PHOENIX & CERBERUS

Botnet tracking via precise DGA characterization

Will be presented at DIMVA 2014 (London)

Stefano Schiavoni, Edoardo Colombo
Federico Maggi
Lorenzo Cavallaro
Stefano Zanero

Politecnico Di Milano & Royal Holloway, University of London

POLITECNICO
DI MILANO

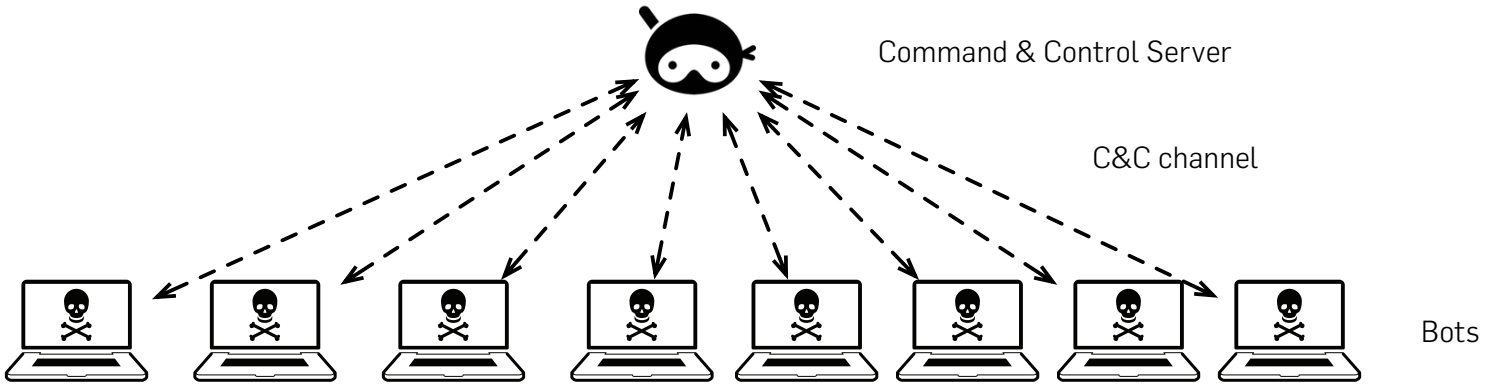
BOTNET: CONTROLLED NETWORK OF COMPROMISED COMPUTERS

BOTNETS (CRYPTOLOCKER CASE)

- Appeared in early Sep 2013
- Earnings estimated at 30 million USD on Dec 18 2013[1]
- FBI takes it down on June 2014 (8 months effort)

[1] http://www.zdnet.com/cryptolockers-crimewave-a-trail-of-millions-in-laundered-bitcoin-7000024579/

MITIGATION



- **C&C channel:**
  - → Fixed IP address
  - → Fixed domain name
  - → **Domain Generation Algorithms (DGAs):**

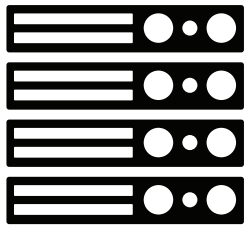
    pseudo-random rendezvous between bots and botmaster
    state of the art centralized botnets

BOTNETS > DOMAIN GENERATION ALGORITHMS

C&C Server, **sjq.info**

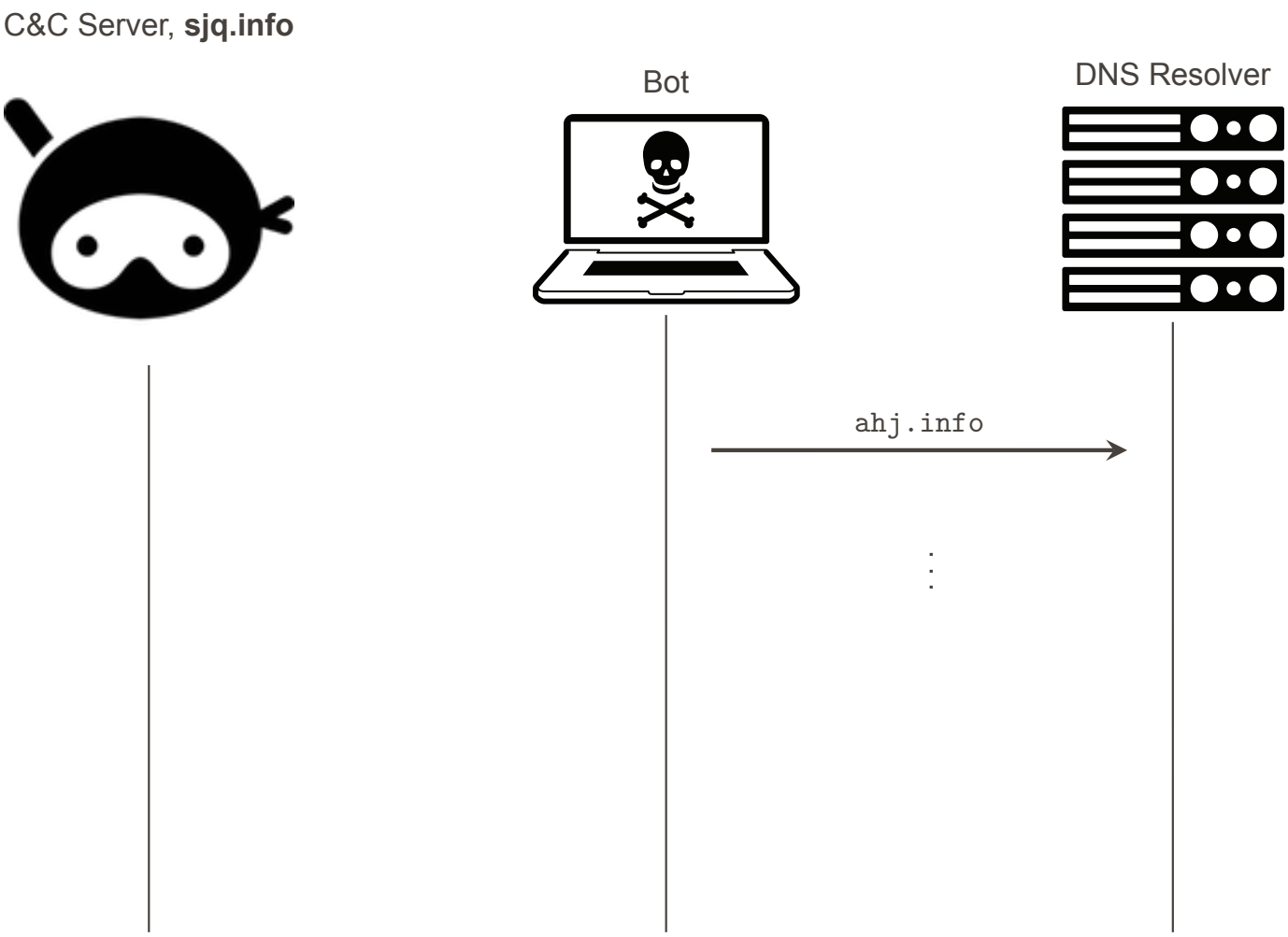
Bot

DNS Resolver

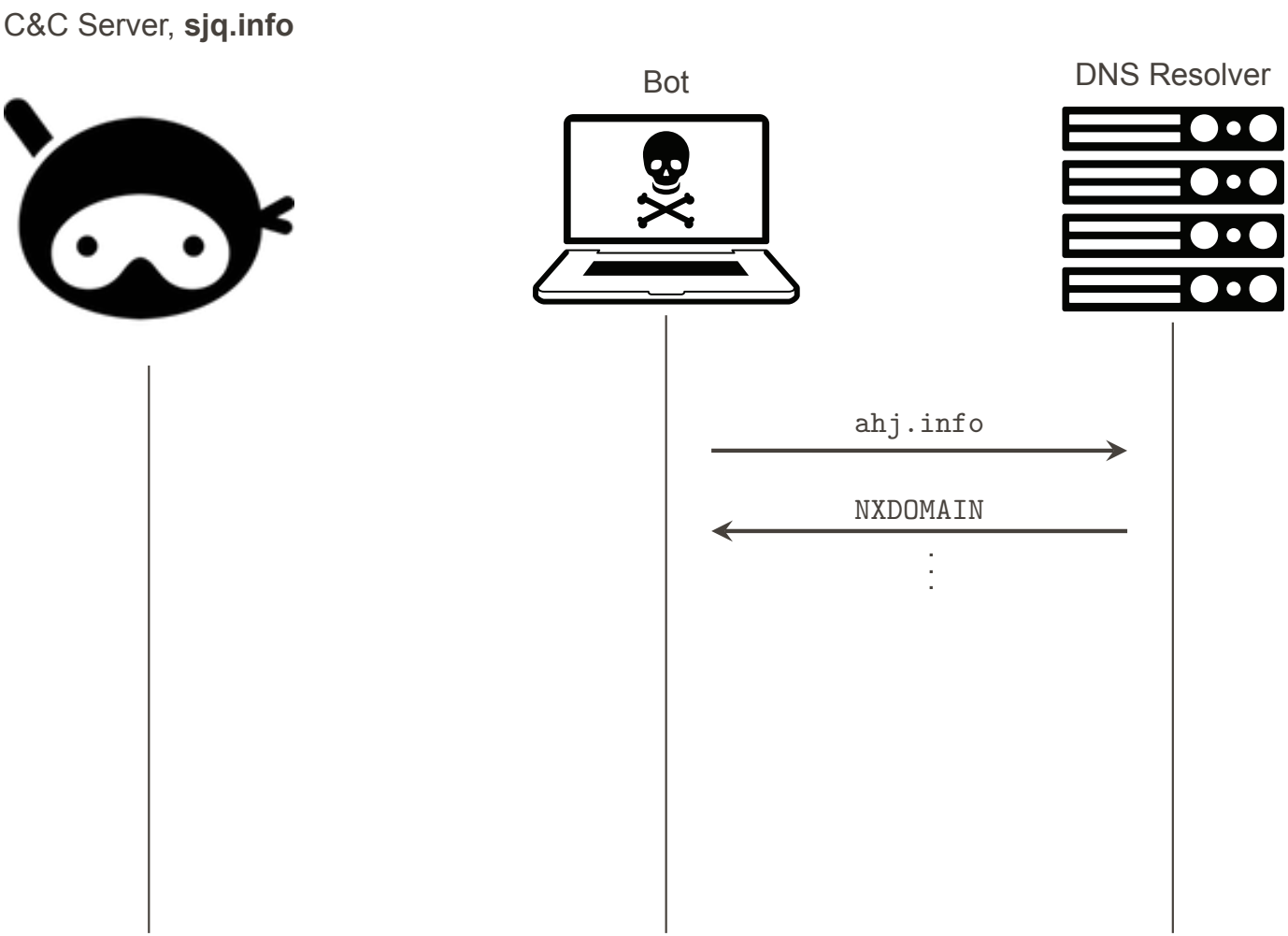


Creative Commons Attribution: Ninja by Anand A Nair from The Noun Project.

BOTNETS > DOMAIN GENERATION ALGORITHMS

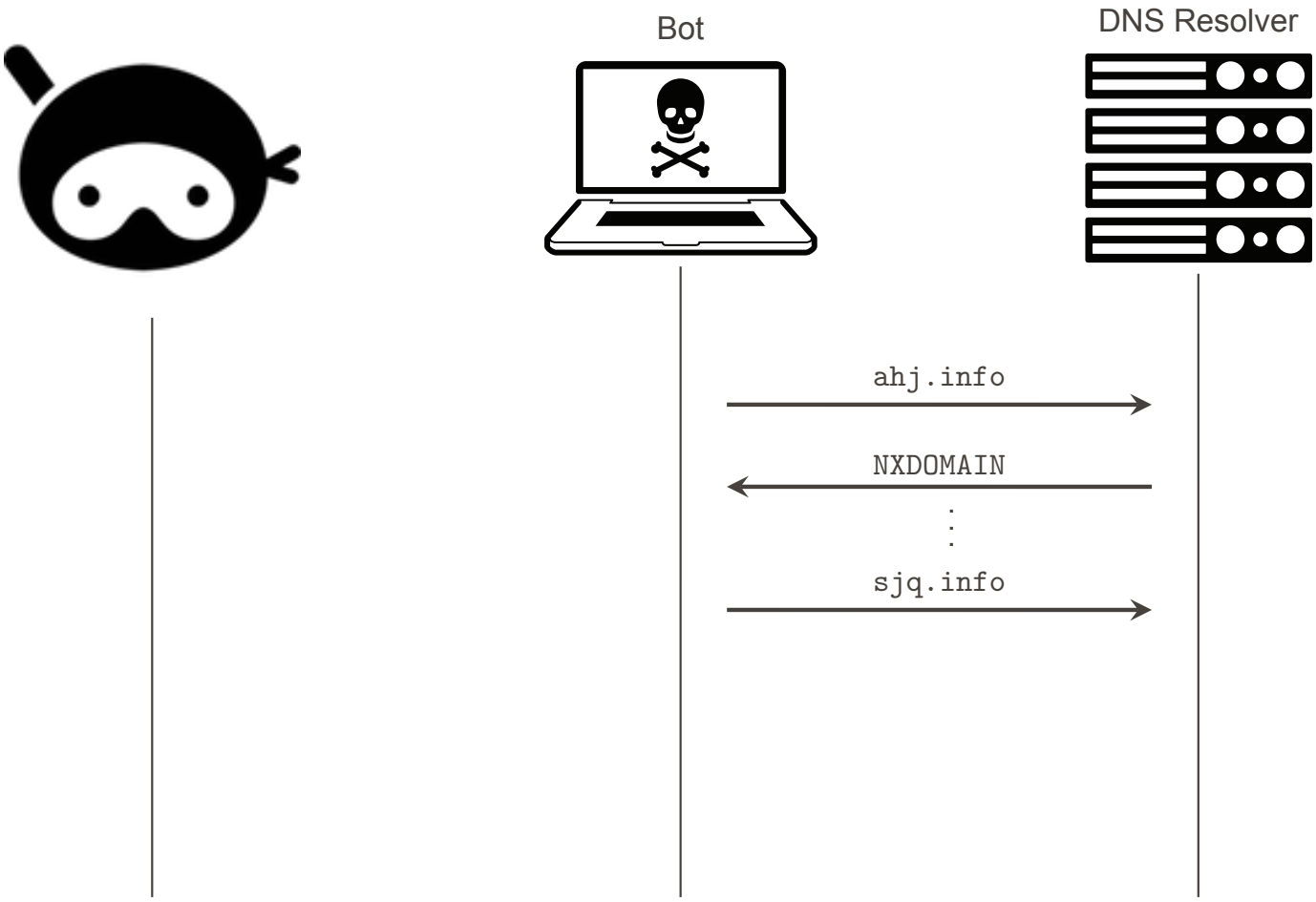


BOTNETS > DOMAIN GENERATION ALGORITHMS



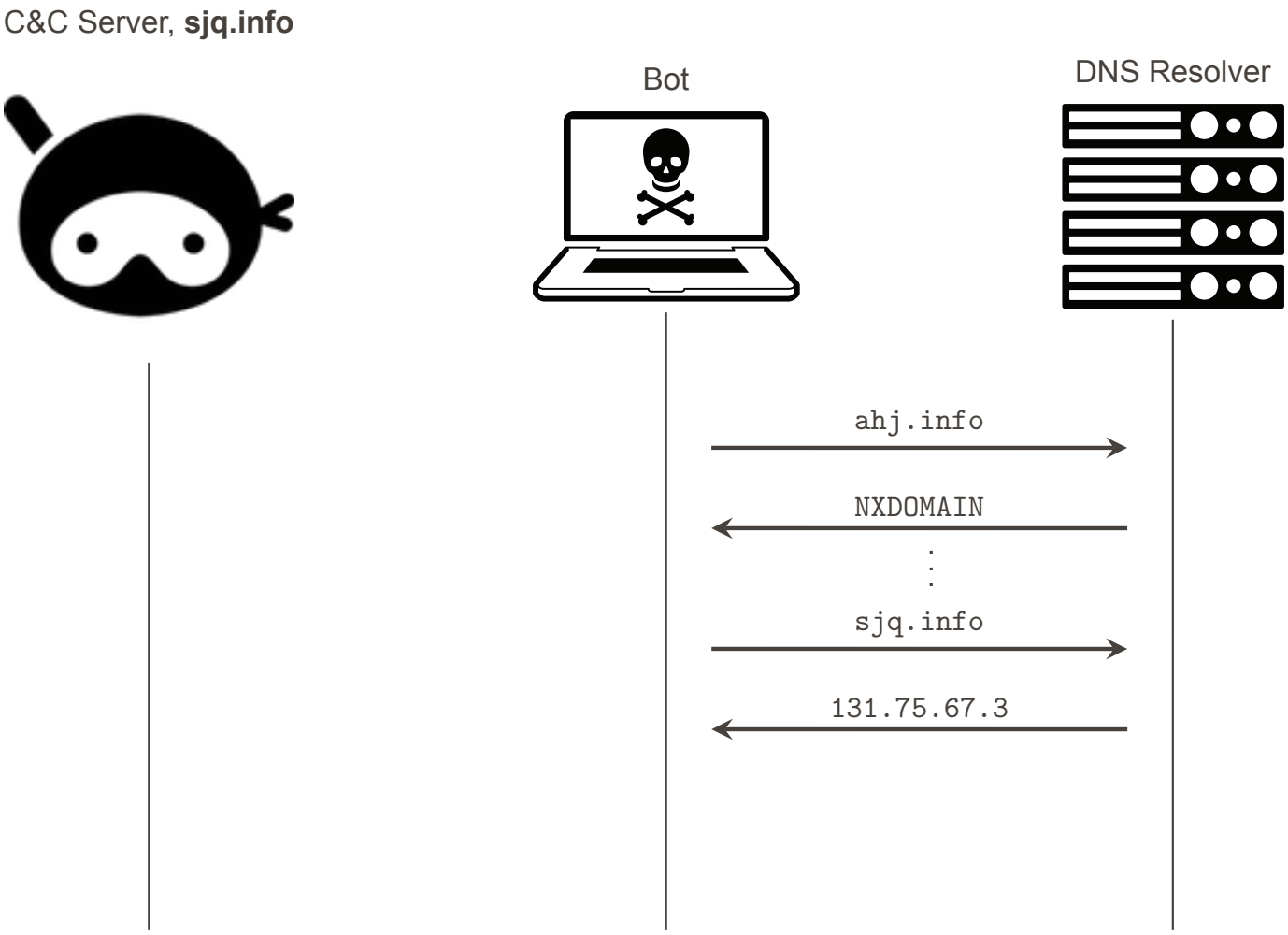
Creative Commons Attribution: Ninja by Anand A Nair from The Noun Project.

BOTNETS > DOMAIN GENERATION ALGORITHMS



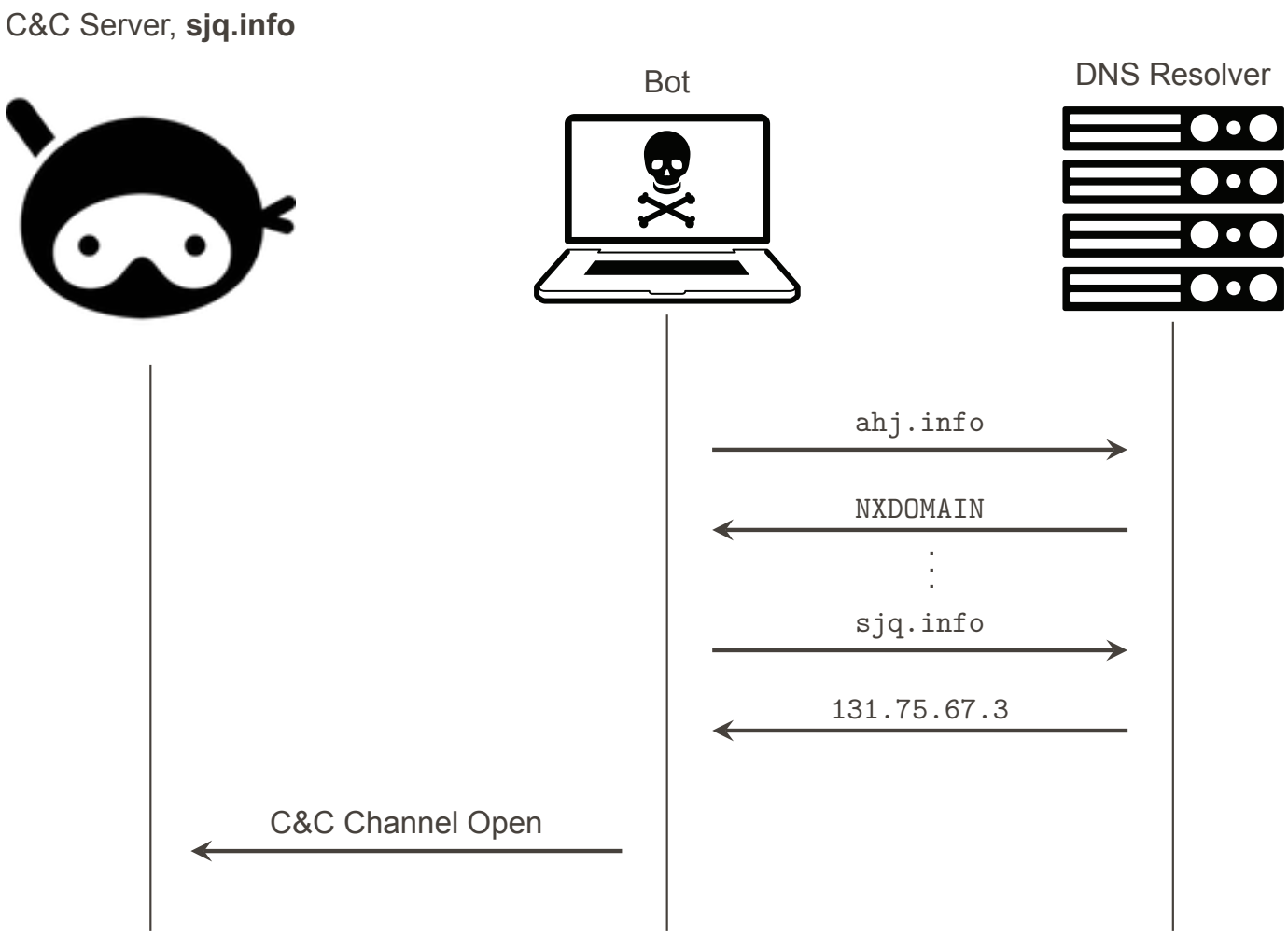
Creative Commons Attribution: Ninja by Anand A Nair from The Noun Project.

BOTNETS > DOMAIN GENERATION ALGORITHMS



Creative Commons Attribution: Ninja by Anand A Nair from The Noun Project.

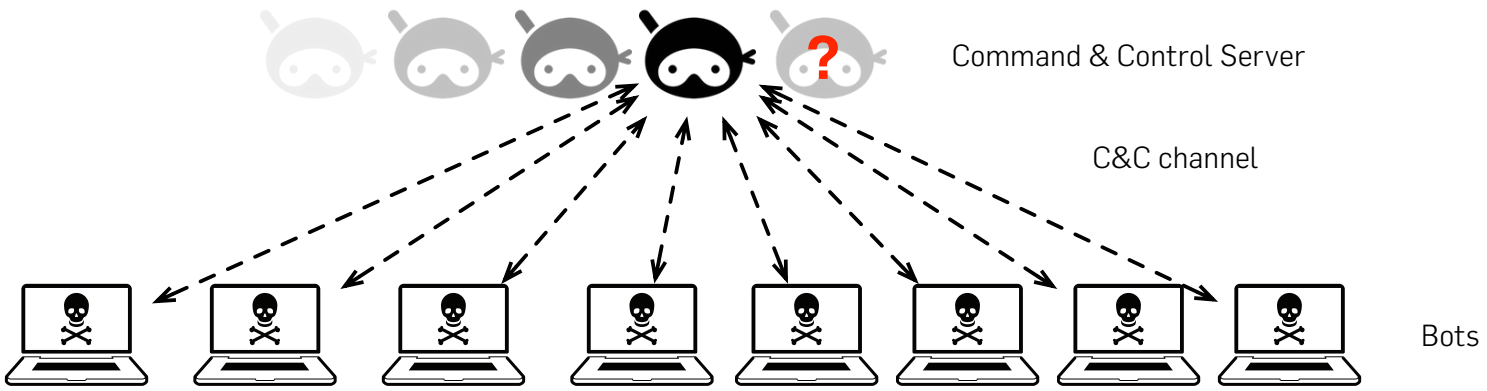
BOTNETS > DOMAIN GENERATION ALGORITHMS



Creative Commons Attribution: Ninja by Anand A Nair from The Noun Project.

DGA > BENEFITS FOR THE BOTMASTERS

- **Asymmetry** Botmasters Vs Defenders
  - → Thousands of domain names,
  - → only one is the active one.
- **Blacklists** quickly obsolete
  - → new domains generated every day,
  - → never re-used (if not after years).



RESEARCH GOALS

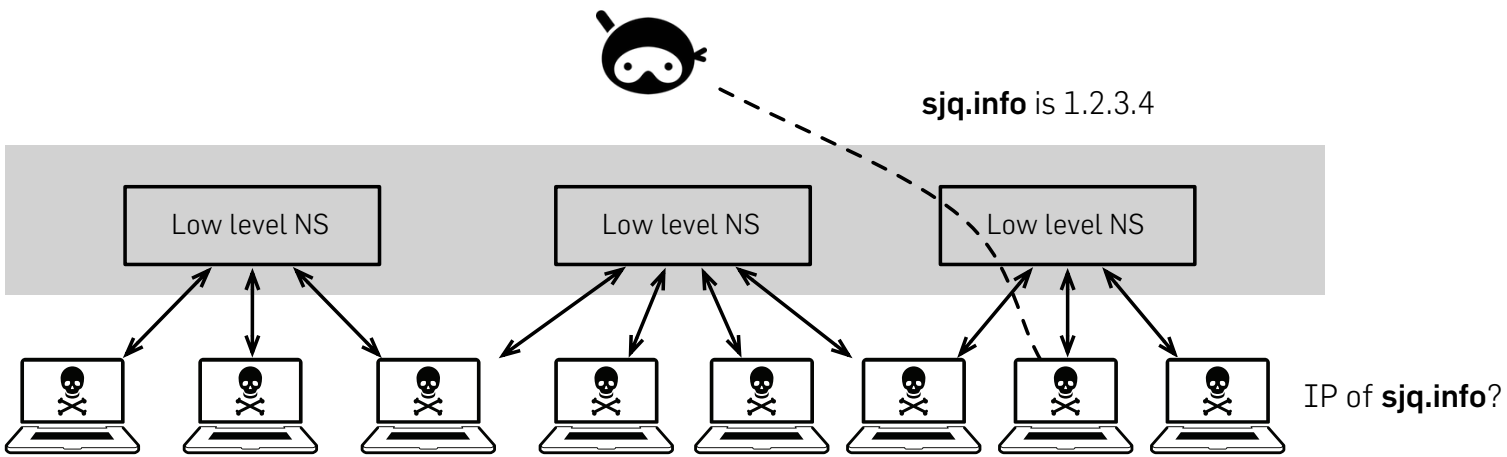
Automatically:

- **distinguish** domains that are generated by a DGA, from those that are not
- **isolate** distinct DGAs ($\simeq$ botnets)

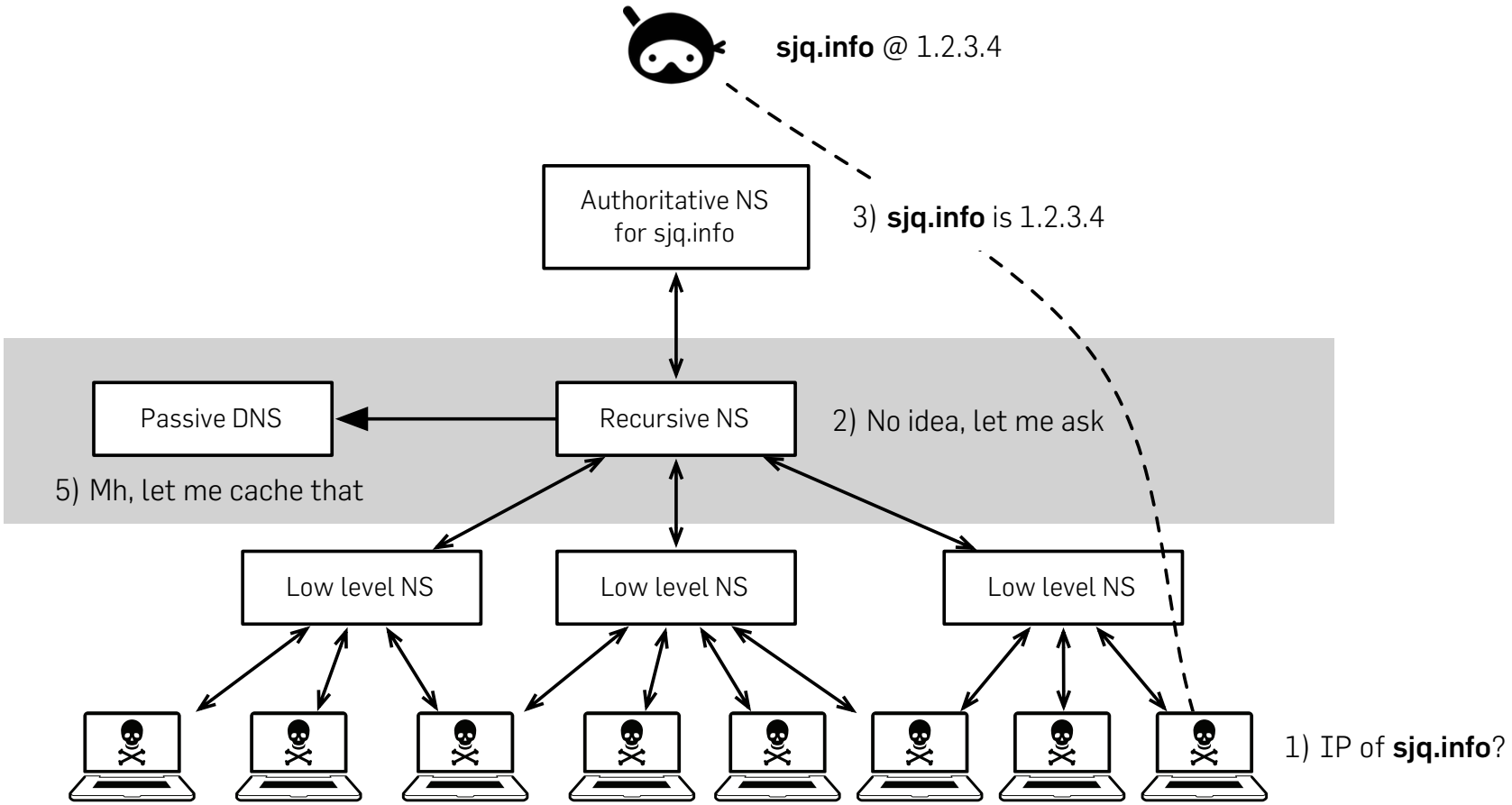
STATE OF THE ART

STATE OF THE ART

- Work at the **lower levels** of the **DNS hierarchy:**
  - → not so easy to deploy,
  - → privacy (visibility of the hosts' IP addresses).



PASSIVE SENSORS



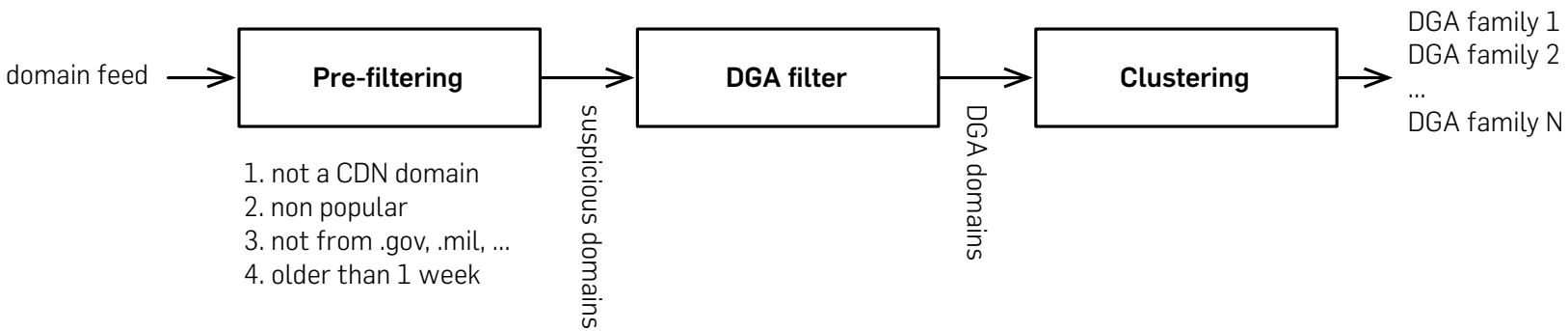
STATE OF THE ART

- **Supervised:** highly dependent on labeled data
  - → "That domain name is known to be DGA generated",
  - → "That other domain is not".

Our approach

Easier to **model benign domains** and detect DGAs as "outliers"

OVERVIEW


domain feed
Pre-filtering
1. not a CDN domain
2. non popular
3. not from .gov, .mil, ...
4. older than 1 week
suspicious domains
DGA filter
DGA domains
Clustering
DGA family 1
DGA family 2
...
DGA family N

OBSERVATION 1: NON-DGA DOMAINS ARE PRONOUNCEABLE

MODEL OF THE "PRONOUNCEABILITY"

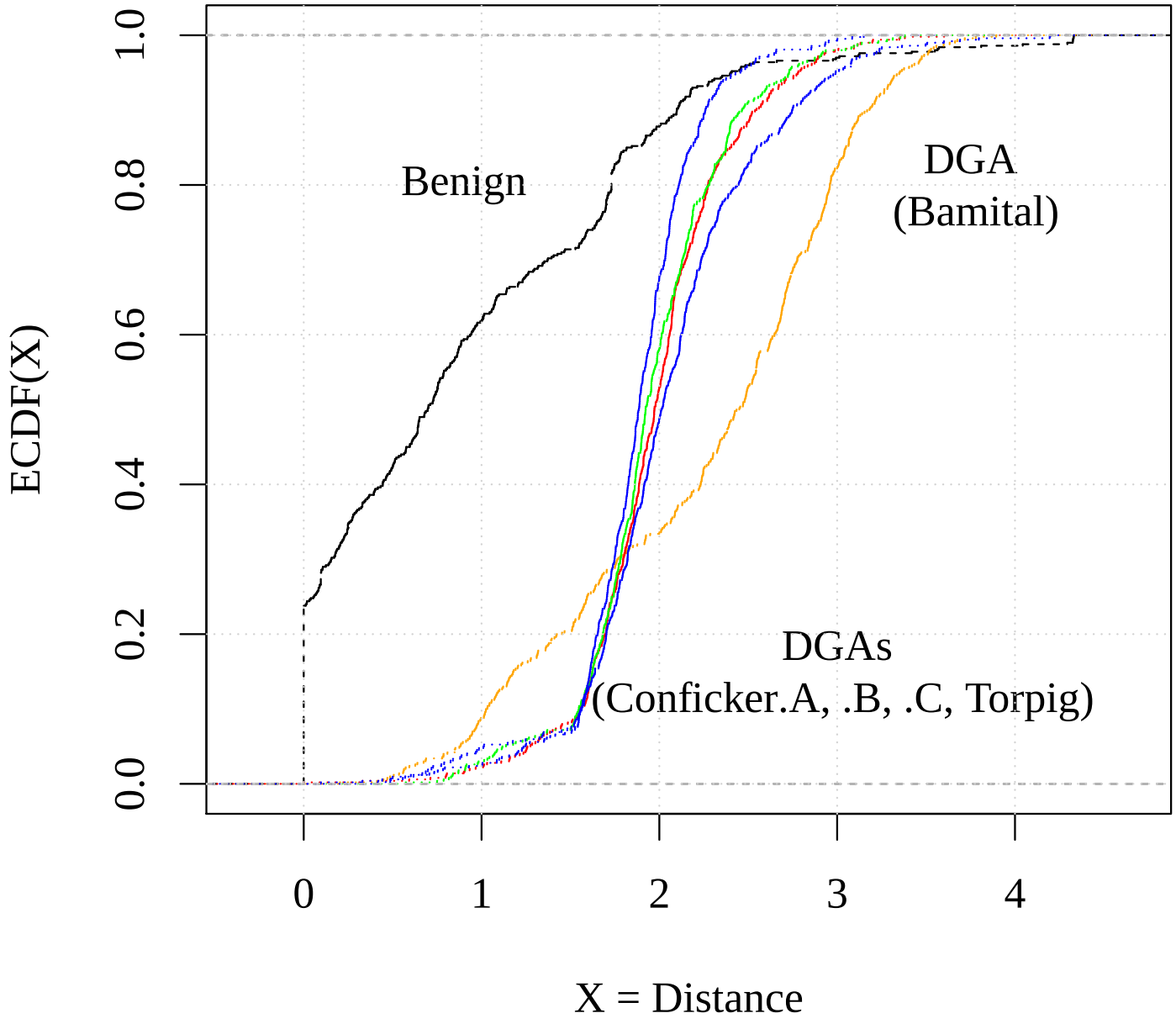
Meaningful Word Ratio (English dict)

$$\frac{\text{len}(\texttt{face}) + \text{len}(\texttt{book})}{\text{len}(\texttt{facebook})} = 1 \qquad \frac{\text{len}(\texttt{pub})}{\text{len}(\texttt{pub03str})} = 0.375.$$

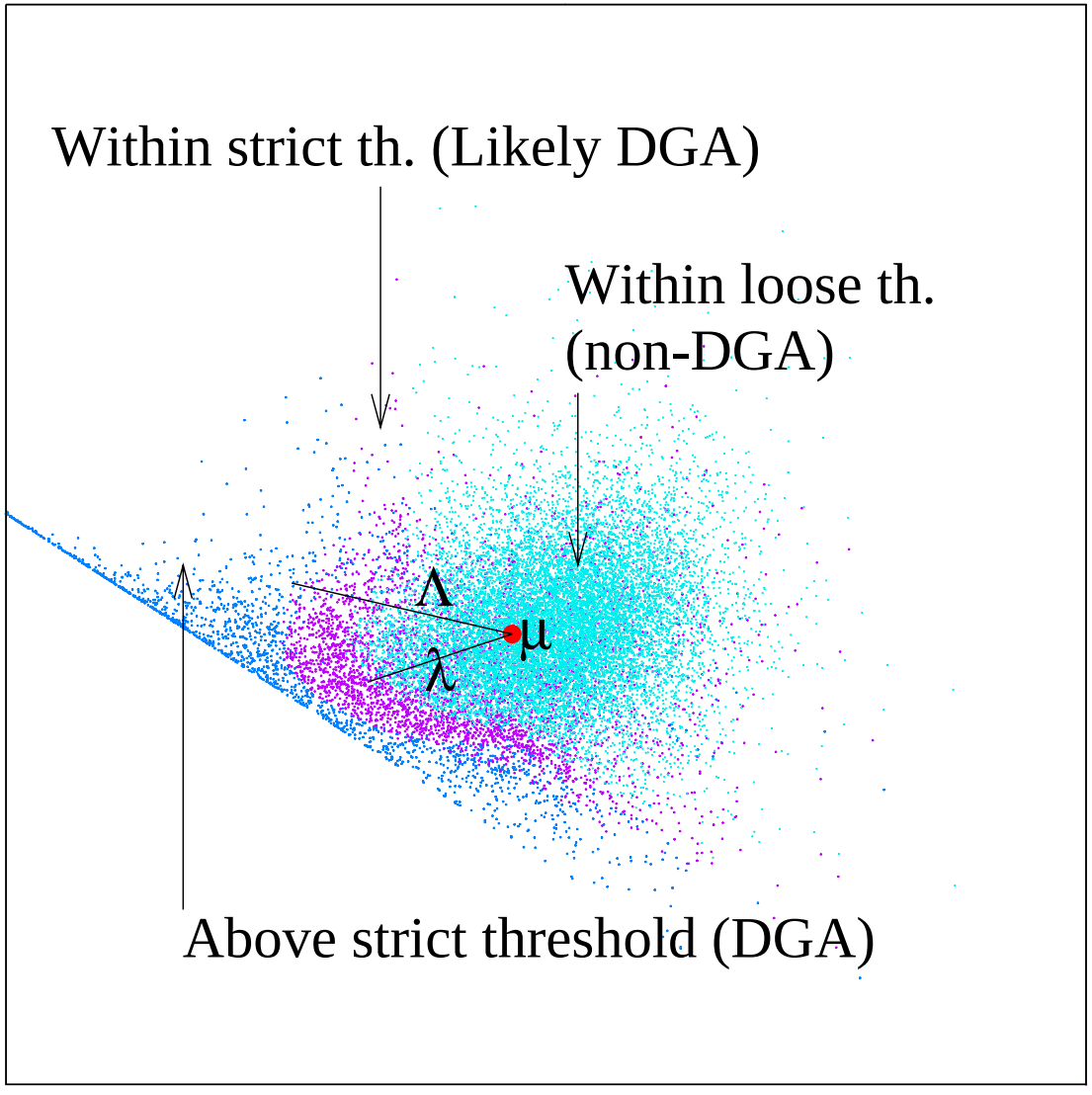
likely **non-DGA generated** likely **DGA generated**

`top10000en.txt` $\rightarrow$ https://books.google.com/ngrams

FEATURE DISTRIBUTION

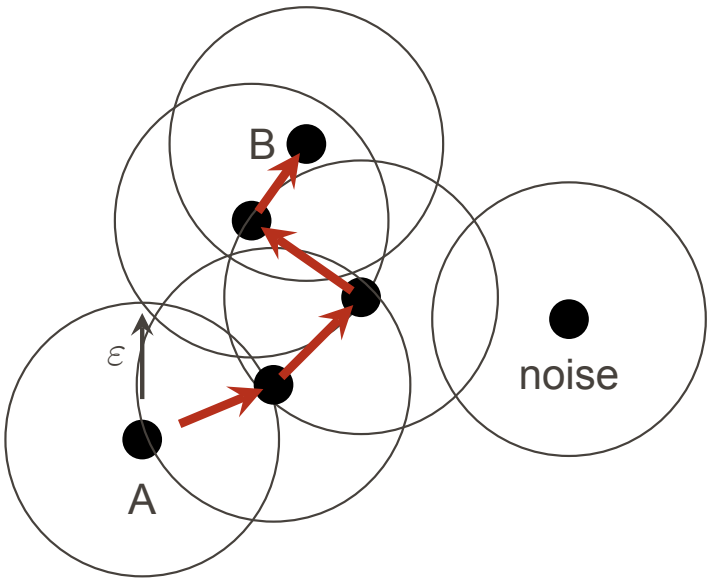


THRESHOLDING



OBSERVATION 2: EACH BOTNET FAMILY WILL USE THE SAME DGA

DBSCAN



Mahalanobis distance

Tuning:

- $minPts$ domains per cluster,
- ε distance threshold.

CLUSTERING > TUNING

$minPts = 7$ domains per cluster

Observation period in days.

Rationale: the bots will contact the C&C server at least **once a day** on a new domain.

$\frac{\text{intra-cluster distances}}{\text{inter-cluster distances}} \rightarrow 0$ (ε to minimize)

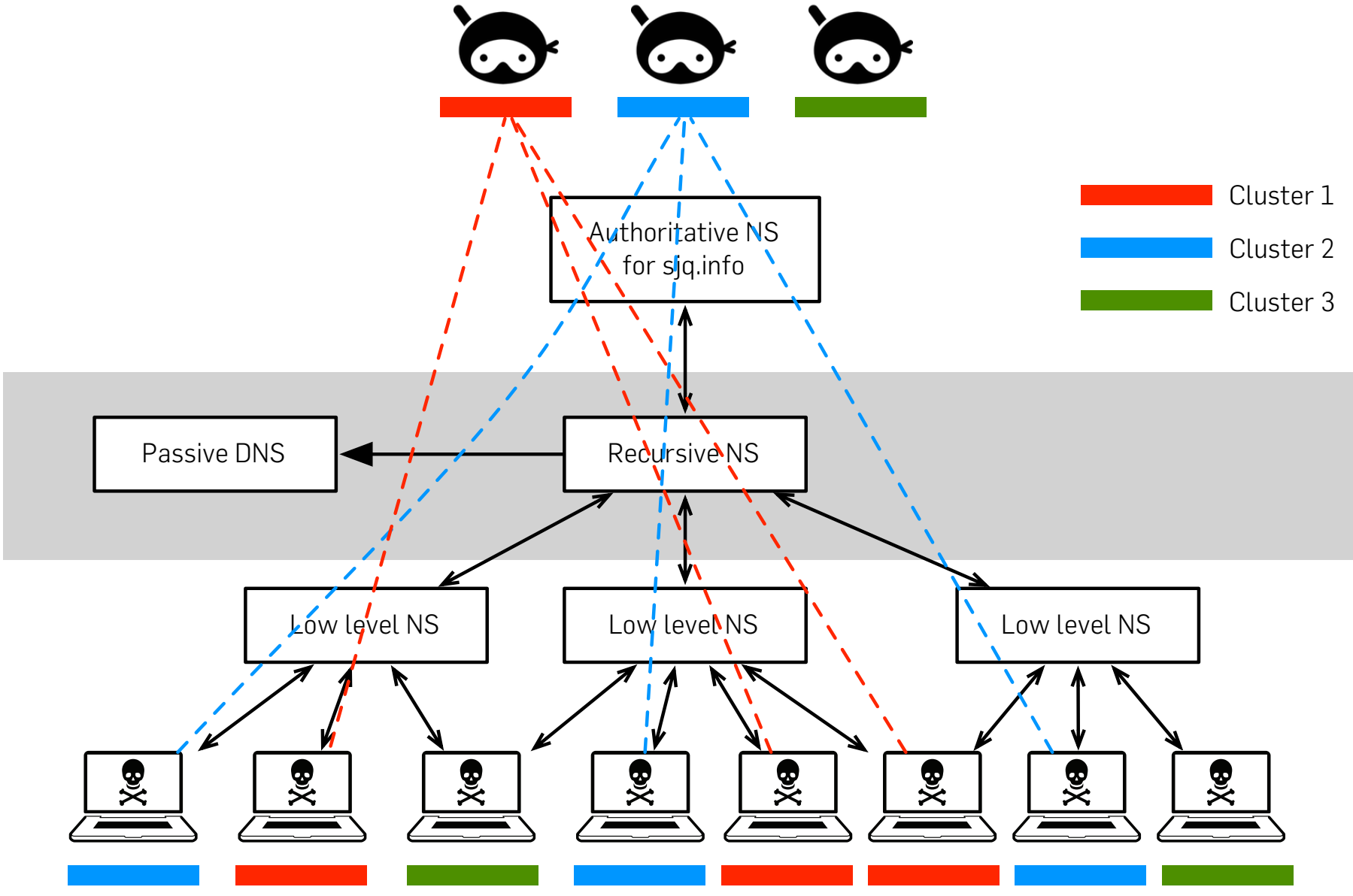
RESULTS (1 WEEK OF TRAFFIC)

Cluster f105c		Cluster 0f468	
IPs:	`176.74.176.175`	IPs:	`217.119.57.22`
	`208.87.35.107`		`91.215.158.57`
			`178.162.164.24`
			`94.103.151.195`
Domains:	`cvq.com`	Domains:	`jhhfghf7.tk`
	`epu.org`		`faukiijjj25.tk`
	`bwn.org`		`pvgvy.tk`
(Botnet:	Palevo)	(Botnet:	Sality)

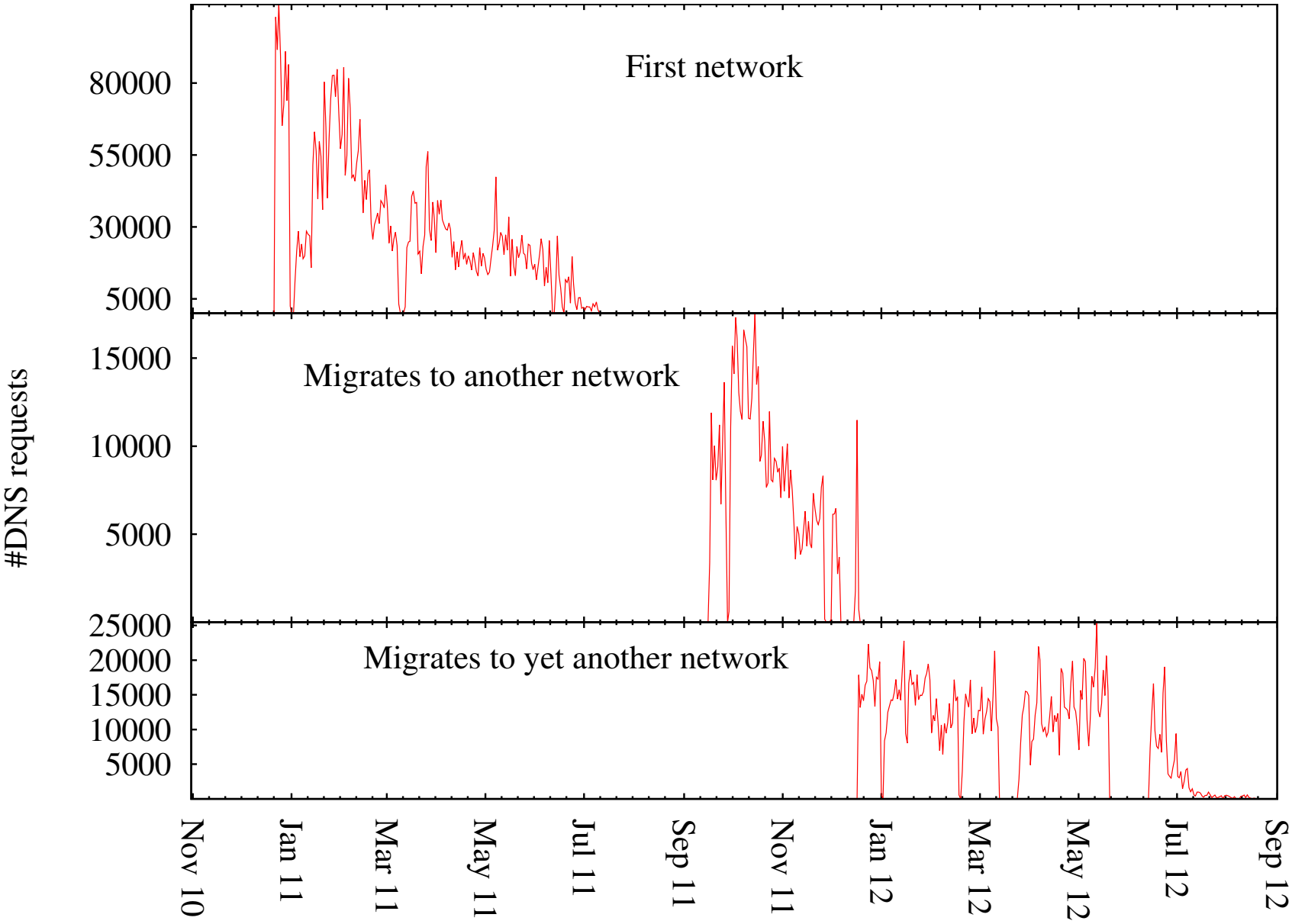
RESULTS (1 WEEK OF TRAFFIC)

Botnet	IPs	Size
Sality	62.116.181.25	26
Palevo	199.59.243.118	40
Jadtre*	69.43.161.180	173
	69.43.161.174	
Jadtre**	69.43.161.180	37
Jadtre***	69.43.161.167	47
Hiloti	69.43.161.167	24
Palevo	82.98.86.171	142
	82.98.86.176	
	82.98.86.175	
Jusabli	69.58.188.49	73
Generic Trojan	82.98.86.169	57
	82.98.86.162	
	82.98.86.178	
	82.98.86.163	

APPLICATION: C&C TRACKING



APPLICATION: C&C TRACKING



APPLICATION: INTELLIGENCE

- February 9th, 2013: **inquiry** from a group of security officers
- **previously unseen** list of domains, which resembled no known botnet
- Phoenix matched it on a **Conficker** cluster
- It was **indeed** Conficker (variant B)

OVERALL RESULTS

PRE-FILTERING

~38% **not** in the **top 1M popular sites**

~30% **not** in the most popular **CDNs**

~14% **no** TLD **clearance**

~11% **likely non DGA** (DGA filter)

~6% **younger than 1 week** (most expensive analysis)

FILTERING (1 WEEK OF TRAFFIC)

- **3,576 suspicious** domains collected
- **187** DGA domains (manually confirmed correct)
- **47 clusters** (matching real families)
- **time (with pre-filtering):** hours $\rightarrow$ minutes

FUTURE CHALLENGES

FUTURE CHALLENGES

- `this-would-evade-the-linguistic-filter.com`
  $\rightarrow$ direction: rely more on external features
- remove language dependency
  $\rightarrow$ direction: model every language
- put it in production (scalability)
  $\rightarrow$ direction: map-reduce implementation of DBSCAN


THANK YOU
federico.maggi@polimi.it
http://maggi.cc
@phretor