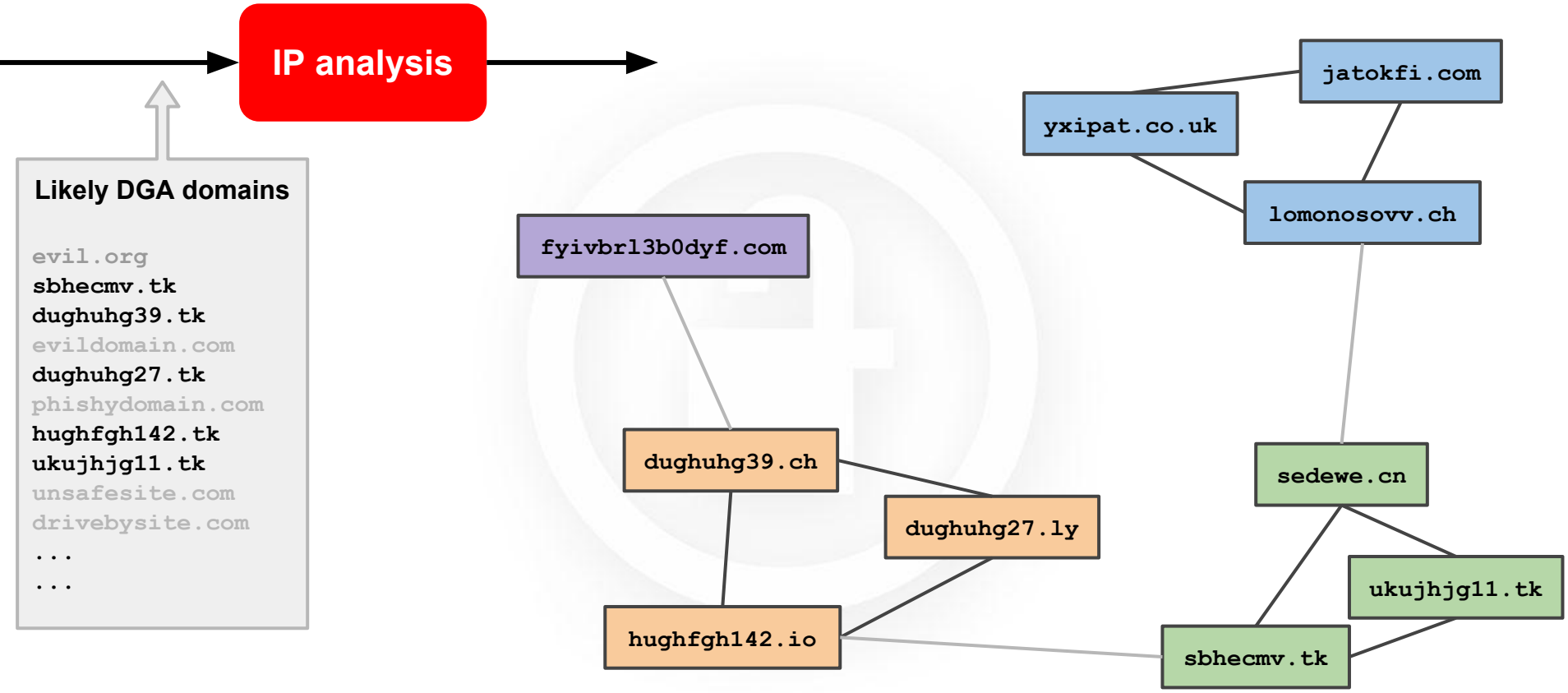
.
.
.

HIGH — **Feature N (n = N)** — LOW

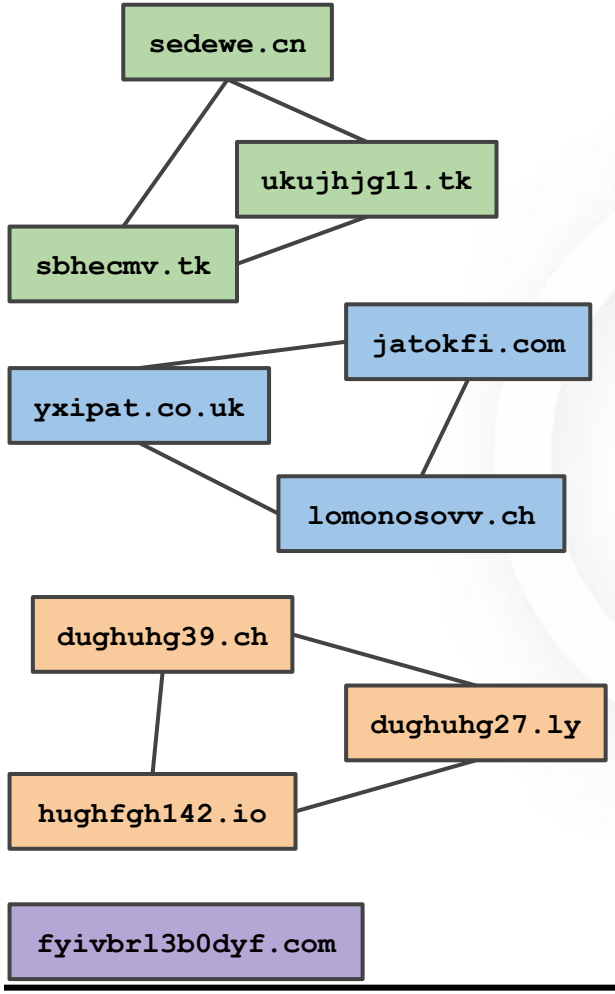
Linguistic features (2D PCA)



Step 2: IP analysis



Step 2: DBSCAN Clustering



Cluster 1

Domains that, in their lifetime, have resolved to the very same IPs

Cluster 2

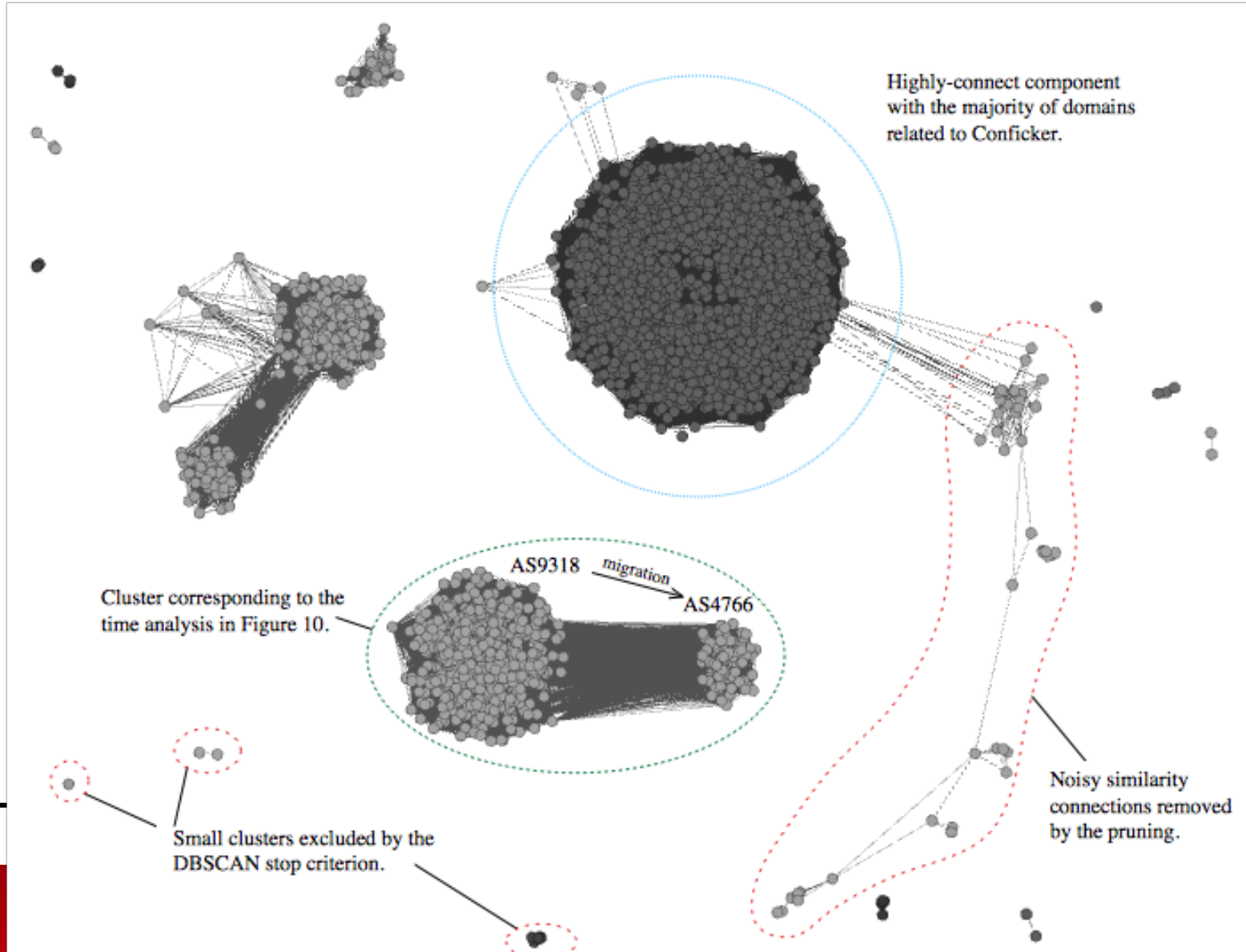
Domains that, in their lifetime, have resolved to the very same IPs

Cluster 3

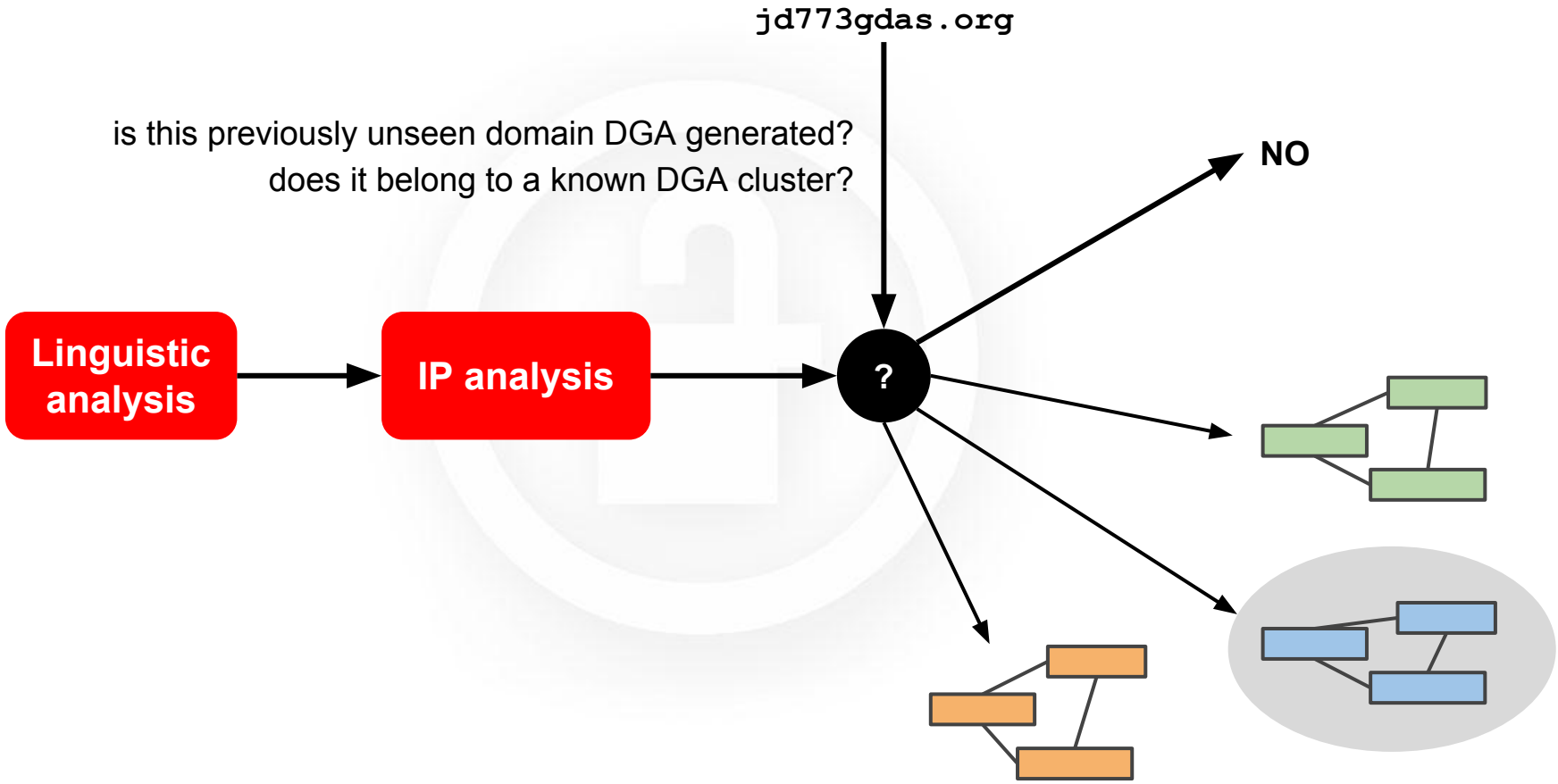
Domains that, in their lifetime, have resolved to the very same IPs

Singleton (removed)

Real output (example)



Classifying new domains



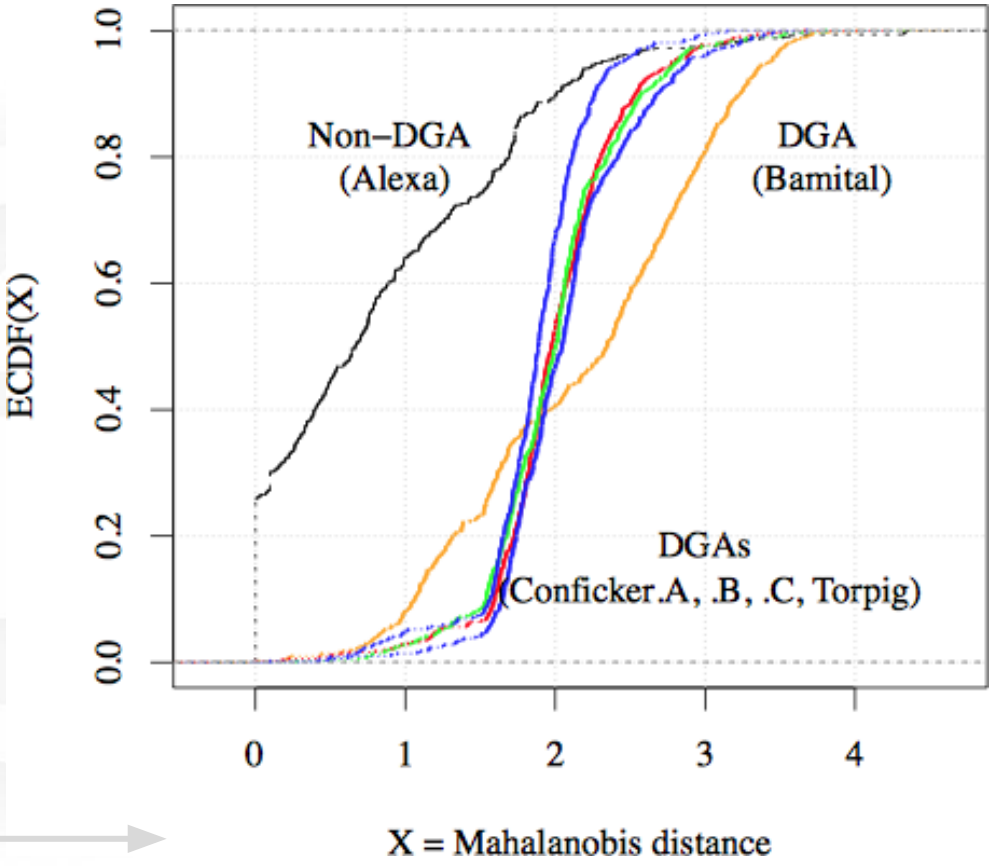
Roadmap

1. Modern cybercrime
2. Botnets
3. Communication channels
4. Domain generation algorithms (DGAs)
5. Detecting DGA-based botnets
6. **Results**

Step 1 on real data

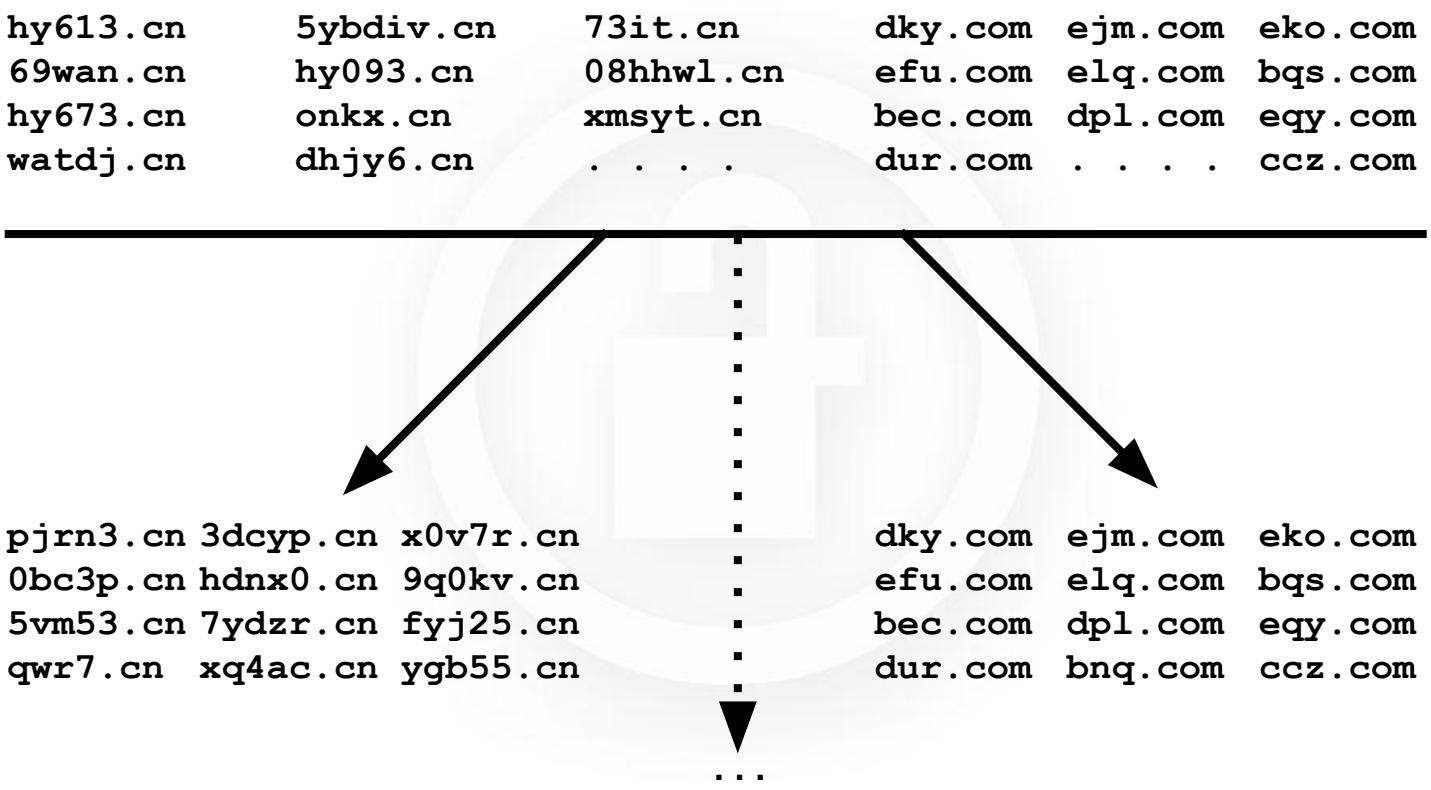
Dataset

- Conficker.A (7,500)
- Conficker.B (7,750)
- Conficker.C (1,101,500)
- Torpig (420)
- Bamital (36,346)



Linguistic analysis → IP analysis

Step 2 on real data

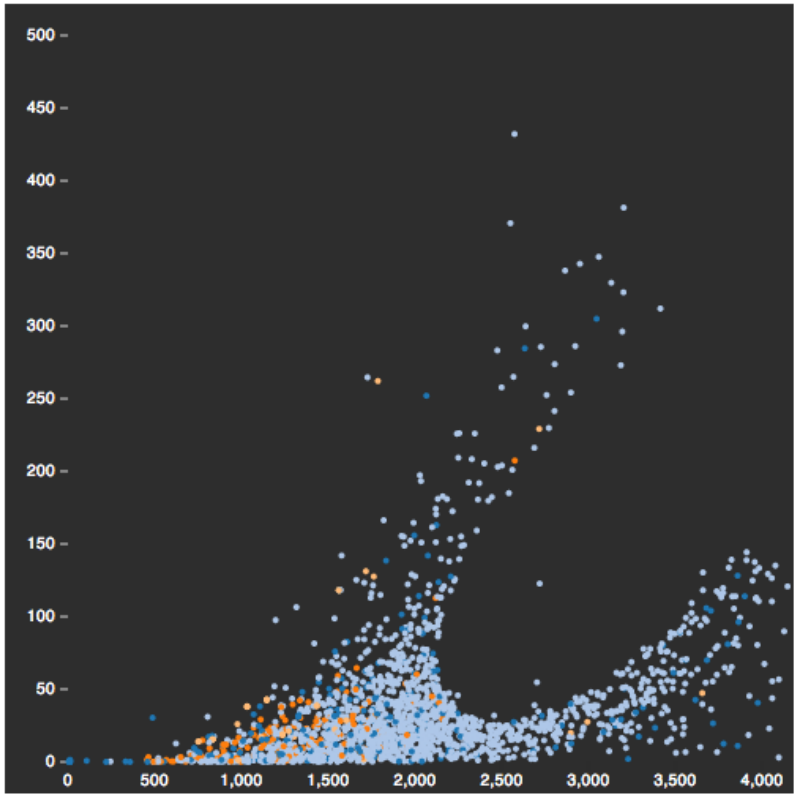


Correct clusters found: **Conficker, Bamital, SpyEye, Palevo**

DEMO (come talk to me offline)

/clustering

DGA Clustering



Dashboard

Select a cluster

- ✓ none
- 5c4cc
- ab6ce
- ad76f

Map

Ongoing research

Non-english baseline

- Italian domain names? Swedish domain names?
- Non-ASCII domains?
  - **`π.com`**
  - **`葉렵ഷ.io`**
  - **`❤★⇄❤.tk`**

Word-based DGAs

- concatenate random, valid words instead of letters
  - **`also-is-dom-yesterday-a-new.com`**

Questions?

http://necst.it

http://maggi.cc

PALSIT

Federico Maggi

federico@maggi.cc

Politecnico di Milano