From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware

Manos Antonakakis

Damballa Inc.
Georgia Institute of Technology, College of Computing
Atlanta, Georgia

08/10/12, Bellevue, WA, USA


Talk Outline

- Preliminaries
  - DNS Abuse
  - Problem Statement
  - Contributions of Pleiades
- Pleiades in Depth
  - Methodology and C&C Detection
  - Results and New DGAs Discovered
  - Latest Discoveries
- Discussion
  - Consideration and Limitations

Credits

My Research Peers:

- Wenke, Roberto and Nick
- Yacin, Saeed and David

Special Thanks:

- Jeremy Demar
- Robert Edmonds


DNS Abuse and Why Should I Care?

- DNS critical Internet protocol
- Illicit operations also rely on DNS
- Important to improve ways to deal with DNS abuse
- **Limiting DNS abuse will help us limit the abuse in the Internet**

Problem Statement

Abuse (malware, DB, PPI, ect.):

- Why DNS is being exploited?
  - Agile and reliable platform
  - Take-down efforts are hard (i.e., MS/Zeus)
  - Basic economics:
    - Domain cost is small
    - Domain space very big

  Defenders: Manual Blacklisting
  Attackers: Easy to evade!
- **We can deal with the tail using Notos**

Malware-related Domain Discovery:

- Malware discovery typically behind the threat, and malware-domain discovery even slower
- DNS Hierarchy: Different levels enable different classification signal
- **We can act proactively at the authority levels using Kopis**

Open Problem: We cannot act proactively at the recursive level!

DGAs ecosystem and DGA discovery

- DGA is an alternative agile platform for modern C&C botnets
- Great success so far: Kraken, Sinowal, Srizbi, Zeus-variants etc.
- Currently, the community detects DGA bots via malware analysis
  - Reverse-engineer the DGA algorithm
  - Pre-compute future candidate C&C domains
  - Register the generated domains before botmaster
- Community can deal only with client side not server side DGAs


DGAs ecosystem and DGA discovery

- DGA is an alternative agile platform for modern C&C botnets
- Great success so far: Kraken, Sinowal, Srizbi, Zeus-variants etc.
- Currently, the community detects DGA bots via malware analysis
  - Reverse-engineer the DGA algorithm
  - Pre-compute future candidate C&C domains
  - Register the generated domains before botmaster
- Community can deal only with client side not server side DGAs
- Several the problem with this:
  - Obtaining the malware
  - Preregistering the domains (non-friendly TLDs)
  - Any binary update could alter the DGA
- Need for a light-weight **ISP-level** detector

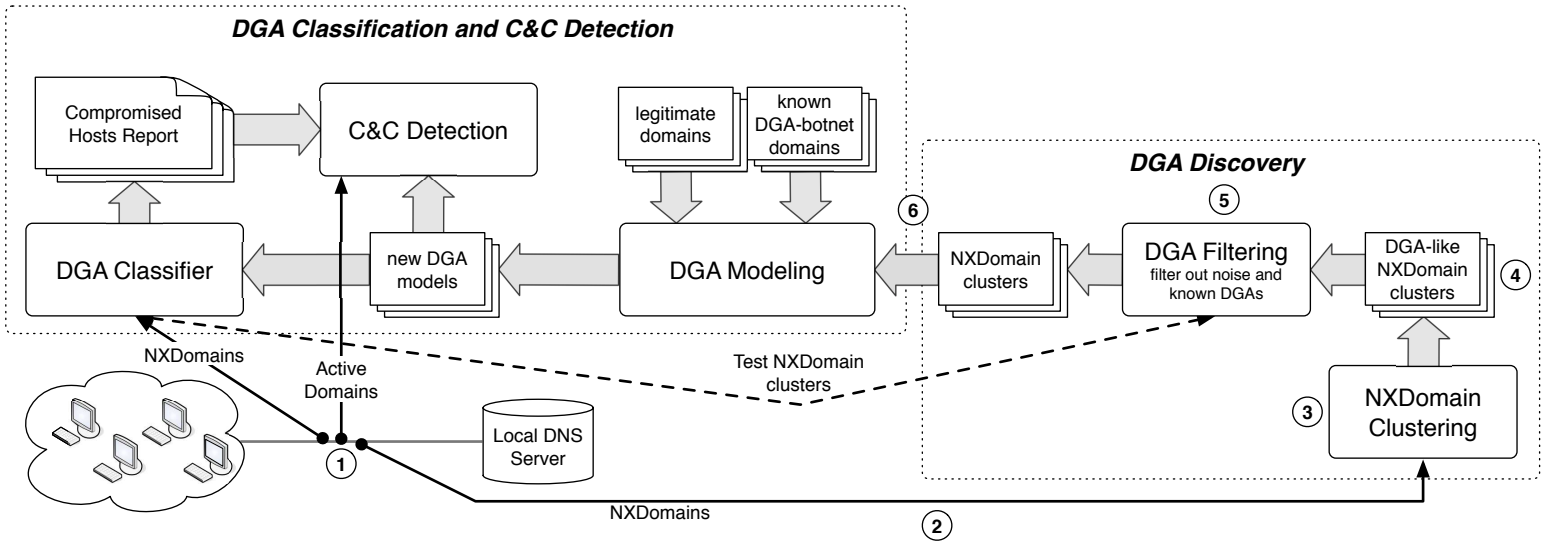

Research Contributions

- Early Malware Domain Detection at the recursive level:
  - Detect rising DGA-botnets prior the malware discovery
  - Model new DGA-bots *without* the related malware
  - Discover the active C&C for the rising DGA-bot

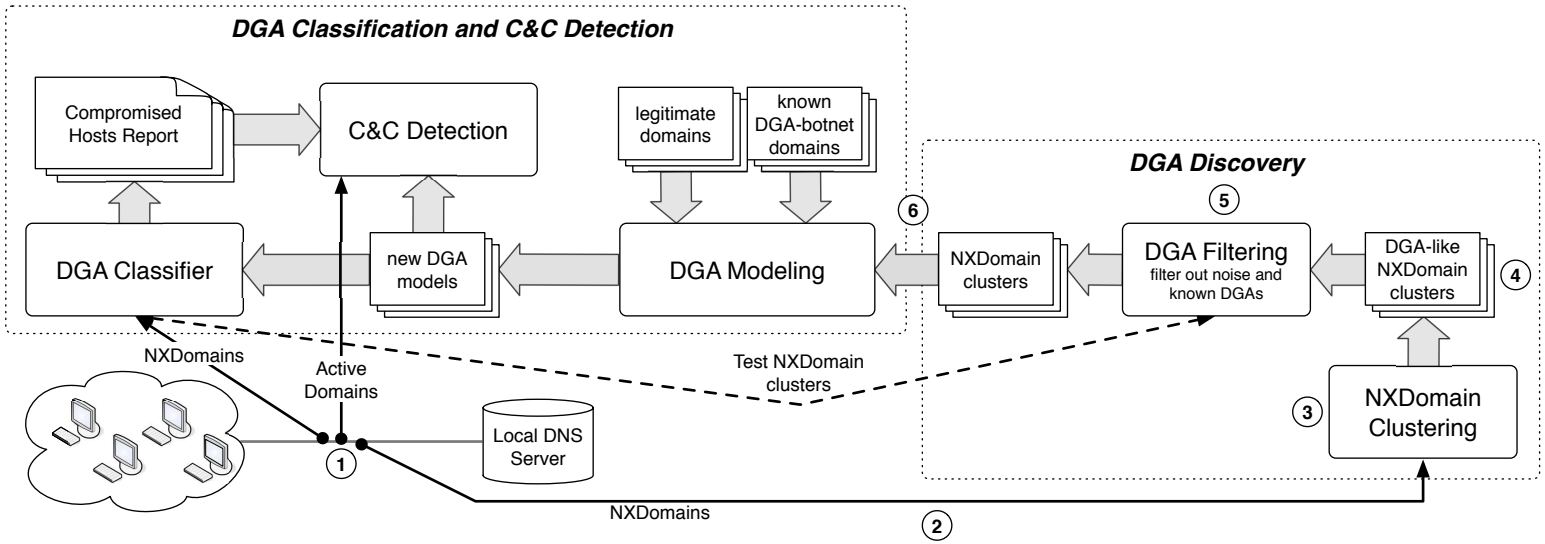
With Pleiades we can now proactively identify malware-related domain names even in the absence of corresponding malware

Looking under the hood

System Overview



System Overview



- **Pleiades**: a DGA-based botnet identification system
- Analyze streams of NXDomains at the recursive level
- Accurately detects and models (new and old) DGA-bots
- Our experimental results allowed us to discover six new DGA-based botnets
- HMM-based active C&C detector for new DGAs

Statistical Features in Pleiades

- Group NXDomains per asset with cardinality α


Statistical Features in Pleiades

- Group NXDomains per asset with cardinality α
- *n*-gram Features
  - Frequency distribution of *n*-grams across domain

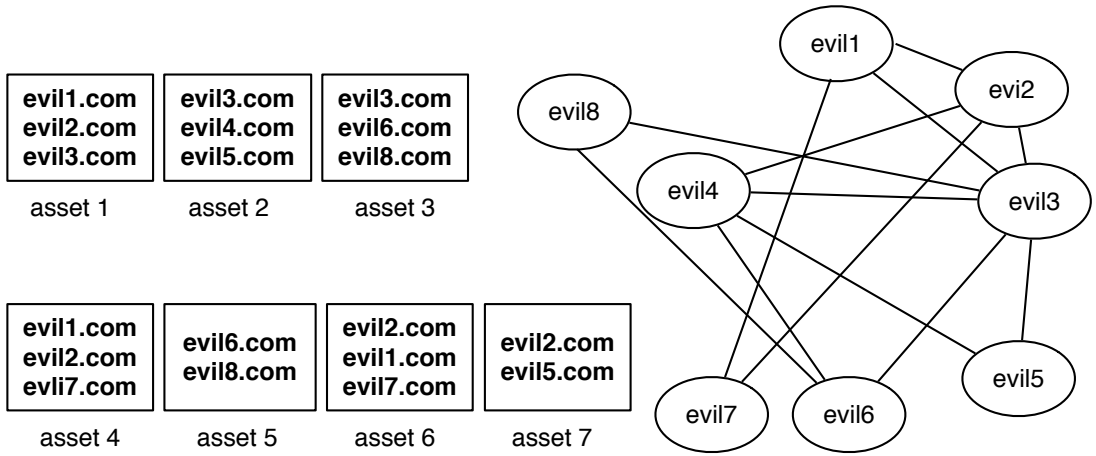
Statistical Features in Pleiades

- Group NXDomains per asset with cardinality α
- *n*-gram Features
  - Frequency distribution of *n*-grams across domain
- Entropy-based Features
  - Entropy of character distribution for separate domain levels, from the domains in the set


Statistical Features in Pleiades

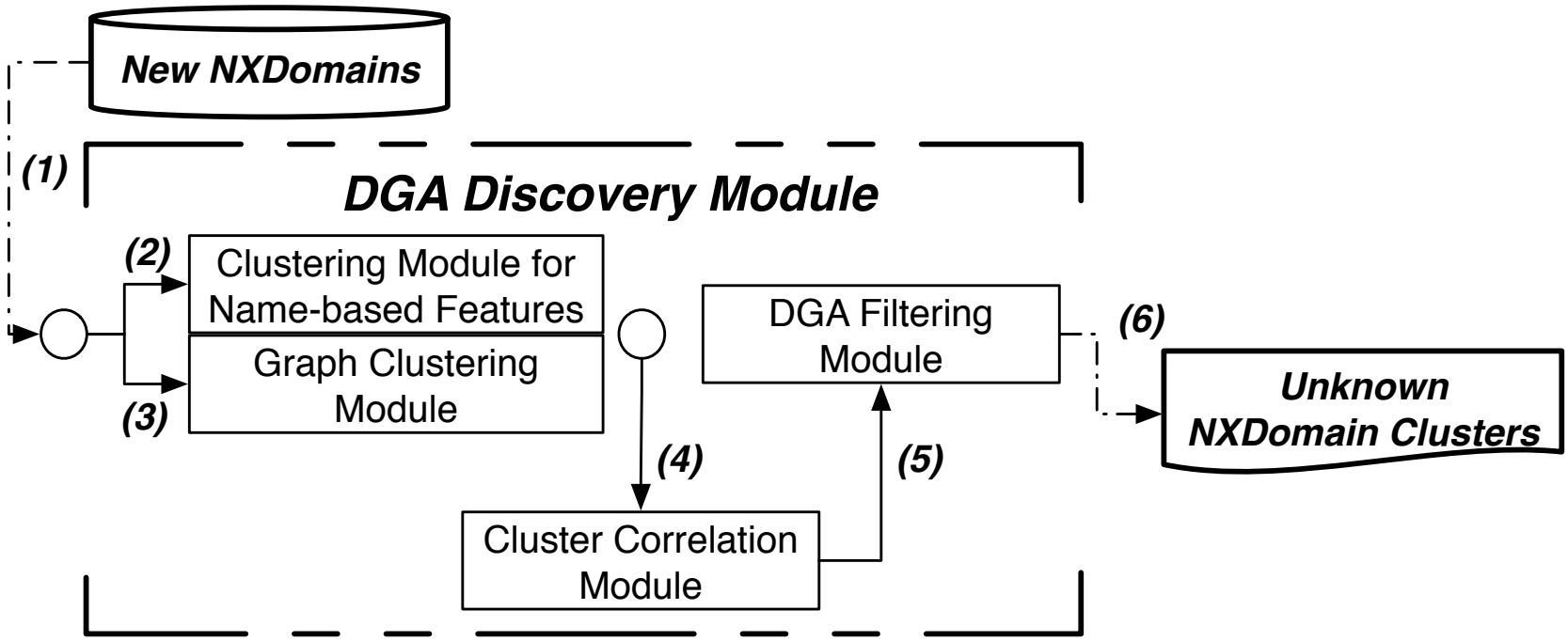
- Group NXDomains per asset with cardinality α
- *n*-gram Features
  - Frequency distribution of *n*-grams across domain
- Entropy-based Features
  - Entropy of character distribution for separate domain levels, from the domains in the set
- Structural Domain Features
  - Summarizes NXDomains structure
    - Length
    - # of unique TLDs
    - # domain levels

NXDomain Graph Construction

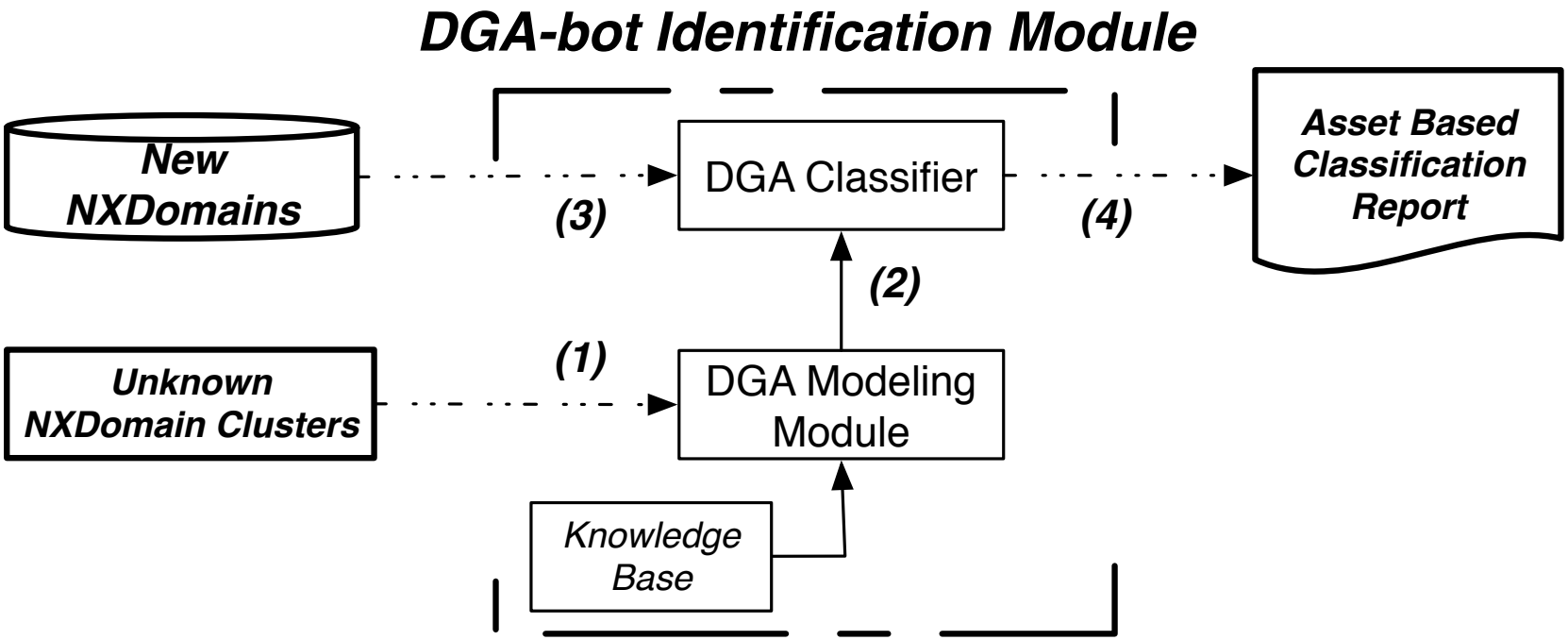


- **Weight:** Inverse document frequency
- **Intuition:** The higher the number of unique NXDomains queried by a host (or asset) the less likely the host is "representative" of the NXDomains it queries

DGA Discovery

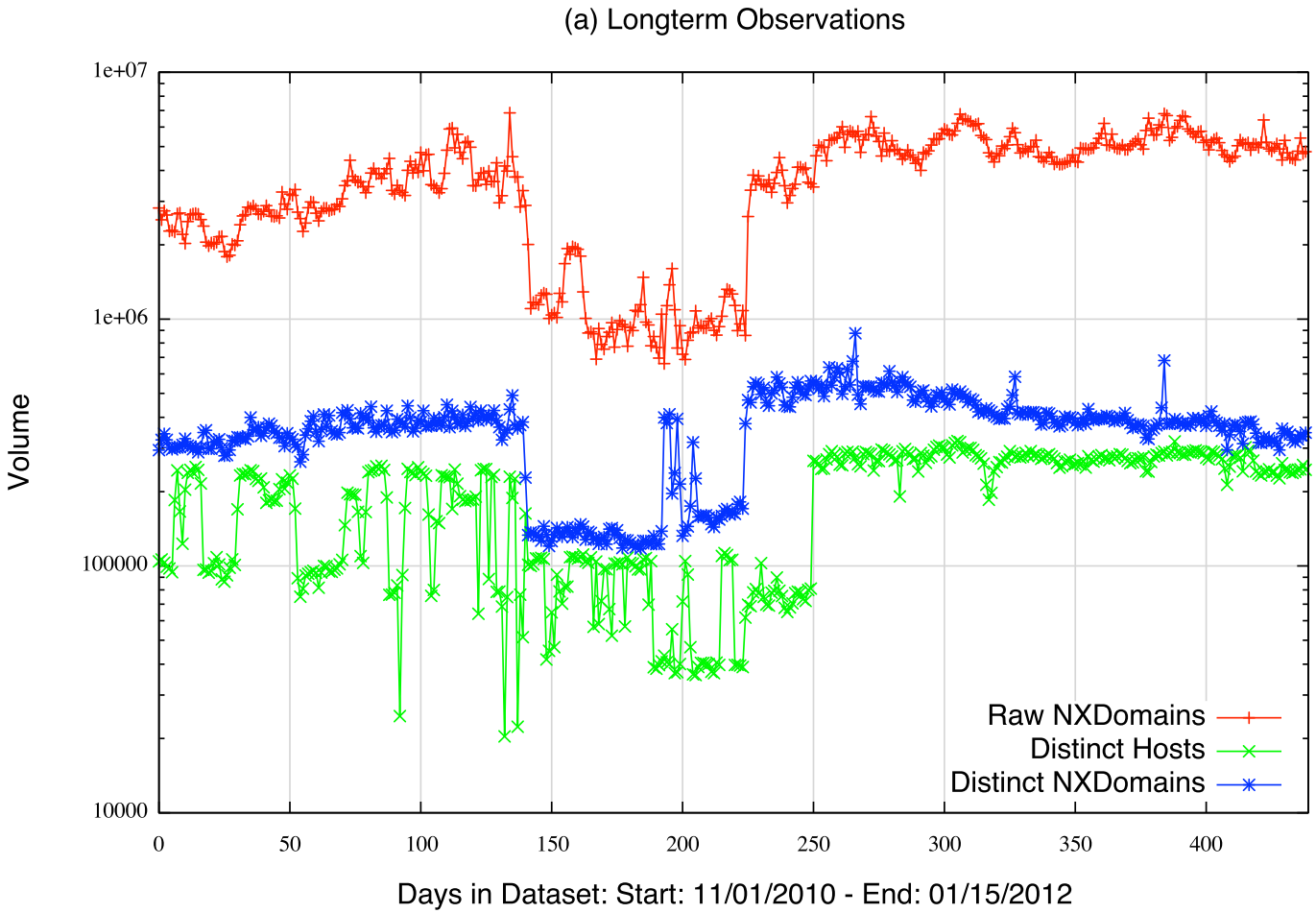


DGA Modeling



Dataset


NXDomains Traffic from ISP



15 months, 5M observations, 187K unique hosts, 360K unique NXDomains, sparse (graph) matrix size 187K $\times$ 360K

Classification Accuracy and New DGAs


DGA Modeling: 5 Classes with RF (percent)

	$\alpha = 5$ NXDomains			$\alpha = 10$ NXDomains		
Class	TP_{rate}	FP_{rate}	ROC	TP_{rate}	FP_{rate}	ROC
Bobax	95	0.4	97	**99**	**0**	**99**
Conficker	98	1.4	98	**99**	**0.1**	**99**
Sinowal	99	0.1	98	**100**	**0**	**100**
Murofet	98	0.7	98	**99**	**0.2**	**99**
Benign	96	0.7	97	**99**	**0.1**	**99**

Case study from the paper


The BankPatch Trojan

- Active since late 2010 — DGA not discovered yet
- Trojan is targeting 187 banks, credit cards and credit unions hard-coded into the binary.
- Some of the US financial entities targeted are BOA, Citi, Wellsfargo, Chase, AmericaFirst, CapitalOne, etc.

Domain Structure: 4 random characters and a domain argument. The arguments are:

seapollo.com, tomvader.com, aulmala.com, apontis.com, fnomosk.com, erhogeld.com, erobots.com, ndsontex.com, rtehedel.com, nconnect.com, edsafe.com, berhogeld.com, musallied.com, newnacion.com, susaname.com, tvolveras.com and dminmont.com.

The C&C Infrastructure for BankPatch

IP addresses	CC	Owner
146.185.250.{89-92}	RU	Petersburg Int.
31.11.43.{25-26}	RO	SC EQUILIBRIUM
31.11.43.{191-194}	RO	SC EQUILIBRIUM
46.16.240.{11-15}	UA	iNet Colocation
62.122.73.{11-14,18}	UA	“Leksim” Ltd.
87.229.126.{11-16}	HU	Webenlet Kft.
94.63.240.{11-14}	RO	Com Frecatei
94.199.51.{25-18}	HU	NET23-AS 23VNET
94.61.247.{188-193}	RO	Vatra Luminoasa
88.80.13.{111-116}	SE	PRQ-AS PeRiQuito
109.163.226.{3-5}	RO	VOXILITY-AS
94.63.149.{105-106}	RO	SC CORAL IT
94.63.149.{171-175}	RO	SC CORAL IT
176.53.17.{211-212}	TR	Radore Hosting
176.53.17.{51-56}	TR	Radore Hosting

IP addresses	CC	Owner
176.53.17.{51-56}	TR	Radore Hosting
31.210.125.{5-8}	TR	Radore Hosting
31.131.4.{117-123}	UA	LEVEL7-AS IM
91.228.111.{26-29}	UA	LEVEL7-AS IM
94.177.51.{24-25}	UA	LEVEL7-AS IM
95.64.55.{15-16}	RO	NETSERV-AS
95.64.61.{51-54}	RO	NETSERV-AS
194.11.16.133	RU	PIN-AS Petersburg
46.161.10.{34-37}	RU	PIN-AS Petersburg
46.161.29.102	RU	PIN-AS Petersburg
95.215.{0-1}.29	RU	PIN-AS Petersburg
95.215.0.{91-94}	RU	PIN-AS Petersburg
124.109.3.{3-6}	TH	SERVENET-AS-TH-AP
213.163.91.{43-46}	NL	INTERACTIVE3D-AS
200.63.41.{25-28}	PA	Panamaserver.com

Sinkhole action showed that this botnet has infected hosts in 270 different networks distributed across 25 different countries

Active today!

The latest look into BankPatch DGA

```
[ 68.35.XXX.XXX 304 0.523465703971 ]
lfnnmobama.com  hhqrmobama.com  brdsmobama.com  wmtqmobama.com  oqtnmobama.com
uiflmobama.com  wyhqmobama.com  kmtcmobama.com  sgvxmobama.com  jtulmobama.com
pzeimobama.com  qedkmobama.com  oavbmobama.com  ywfhmobama.com  gmpvmobama.com
dxvnmobama.com  wuvsmobama.com  tbtimobama.com  aybhmobama.com  qutomobama.com
xxifmobama.com  ltjwmobama.com  lhuamobama.com  iqzwmobama.com  blxvmobama.com
tnxpmobama.com  pbbmmobama.com  hhnumobama.com  zngmmobama.com  vxygmobama.com
trtnmobama.com  gmjdmobama.com  dzxtmobama.com  ltcymobama.com  uyoamobama.com
fdddmobama.com  nrwumobama.com  fpqtmobama.com  lxdbmobama.com  hlcfmobama.com
ocxrmobama.com  ucxumobama.com  eeqgmobama.com  flaymobama.com  segemobama.com
tfonmobama.com  gmtrmobama.com  ogkamobama.com  wkccmobama.com  dvlzmobama.com
rnwimobama.com  cqfmmobama.com  xdvqmobama.com  pjnmmobama.com  dxuwmobama.com
ljydmobama.com  tzfjmobama.com  ecpzmobama.com  kmnemobama.com  ffrnmobama.com
ltpumobama.com  bxgymobama.com  wmdamobama.com  briumobama.com  igvkmobama.com

[ 174.56.XXX.XXX 876 0.411764705882 ]
ggsgmobama.com  msrsmobama.com  yuxtmobama.com  ushjmobama.com  iousmobama.com
trnpmobama.com  htabmobama.com  pbjrmobama.com  kcjkmobama.com  nrmvmobama.com
iypimobama.com  wsdwmobama.com  dxvnmobama.com  aybhmobama.com  oabdmobama.com
ukmumobama.com  zrcmmobama.com  ltjwmobama.com  rztamobama.com  acjymobama.com
hhnumobama.com  awymmobama.com  kaqmmobama.com  xtdwmobama.com  oskcmobama.com
fbgzmobama.com  smaemobama.com  xvtqmobama.com  dxcjmobama.com  xfskmobama.com
eeqgmobama.com  kuvbmobama.com  bztfmobama.com  uascmobama.com  hndlmobama.com
wyhqmobama.com  iqlzmobama.com  vvmemobama.com  xtjrmobama.com  dbxrmobama.com
gsjgmobama.com  vbilmobama.com  wmdamobama.com  gclemobama.com  ysdmmobama.com
rfvpmobama.com  rpgsmobama.com  vblcmobama.com  zvmimobama.com  jrqbmobama.com
ccpumobama.com  bfudmobama.com  suvrmobama.com  jlibmobama.com  htcxmobama.com
uofqmobama.com  pnvbmobama.com  nfiemobama.com  xfmamobama.com  dnapmobama.com
oaiymobama.com  znaqmobama.com  dpbamobama.com  bvyfmobama.com  ukvgmobama.com
```

Some of the unknown at the time DGAs ...

New-DGA-v1	New-DGA-v2	New-DGA-v3
71f9d3d1.net	clfnoooqfpdc.com	uwhornfrqsdbrbnbuhjt.com
a8459681.com	slsleujrrzwx.com	epmsgxuotsciklvywmck.com
a8459681.info	qzycprhfiwfb.com	nxmglieidfsdolcakggk.com
a8459681.net	uvphgewngjiq.com	ieheckbkkkoibskrqana.com
1738a9aa.com	gxnbtlvvwmyg.com	qabgwxmkqdeixsqavxhr.com
1738a9aa.info	wdlmurglkuxb.com	gmjvfbhfcfkfyotdvbtv.com
1738a9aa.net	zzopaahxctfh.com	sajltlsbigtfexpxvsri.com
84c7e2a3.com	bzqbcftfcrqf.com	uxyjfflvoqoephfywjcq.com
84c7e2a3.info	rjvmrkkycfuh.com	kantifyosseefhdgilha.com
84c7e2a3.net	itzbkyunmzfv.com	lmklwkkrficnnqugqlpj.com

New-DGA-v4	New-DGA-v5	New-DGA-v6
semk1cquvjufayg02orednzdfg.com	zpdyaislnu.net	lymylorozig.eu
invfgg4szr22sbjbmdqm51pdtf.com	vvbmjfxpyi.net	lyvejujolec.eu
0vqbqcuqdv0i1fadodtm5iumye.com	oisbyccilt.net	xuxusujenes.eu
np1r0vnqjr3vbs3c3iqyuwe3vf.com	vgkblzdsde.net	gacezobeqon.eu
s3fhkbdu4dmc00ltmxskleeqrf.com	bxrvftzvoc.net	tufecagemyl.eu
gup1iapsm2xiedyefet21sxete.com	dlftozdnxn.net	lyvitexemod.eu
y5rk0hgujfgo0t4sfers2xolte.com	gybszkmpse.net	mavulymupiv.eu
me5oclqrfano4z0mx4qsbpdufc.com	dycsmcfwwa.net	jenokirifux.eu
jwhnr2uu3zp0ep40cttq3oyeed.com	dpwxwmkbxl.net	fotyriwavix.eu
ja4baqnv02qoxlsjxqrszdziwb.com	ttbkuogzum.net	vojugycavov.eu

Some of them are malware related: New-DGA-v1 is EnviServ.A and New-DGA-v6 is Simba-F, while others not active any more.

Closing Remarks


Pleiades in a nutshell

- DGA-based C&C discovery is very resilient mechanism
- Malware increasingly use DGAs (i.e., Zeus)
- The community cannot detect DGA bot without malware
- Key contributions:
  - Detect rising DGA-botnets before responsible malware is found
  - Accurately model DGA-bots without malware
  - Identify active C&C domains for new DGA even in the absence of corresponding malware

Limitations

- HMM C&C detection of e2LDs domains (i.e., Bonnana)
- C&C detection on triple flux networks: IP & C&C fluxing simultaneously
- DGA to MD5: Automata Induction and Grammar Inference. Hard to infer the grammar based on a few “characters” of the alphabet

Thanks! Questions?

Manos Antonakakis (manos@antonakakis.org)