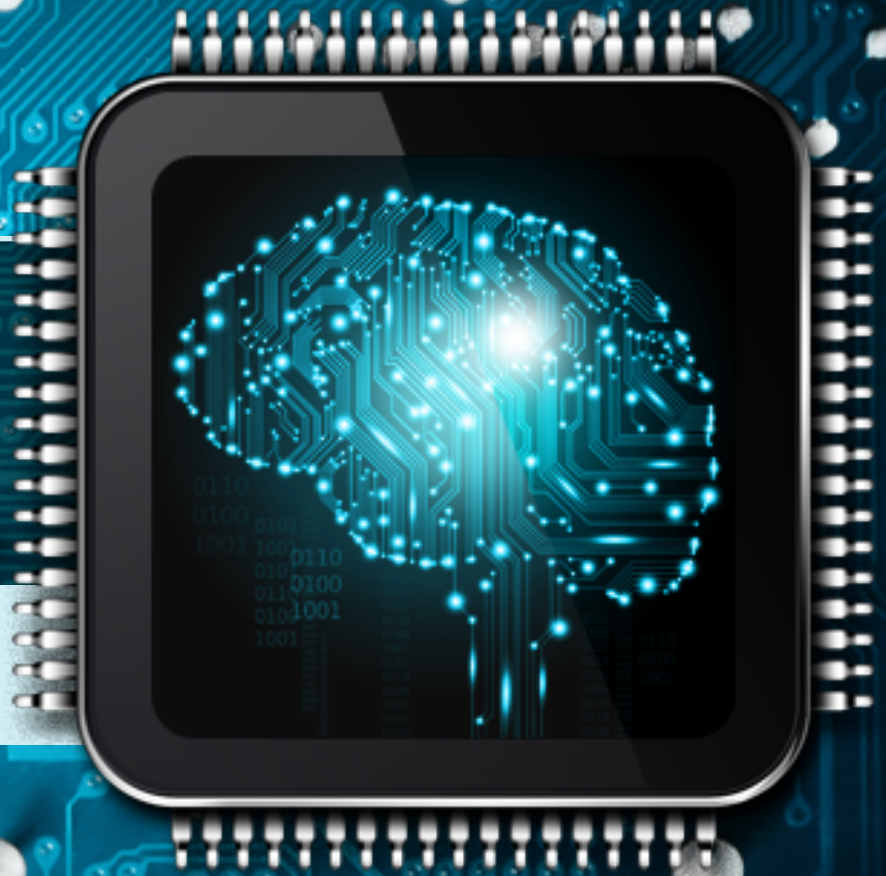
anubisnetworks™
a BITSIGHT company

CONDENSER: A Graph-based Approach for Detecting Botnets



Pedro Camelo, João Moura and Ludwig Kriphall

botconf 2014

Agenda

- Introduction
- Botnets and Machine Learning
- Detection and Correlation
- Conclusion

Introduction

Pedro Camelo

- Took my MSc in Computer Engineering at FCT/UNL (last week, yay!)
- This is the outcome of my MSc Thesis
- Always liked the computer security field
- R&D Team from AnubisNetworks

Introduction

Collaboration

- João Moura
  - Taking his PhD in Artificial Intelligence
  - R&D Team from AnubisNetworks
- Prof. Ludwig Krippahl
  - PhD on Biochemistry
  - MsC on Applied Artificial Intelligence

Botnets

Botnets and Its Tails

Core Concepts

- Botnets are group of infected devices controlled by one or more operators (botmasters)
- May have one or more command and control nodes (aka C2, C&C)
- Evasion techniques to evade from takedowns and detection

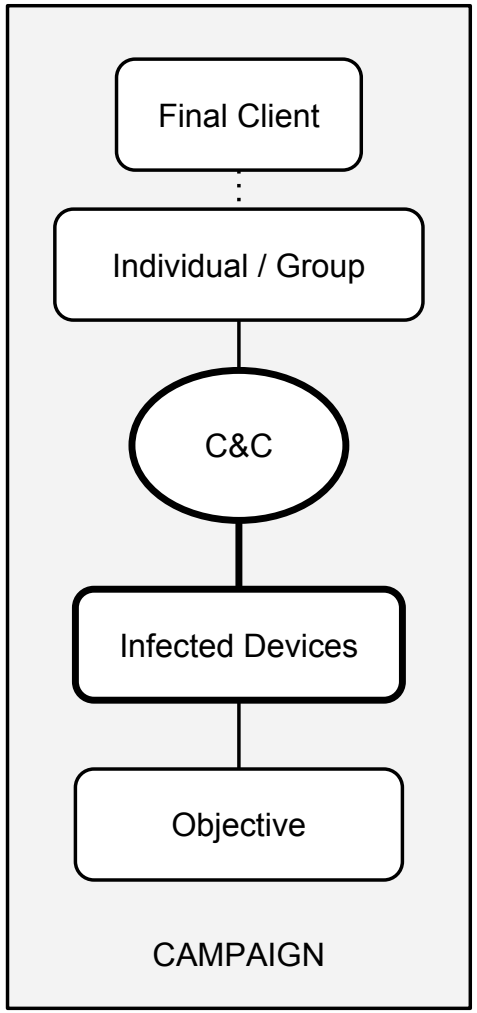
Campaigns

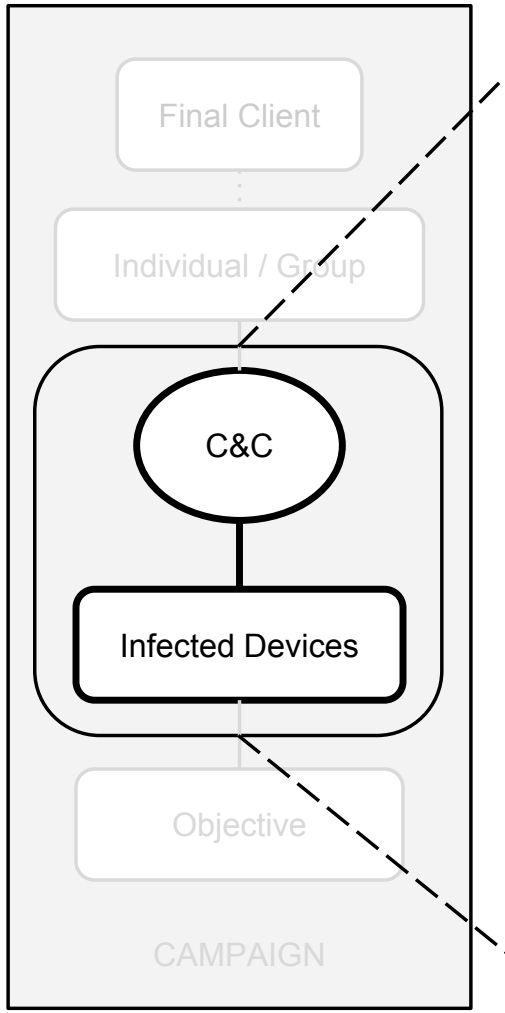
What are them

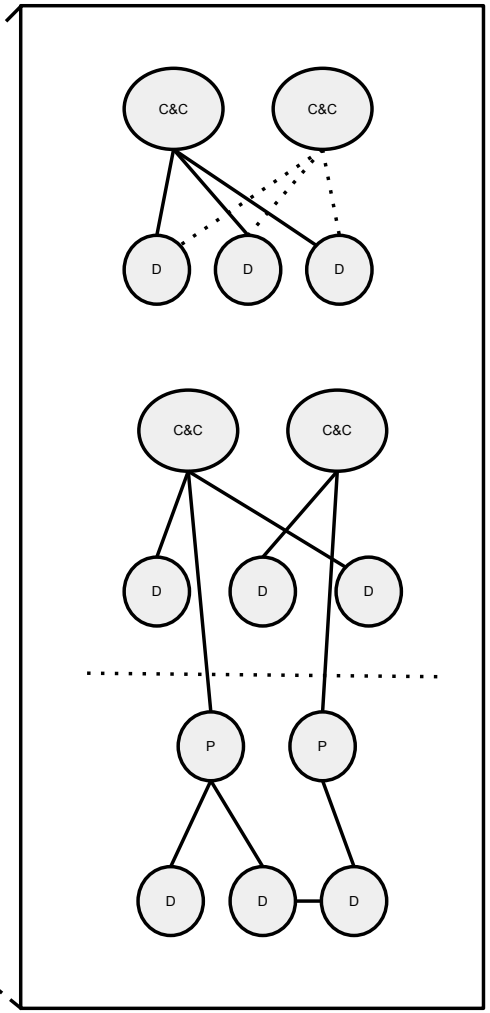
- Are requested by a (blackhat) client
- Blackhat groups implement or use an existing botnet
- They use one or more C2 to control infected devices
- Infected devices execute commands ordered by the C2 to fulfil its (client) objective.

Architecture & Topologies

Definition and Architecture





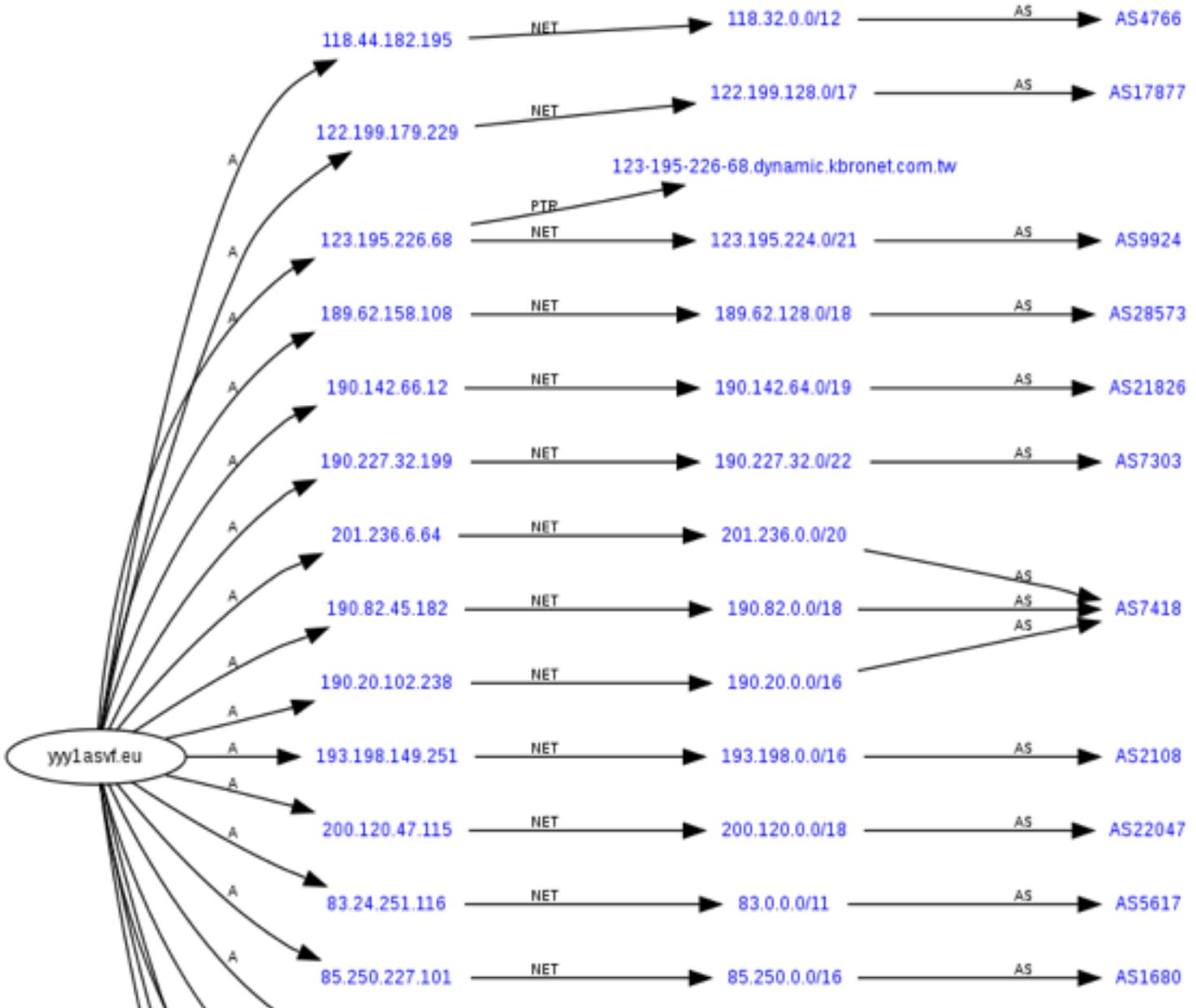


Evasion Techniques

- Encrypted comm
- (Double) Fast-Flux
- Domain Generation Algorithms (aka DGAs)
- **Others**
  - Legitimate domain resolutions to get C&C IP

Fast Flux and Double Fast-Flux

What is it ...



Domain Generation Algorithms

What is it ...

jedisct1 / **g01exploit-dga.rb**

Last active on Mar 13, 2013

g01 exploit kit DGA names generator

```
#! /usr/bin/env ruby

DOMAINS = %w(.doesntexist.com .dnsalias.com .dynalias.com)

DICT = %w(as un si speed no r in me da a o c try to n h call us why q
          k old j g how ri i net t ko tu host on ad portal na order b ask l s d
          po cat for m off own e f p le is)

DICT_LEN = DICT.length

ts = Time.now.utc

c0 = ts.hour
c1 = ts.day   + c0
c2 = ts.month + c1 - 1
c3 = ts.year  + c2

d0 = c0 % DICT_LEN
d1 = c1 % DICT_LEN
d2 = c2 % DICT_LEN
d3 = c3 % DICT_LEN

d1 = (d1 + 1) % DICT_LEN if d0 == d1
d2 = (d2 + 1) % DICT_LEN if d1 == d2
d3 = (d3 + 1) % DICT_LEN if d2 == d3

domain    = DOMAINS[c0 % DOMAINS.length]
subdomain = [ d0, d1, d2, d3 ].map { |x| DICT[x] }.join

name = subdomain + domain

puts name
```

Legitimate domain resolutions to get C&C IP

Necurs PoC


```
$ dig -t A example.com @a.iana-servers.net

; <<>> DiG 9.8.3-P1 <<>> -t A example.com
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 4622
;; flags: qr rd ra; QUERY: 1, ANSWER: 1, AUTHORITY: 2, ADDITIONAL: 0

;; QUESTION SECTION:
;example.com.                   IN   A

;; ANSWER SECTION:
example.com.        86400      IN   A    93.184.216.119
```

93.184.216.119 - 01011101.10111000.11011000.01110111

93.184.202.119 - 01011101.10111000.11001010.01110111

C&C IP

Detection Methods

Detection Methods

- **Passive Detection**
  - Packet Inspection
  - Network Flows (discarded for this research)
  - Domain Name Syntax
- **Active Detection**
  - Domain Name Resolutions
- **Information Correlation**
- **Graph Oriented Queries**

Detection Methods

- Packet Analysis

```
{
  "ua": "Mozilla/5.0 (...)",
  "httpcode": "200",
  "ref": "http://ref.example.com/",
  "uri": "http://example.com/checkin.php",
  "method": "GET",
  "ip": "127.0.0.1",
  "httpversion": "HTTP/1.0",
  "sz": “1212"
}
```

A. Group same pattern traffic discarding destination information
B. Correlate IP connections by common destinations

Detection Methods

- Domain Name Syntax

```
jyyfmnefedjogsh.biz
dxejhopldgymgld.com
oiokidamwjythao.info
qayuttfyvdsufol.net
bgtyvxkyemflyjo.co.uk
ujtohypxdfvrtor.org
fposjduxloiiurh.net
srjeviklelcqdbl.biz
hhydutakkicjusf.ru
```

Random Chars

```
asiorderb.doesntexist.com
unnetbask.dnsalias.com
sitaskl.dynalias.com
speedkols.doesntexist.com
notusd.dnsalias.com
rhostdpo.dynalias.com
inonpocat.doesntexist.com
meadcatfor.dnsalias.com
daportalform.dynalias.com
```

Dictionary Based

- Vowels Ratio
- Consonants Ratio
- Domain Name Length
- Vowel Consonant Ratio
- English Dictionary Words
- Known Words used by Malware Samples
- …

Detection Methods

- Domain Name Resolutions

```
dig -t A example.com @a.iana-servers.net

; <<>> DiG 9.8.3-P1 <<>> -t A example.com @a.iana-servers.net
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 47748
;; flags: qr aa rd; QUERY: 1, ANSWER: 1, AUTHORITY: 2, ADDITIONAL: 0
;; WARNING: recursion requested but not available

;; QUESTION SECTION:
;example.com.                   IN   A

;; ANSWER SECTION:
example.com.            60     IN   A    93.184.216.119
```

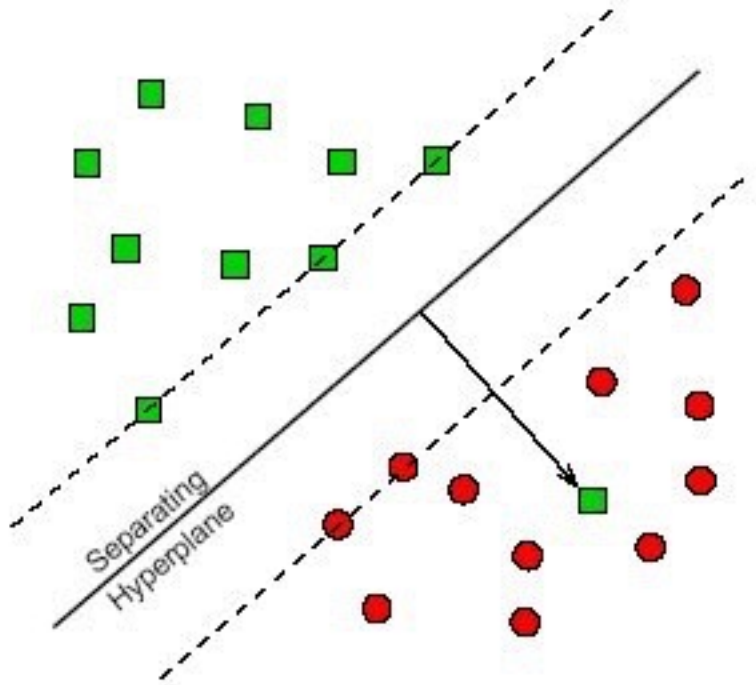
```
example.com.      60  IN  A   127.0.0.1
example.com.      60  IN  A   127.0.0.2
example.com.      60  IN  A   127.0.1.3
example.com.      60  IN  A   127.0.2.4
example.com.      60  IN  A   127.0.3.5
example.com.      60  IN  A   127.1.0.6
example.com.      60  IN  A   127.2.0.1
example.com.      60  IN  A   127.3.0.1
example.com.      60  IN  A   127.0.8.1
example.com.      60  IN  A   127.0.7.2
example.com.      60  IN  A   127.0.0.8
example.com.      60  IN  A   127.0.0.9
example.com.      60  IN  A   127.0.9.1
example.com.      60  IN  A   127.1.1.1
example.com.      60  IN  A   127.1.2.1
```

Detection Methods

- Support Vector Machines for DGA Domain Classification

Non-separable training sets

Use linear separation, but admit training errors.



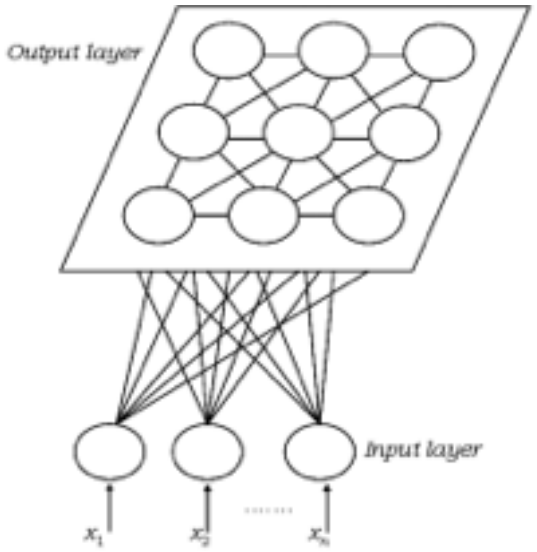
Penalty of error: distance to hyperplane multiplied by *error cost* C.

jyyfmnefedjogsh.biz
dxejhopldgymgld.com
oiokidamwjythao.info
foobar.com
superawesome.co.uk
ujtohypxdfvrtor.org
fposjduxloiiurh.net
facebook.com
rbnbusiness.ru

Domain Classification Example

Detection Methods

- Neural Networks (Self Organising Maps) for Traffic Grouping




```
{
        "ips": {
          "extended": [{
              "hits": 4,
              "intersections": 1,
              "ip": "127.0.0.1"
            },
            {
              "hits": 5,
              "intersections": 1,
              "ip": "127.0.0.2"
            }],
          "values": ["127.0.0.1","127.0.0.2"]
        },
        "metrics": {
          "avg_hits": 4.5,
          "avg_intersections": 1.0,
          "avg_ips": 1.0,
          "hits": 9,
          "avg_ip_per_pattern": 1.0,
          "ips": 2
        },
```



```
"patterns": [{
            "hits": 4,
            "value": {
              "host": "jhia2iu6skja9.com",
              "httpcode": 200,
              "httpversion": "HTTP/1.0",
              "method": "GET",
              "size": 1,
              "uri_path": "/update",
              "seems_dga": true
            }},{
            "hits": 5,
            "value": {
              "host": "jsia5iueseja0.com",
              "httpcode": 200,
              "httpversion": "HTTP/1.0",
              "method": "GET",
              "size": 1,
              "uri_path": "/update",
              "seems_dga": true
            }}]
      }
```

Information Correlation

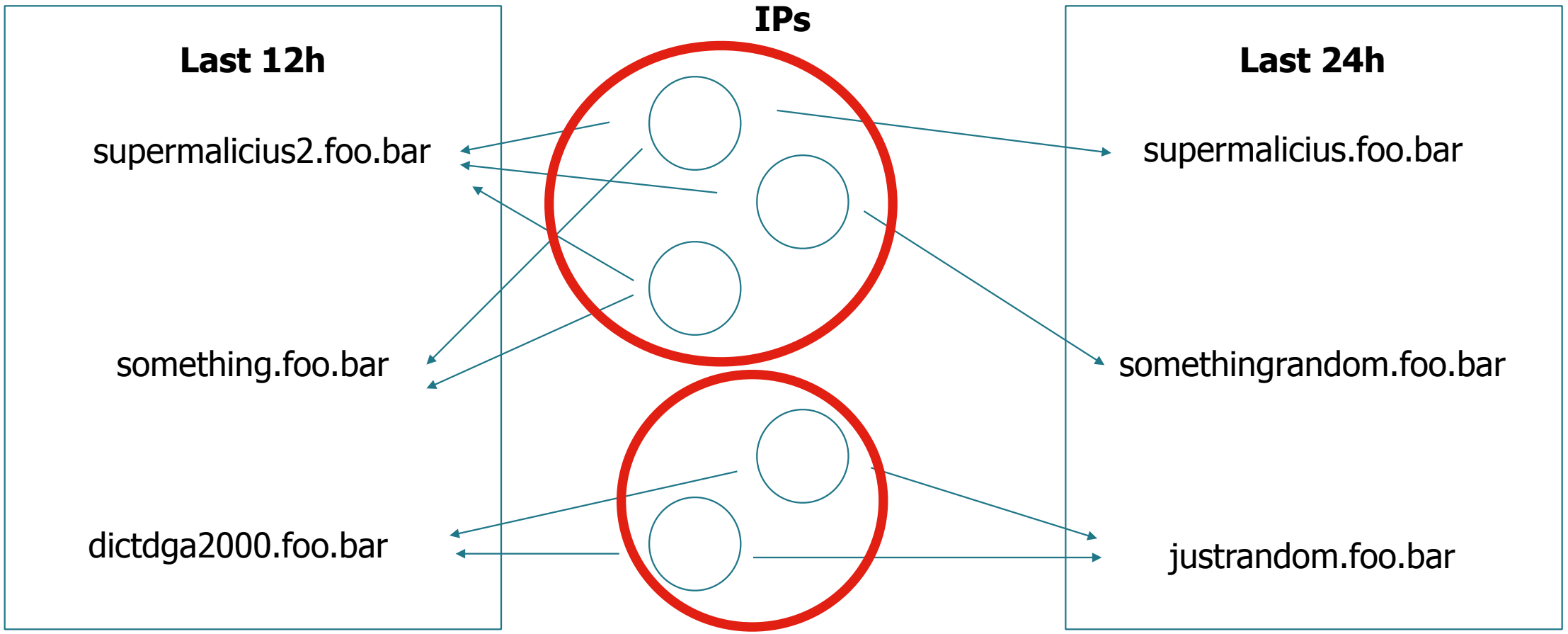
- **Infected machines**
  - Sinkholes
- **IP Reputation**
  - Mail Spike (mailspike.org)
  - Spamhaus
- **Malware Analysis**
  - Maltracker (maltracker.net)
  - Virus Total
- **Historic (Passive) DNS Information**
  - DNS Crawler

Graph Oriented Queries

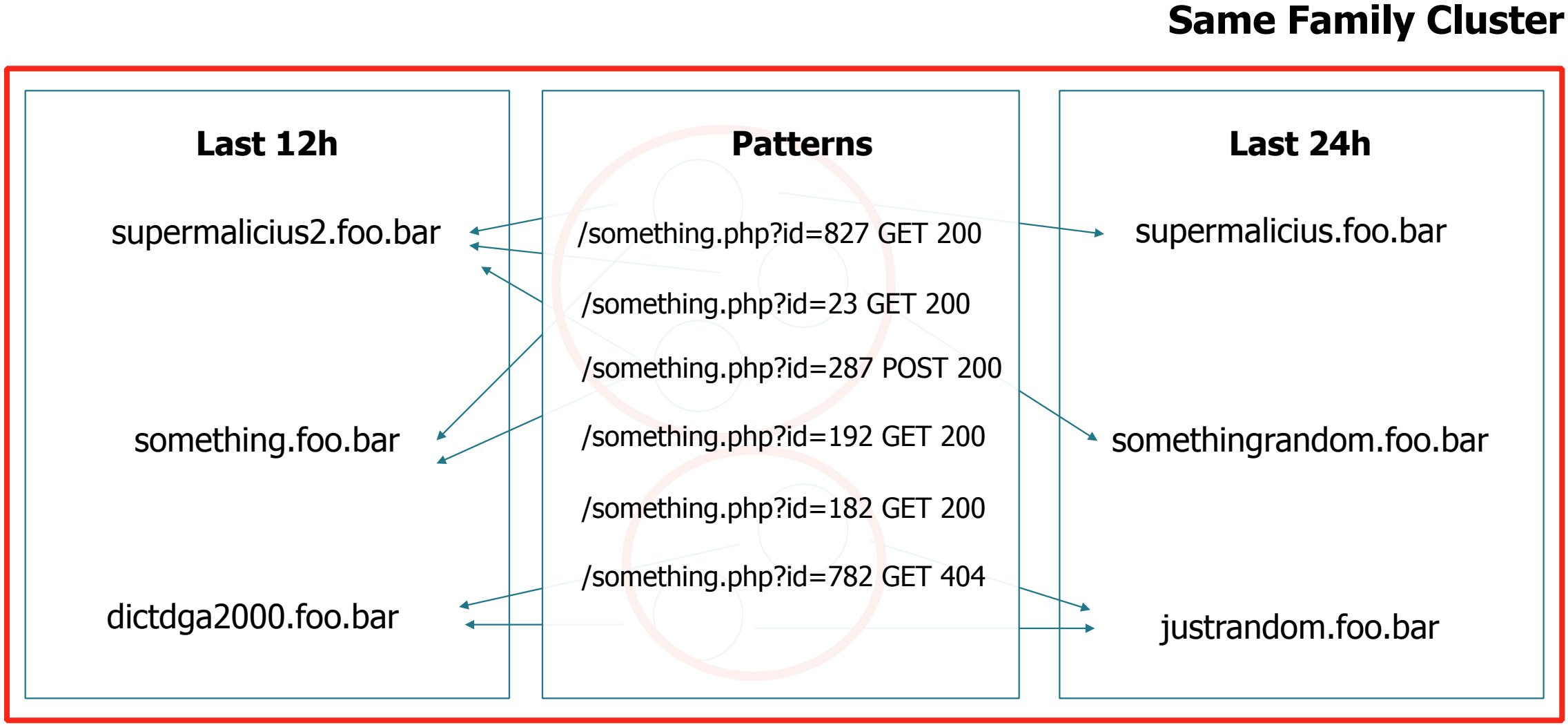
- **Discover related Domains by Graph Sub-Topologies**
  - Suspect Host <- IP(s) -> Related Domains
- **Discover new domains for a time frame …**
  - Last 6h / 12h / 24h / …
- **… looking for an even bigger time frame.**
  - Last 12h / 24h / 48h
- **Sky is the limit**
  - Interesting relations give interesting results ;)

Graph Oriented Queries

Same Campaign/Group Clusters



Graph Oriented Queries



Wrap-up

All pieces together

- **Classify DGA Like domains** -> Support Vector Machine
- **Past and Live C&C Info** -> DNS Information
- **Fast-flux and Double Fast-flux** -> DNS Information
- **DGA Rotations** -> Graph DB Queries
- **Last hour alerts** -> Graph DB Queries
- **Group Similar Traffic** -> Self Organising Map (Neural Network)
  - (discarding traffic source and destination)

All pieces together

- **Same botnet traffic** -> Graph DB Query
  - (correlate IP connections by common destinations)
- **Same botnet family traffic** -> Graph DB Query
  - (correlate IP connections by machine learning clustering process)
- **Malware Analysis** -> Direct Classification
- **IP Rep** -> Interesting Indicator

All pieces together

DEMO

Conclusion

Conclusion

Problems:

- Distinct Topologies
- Evasion Techniques
- Hu**mongo**us Traffic
- Bad actors creativity

Solutions:

- Correlate Relevant Information
  - Present VS Past
  - Real Samples
  - InfoSec Community
- Academia
  - Machine Learning
  - (your contribution)

Bonus!


Wi-Fi: On
Turn Wi-Fi Off
✓ eduroam
FBI Surveillance Van
Infect3d
INRIA-guest
It hurts when IP
Trend Micro iPhone
Join Other Network...
Create Network...
Open Network Preferences...

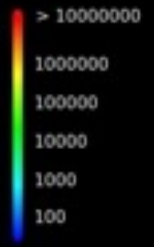
Thank You

pedro.camelo@anubisnetworks.com

joao.moura@anubisnetworks.com

@PCams

@jmgmoura



460 Million reachable IPv4 addresses observed from June 2012 to October 2012 using ICMP Ping requests and Port Scans.

Source: Carna Botnet