

FASTER C&C DETECTION - STRATEGIES FOR FINDING ALGORITHMICALLY GENERATED DOMAIN NAMES

Malgorzata Debska

September 22, 2015

CERT Polska

LIST OF TOPICS

Introduction - what is DGA?

Malicious usage in botnets

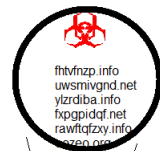
Benign DGA - false alarms in
detection systems

Current detection techniques - classification

Challenges and conclusion

INTRODUCTION - WHAT IS DGA?

ALGORITHMICALLY GENERATED DOMAIN NAMES



→ fhtvfnzp.info
← NX

→ uwsmivgnd.net
← NX

→ ylzrdiba.info
← NX



→ fxpgpidgf.net
← 82.81.120.11



DIFFERENCIES IN GENERATED DOMAINS:

- randomness of characters



DIFFERENCIES IN GENERATED DOMAINS:



- randomness of characters
- characters set

DIFFERENCIES IN GENERATED DOMAINS:



- randomness of characters
- characters set
- distribution of frequency of character usage

DIFFERENCIES IN GENERATED DOMAINS:



- randomness of characters
- characters set
- distribution of frequency of character usage
- length of generated domains

DIFFERENCIES IN GENERATED DOMAINS:



- randomness of characters
- characters set
- distribution of frequency of character usage
- length of generated domains
- level of domain generation

DIFFRENCIES IN GENERATED DOMAINS:



- randomness of characters
- characters set
- distribution of frequency of character usage
- length of generated domains
- level of domain generation
- utilized set of top level domains

EXAMPLES

tinba-dga

jmqvlmmbred2e.com

fg4zstnd3ftwh.net

qeh2p2u9pd3i1.com

ptthldqrtdt.net

dircrypt

mhrmhuxlcvkxay.com

ctskthnhq.com

safkylboxhb.com

simda

lykef.eu

qekol.eu

puzej.eu

galin.eu

EXAMPLES

tinba-dga

jmqvlmmbred2e.com
fg4zstnd3ftwh.net
qeh2p2u9pd3i1.com
pttthldqrdt.net

dircrypt

mhrmhuxlcvkxay.com
ctskthnhq.com
safkylboxhb.com

simda

lykef.eu
qekol.eu
puzej.eu
galin.eu

dyre

a3f6e2d182a40304a8874e994a294ec314.cc
b5191b0ad53da1f1fa66653610e7601856.ws
cc466dc54278d8e0fe14bdd2038b927e6f.to

gameover-zeus

1g22l018lpt4alpeypioqq24k.com
1yz3uuo1yg5zmf1u7goe81sy0xy9.net
1fhvdfa1hr7na1gu9vmv6r710j.biz
5bpzt0njqbkqlbwupc8vi3yt.org

banjori

xjsrrsensinaix.com
hlrfrsensinaix.com
antisemitismgavenuteq.com
bnmtsemitismgavenuteq.com

MALICIOUS USAGE IN BOTNETS

C&C SERVER'S NAME EXAMPLE

Every second infected host try to connect with hundreds or thousands alghoritmically generated domain name

- most of domains return NX response

```
Standard query A tmcmeffxxsr1.info.localdomain
Standard query response, No such name
Standard query A btchfxgupqyg.com
Standard query response, No such name
Standard query A btchfxgupqyg.com.localdomain
Standard query response, No such name
Standard query A oxwqodvowcgr.net
Standard query response, No such name
Standard query A oxwqodvowcgr.net.localdomain
Standard query response, No such name
Standard query A cxmihhlfwott.biz
Standard query A cxmihhlfwott.biz
Standard query response, No such name
Standard query A cxmihhlfwott.biz.localdomain
Standard query A cxmihhlfwott.biz.localdomain
Standard query response, No such name
Standard query response, No such name
Standard query A pchrqmbyeacf.ru
Standard query A pchrqmbyeacf.ru
Standard query A pchrqmbyeacf.ru
Standard query response, No such name
Standard query A pchrqmbyeacf.ru.localdomain
Standard query response, No such name
Standard query A pchrqmbyeacf.ru.localdomain
```

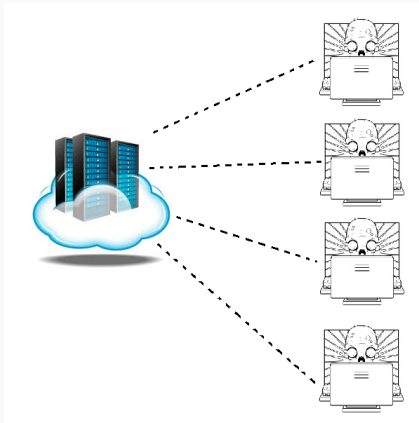
C&C SERVER'S NAME EXAMPLE

Every second infected host try to connect with hundreds or thousands alghoritmically generated domain name

```
Standard query A tmcmeffxxsr1.info.localdomain
Standard query response, No such name
Standard query A btchfxgupqyg.com
Standard query response, No such name
Standard query A btchfxgupqyg.com.localdomain
Standard query response, No such name
Standard query A oxwqodvowcgr.net
Standard query response, No such name
Standard query A oxwqodvowcgr.net.localdomain
Standard query response, No such name
Standard query A cxmihhlfwott.biz
Standard query A cxmihhlfwott.biz
Standard query response, No such name
Standard query A cxmihhlfwott.biz.localdomain
Standard query A cxmihhlfwott.biz.localdomain
Standard query response, No such name
Standard query response, No such name
Standard query A pchrqmbyeacf.ru
Standard query A pchrqmbyeacf.ru
Standard query A pchrqmbyeacf.ru
Standard query response, No such name
Standard query A pchrqmbyeacf.ru.localdomain
Standard query response, No such name
Standard query A pchrqmbyeacf.ru.localdomain
```

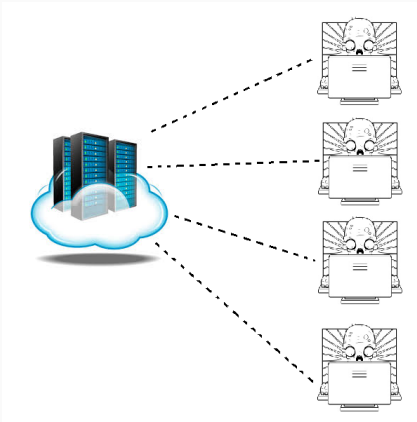
- most of domains return NX response
- attacker needs to have a couple of registered domains

- DNS communication

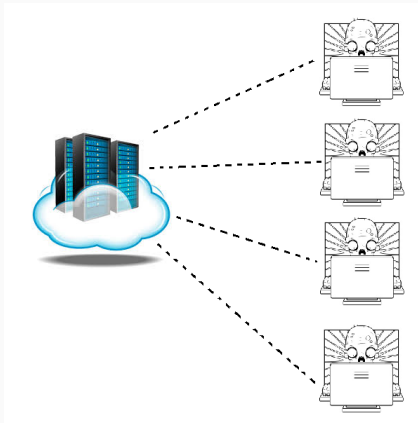


DGA BOTNET COMMUNICATION

- DNS communication
- algorithm that generates domain names

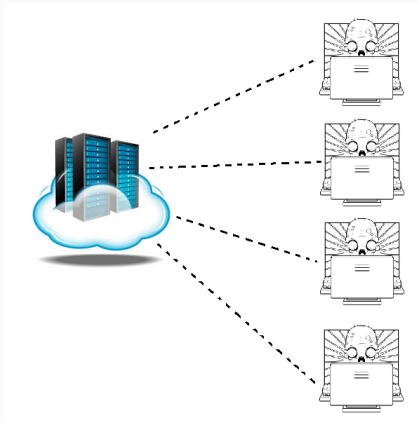


DGA BOTNET COMMUNICATION



- DNS communication
- algorithm that generates domain names
- shared seed between botmaster and clients

DGA BOTNET COMMUNICATION



- DNS communication
- algorithm that generates domain names
- shared seed between botmaster and clients
- victims search C&C server by DNS query

Is it easy to predict and sinkhole DGA domains ahead?

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)
- ...

GENERATOR'S SEED

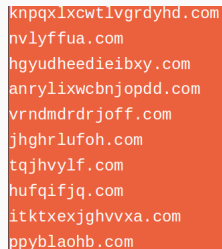
All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)
- ...

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)
- ...



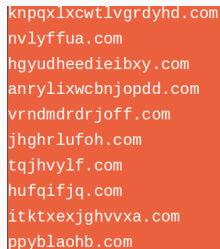
```
knpqxlxcwtlvgrdyhd.com  
nvlyffua.com  
hgyudheedieibxy.com  
anrylixwcbnjopdd.com  
vrndmdrdrjoff.com  
jhghrlufoh.com  
tqjhvylf.com  
hufqifjq.com  
itktexjghvvxa.com  
ppyblaohb.com
```

Figure 1: Ramnit

GENERATOR'S SEED

All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)
- ...



knpqxlxcwtlvgrdyhd.com
nvlyffua.com
hgyudheedieibxy.com
anrylixwcbnjopdd.com
vrndmdrdrjoff.com
jhghrlufoh.com
tqjhhvylf.com
hufqifjq.com
itktexjghvvxa.com
ppyblaohb.com

Figure 1: Ramnit

Seeds are globally consistent - victims use the same one at the same time

IS IT A SERIOUS PROBLEM? WHAT MALWARE USE DGA?

- Dyre
- GameoverZeus
- Banjori
- Matsu
- Pushdo
- Emotet
- Pykpsa
- Ramnit
- Conficker
- Bobax
- Murofet
- Necurs
- Shiotob
- Pykspa
- Cryptolocker
- Rovnix
- Emotet
- Gozi
- BankPatch
- Qakbot
- DirCrypt
- Gozi
- Flashback
- Necrus
- Ramdo

AND MORE ...

DIFFERENT TECHNIQUES BUT STILL DGA

- domain name contains random alphanumeric characters and words from dictionary

Verbs

```
seg000:7FF964F6 verbs db 'is',00h,0Ah ; DATA XREF: ; Start+CCFo
seg000:7FF964F6 db 'are',00h,0Ah
seg000:7FF964F6 db 'has',00h,0Ah
seg000:7FF964F6 db 'get',00h,0Ah
seg000:7FF964F6 db 'see',00h,0Ah
seg000:7FF964F6 db 'need',00h,0Ah
seg000:7FF964F6 db 'know',00h,0Ah
seg000:7FF964F6 db 'would',00h,0Ah
seg000:7FF964F6 db 'find',00h,0Ah
seg000:7FF964F6 db 'take',00h,0Ah
seg000:7FF964F6 db 'want',00h,0Ah
seg000:7FF964F6 db 'does',00h,0Ah
seg000:7FF964F6 db 'learn',00h,0Ah
seg000:7FF964F6 db 'become',00h,0Ah
seg000:7FF964F6 db 'come',00h,0Ah
seg000:7FF964F6 db 'include',00h,0Ah
seg000:7FF964F6 db 'thank',00h,0Ah
seg000:7FF964F6 db 'provide',00h,0Ah
seg000:7FF964F6 db 'create',00h,0Ah
seg000:7FF964F6 db 'add',00h,0Ah
seg000:7FF964F6 db 'understand',00h,0Ah
seg000:7FF964F6 db 'consider',00h,0Ah
seg000:7FF964F6 db 'choose',00h,0Ah
seg000:7FF964F6 db 'develop',00h,0Ah
seg000:7FF964F6 db 'remember',00h,0Ah
seg000:7FF964F6 db 'determine',00h,0Ah
seg000:7FF964F6 db 'grow',00h,0Ah
seg000:7FF964F6 db 'allow',00h,0Ah
seg000:7FF964F6 db 'supply',00h,0Ah
seg000:7FF964F6 db 'bring',00h,0Ah
seg000:7FF964F6 db 'improve',00h,0Ah
seg000:7FF964F6 db 'maintain',00h,0Ah
```

DIFFERENT TECHNIQUES BUT STILL DGA

- domain name contains random alphanumeric characters and words from dictionary
- names are builds from english syllables

Verbs

```
seg000:7FF964F6 verbs db 'is',00h,0Ah ; DATA XREF: ; Start+CCFo
seg000:7FF964F6 db 'are',00h,0Ah
seg000:7FF964F6 db 'has',00h,0Ah
seg000:7FF964F6 db 'get',00h,0Ah
seg000:7FF964F6 db 'see',00h,0Ah
seg000:7FF964F6 db 'need',00h,0Ah
seg000:7FF964F6 db 'know',00h,0Ah
seg000:7FF964F6 db 'would',00h,0Ah
seg000:7FF964F6 db 'find',00h,0Ah
seg000:7FF964F6 db 'take',00h,0Ah
seg000:7FF964F6 db 'want',00h,0Ah
seg000:7FF964F6 db 'does',00h,0Ah
seg000:7FF964F6 db 'learn',00h,0Ah
seg000:7FF964F6 db 'become',00h,0Ah
seg000:7FF964F6 db 'come',00h,0Ah
seg000:7FF964F6 db 'include',00h,0Ah
seg000:7FF964F6 db 'thank',00h,0Ah
seg000:7FF964F6 db 'provide',00h,0Ah
seg000:7FF964F6 db 'create',00h,0Ah
seg000:7FF964F6 db 'add',00h,0Ah
seg000:7FF964F6 db 'understand',00h,0Ah
seg000:7FF964F6 db 'consider',00h,0Ah
seg000:7FF964F6 db 'choose',00h,0Ah
seg000:7FF964F6 db 'develop',00h,0Ah
seg000:7FF964F6 db 'remember',00h,0Ah
seg000:7FF964F6 db 'determine',00h,0Ah
seg000:7FF964F6 db 'grow',00h,0Ah
seg000:7FF964F6 db 'allow',00h,0Ah
seg000:7FF964F6 db 'supply',00h,0Ah
seg000:7FF964F6 db 'bring',00h,0Ah
seg000:7FF964F6 db 'improve',00h,0Ah
seg000:7FF964F6 db 'maintain',00h,0Ah
```

BENIGN DGA - FALSE ALARMS IN DETECTION SYSTEMS

Example¹

0.0.0.0.1.0.0.4e.135jg5e1pd7s4735ftrqweufm5.avqs.mcafee.com
0.0.0.0.1.0.0.4e.13cfus2drmdq3j8cafidezr8l6.avqs.mcafee.com
0.0.0.0.1.0.0.4e.13kqas3qjj46ttkdhastkrds6.avqs.mcafee.com
0.0.0.0.1.0.0.4e.13pq3hfpunqn1d51pmvbdkk5s6.avqs.mcafee.com
0.0.0.0.1.0.0.4e.13qh71bf782qb54uzz9uhdz4mq.avqs.mcafee.com

¹DNS Noise: Measuring the Pervasiveness of Disposable Domains in Modern DNS Traffic, Yizheng Chen et al.

Example¹

```
0.0.0.0.1.0.0.4e.135jg5e1pd7s4735ftrqweufm5.avqs.mcafee.com  
0.0.0.0.1.0.0.4e.13cfus2drmdq3j8cafidezr8l6.avqs.mcafee.com  
0.0.0.0.1.0.0.4e.13kqas3qjj46ttkdhastkrds6.avqs.mcafee.com  
0.0.0.0.1.0.0.4e.13pq3hfpunqn1d51pmvbdkk5s6.avqs.mcafee.com  
0.0.0.0.1.0.0.4e.13qh71bf782qb54uzz9uhdz4mq.avqs.mcafee.com
```

This higher level domain contains basic information about the file, its hash, version of the McAfee system and information about the execution environment

¹DNS Noise: Measuring the Pervasiveness of Disposable Domains in Modern DNS Traffic, Yizheng Chen et al.

INTERNATIONALIZED DOMAIN NAME

Internationalized Domain Name (IDN)	Corresponding punycode
xn--pck3c7di8db7044cmuho9yzlar63kub1ad89aj9u	千葉市看護師求人ハローワーク
xn--gmq274an8aye736be3fesb37bke3a3cn	侍武渢村涯山市師求洵人
xn--pck3c7di8db3144cqixa1uczlr63kub1a3q9cy08a	ハローワーク静岡市看護師求人

- IDNs always begin with 'xn-' prefix

INTERNATIONALIZED DOMAIN NAME

Internationalized Domain Name (IDN)	Corresponding punycode
xn--pck3c7di8db7044cmuho9yzlar63kub1ad89aj9u	千葉県看護師求人ハローワーク
xn--gmq274an8aye736be3fesb37bke3a3cn	埼玉県村涯山市師求消湊人
xn--pck3c7di8db3144cqixa1uczlr63kub1a3q9cy08a	ハローワーク静岡県看護師求人

- IDNs always begin with 'xn-' prefix
- Now, IDNs are also used for malicious purposes

INTERNATIONALIZED DOMAIN NAME

Internationalized Domain Name (IDN)	Corresponding punycode
xn--pck3c7di8db7044cmuho9yzlar63kub1ad89aj9u	千葉市看護師求人ハローワーク
xn--gmq274an8aye736be3fesb37bke3a3cn	詩武渢村涯山市師求洵人
xn--pck3c7di8db3144cqixa1uczlr63kub1a3q9cy08a	ハローワーク静岡市看護師求人

- IDNs always begin with 'xn-' prefix
- Now, IDNs are also used for malicious purposes

INTERNATIONALIZED DOMAIN NAME

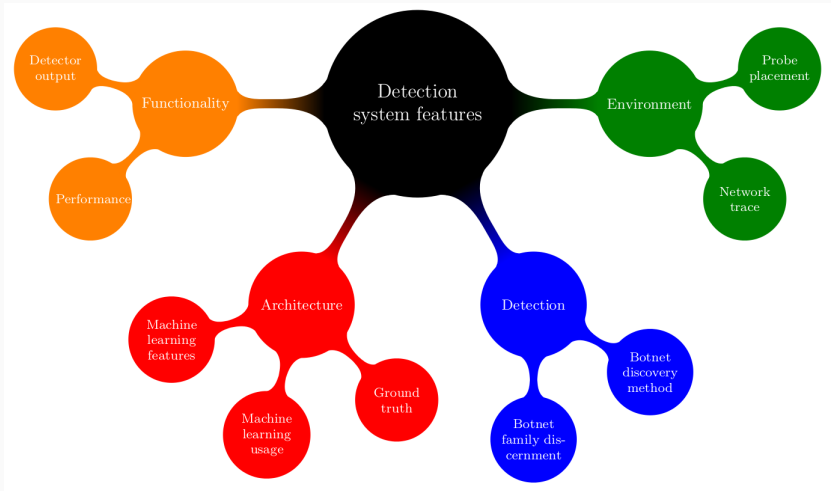
Internationalized Domain Name (IDN)	Corresponding punycode
xn--pck3c7di8db7044cmuho9yylar63kub1ad89aj9u	千葉県看護師求人ハローワーク
xn--gmq274an8aye736be3fesb37bke3a3cn	詩武渢村涯山市師求洵人
xn--pck3c7di8db3144cqixa1uczl63kub1a3q9cy08a	ハローワーク静岡市看護師求人

- IDNs always begin with 'xn--' prefix
- Now, IDNs are also used for malicious purposes

```
Queries
  xn--81by1b3b.xn--h2brj9c: type A, class IN
    Name: xn--81by1b3b.xn--h2brj9c
    [Name Length: 24]
    [Label Count: 2]
    Type: A (Host Address) (1)
    Class: IN (0x0001)
```

CURRENT DETECTION TECHNIQUES - CLASSIFICATION

DETECTION TECHNIQUES CLASSIFICATION



WHAT KIND OF DATA WE HAVE?

Environment:

- Probe placement (LAN, upper levels of DNS hierarchy)



```
doordesk_com_zone.txt - Notepad
File Edit Format View Help
$ORIGIN doordesk.com.
$TTL 1h

doordesk.com. NS ns-1395.awsdns-46.org.
doordesk.com. NS ns-710.awsdns-24.net.
doordesk.com. NS ns-11.awsdns-01.com.
doordesk.com. NS ns-1977.awsdns-55.co.uk.

doordesk.com.    A 107.20.138.116
www.doordesk.com. A 107.20.138.116
```

WHAT KIND OF DATA WE HAVE?

Environment:

- Probe placement (LAN, upper levels of DNS hierarchy)
- Network trace (NX, XD, NX+DX, only requests)



```
doordesk_com_zone.txt - Notepad
File Edit Format View Help
$ORIGIN doordesk.com.
$TTL 1h

doordesk.com. NS ns-1395.awsdns-46.org.
doordesk.com. NS ns-710.awsdns-24.net.
doordesk.com. NS ns-11.awsdns-01.com.
doordesk.com. NS ns-1977.awsdns-55.co.uk.

doordesk.com.    A 107.20.138.116
www.doordesk.com. A 107.20.138.116
```

WHAT KIND OF DATA WE HAVE?

Environment:

- Probe placement (LAN, upper levels of DNS hierarchy)
- Network trace (NX, XD, NX+DX, only requests)
- Raw text data (eg. Zone dump)



```
doordesk_com_zone.txt - Notepad
File Edit Format View Help
$ORIGIN doordesk.com.
$TTL 1h

doordesk.com. NS ns-1395.awsdns-46.org.
doordesk.com. NS ns-710.awsdns-24.net.
doordesk.com. NS ns-11.awsdns-01.com.
doordesk.com. NS ns-1977.awsdns-55.co.uk.

doordesk.com.    A 107.20.138.116
www.doordesk.com. A 107.20.138.116
```

WHAT KIND OF DATA WE HAVE?

Environment:

- Probe placement (LAN, upper levels of DNS hierarchy)
- Network trace (NX, XD, NX+DX, only requests)
- Raw text data (eg. Zone dump)



```
doordesk_com_zone.txt - Notepad
File Edit Format View Help
$ORIGIN doordesk.com.
$TTL 1h

doordesk.com. NS ns-1395.awsdns-46.org.
doordesk.com. NS ns-710.awsdns-24.net.
doordesk.com. NS ns-11.awsdns-01.com.
doordesk.com. NS ns-1977.awsdns-55.co.uk.

doordesk.com.    A 107.20.138.116
www.doordesk.com. A 107.20.138.116
```

WHAT KIND OF DATA WE HAVE?

Environment:

- Probe placement (LAN, upper levels of DNS hierarchy)
- Network trace (NX, XD, NX+DX, only requests)
- Raw text data (eg. Zone dump)



```
doordesk_com_zone.txt - Notepad
File Edit Format View Help
$ORIGIN doordesk.com.
$TTL 1h

doordesk.com. NS ns-1395.awsdns-46.org.
doordesk.com. NS ns-710.awsdns-24.net.
doordesk.com. NS ns-11.awsdns-01.com.
doordesk.com. NS ns-1977.awsdns-55.co.uk.

doordesk.com.    A 107.20.138.116
www.doordesk.com. A 107.20.138.116
```

Input data always enforce solution

GROUND TRUTH



Methods based on :

correlations

DGA features

hybrid

- analysis of DNS traffic between all hosts at network

BOTNET DETECTION - ARCHITECTURE SOLUTIONS

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	1. DNS traffic features	

BOTNET DETECTION - ARCHITECTURE SOLUTIONS

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features<ul style="list-style-type: none">• domain length	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features<ul style="list-style-type: none">• domain length• number of n-grams	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">DNS traffic features<ul style="list-style-type: none">TTL value-based featuresDNS answer-based featuresSOA recordLexical features<ul style="list-style-type: none">domain lengthnumber of n-gramscharacter's entropy	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features<ul style="list-style-type: none">• domain length• number of n-grams• character's entropy• frequency of character usage	

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features<ul style="list-style-type: none">• domain length• number of n-grams• character's entropy• frequency of character usage• domain level	

BOTNET DETECTION - ARCHITECTURE SOLUTIONS

Methods based on :

correlations	DGA features	hybrid
<ul style="list-style-type: none">• analysis of DNS traffic between all hosts at network	<ol style="list-style-type: none">1. DNS traffic features<ul style="list-style-type: none">• TTL value-based features• DNS answer-based features• SOA record2. Lexical features<ul style="list-style-type: none">• domain length• number of n-grams• character's entropy• frequency of character usage• domain level	<ul style="list-style-type: none">• mix of correlations and DGA features

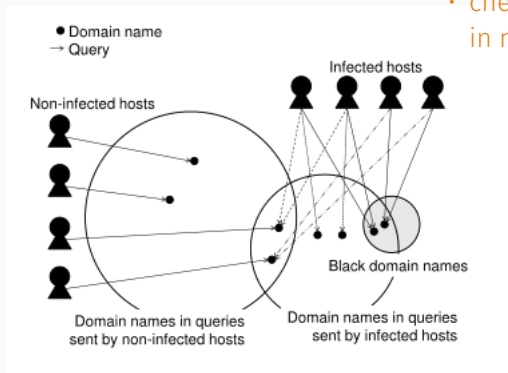
Example of method based on correlations between hosts in network is K. Sato solution in Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries, K. Sato.²

²Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries

CORRELATION METHOD

Example of method based on correlations between hosts in network is K. Sato solution in Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries, K. Sato.²

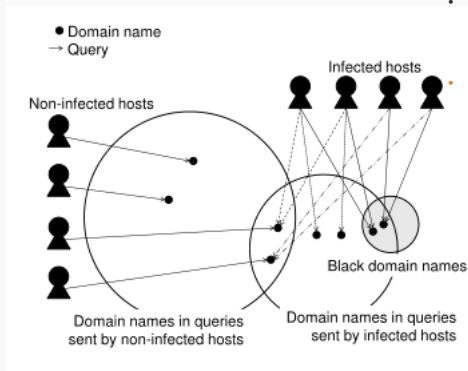
- check DNS traffic for all hosts in network



²Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries

CORRELATION METHOD

Example of method based on correlations between hosts in network is K. Sato solution in Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries, K. Sato.²

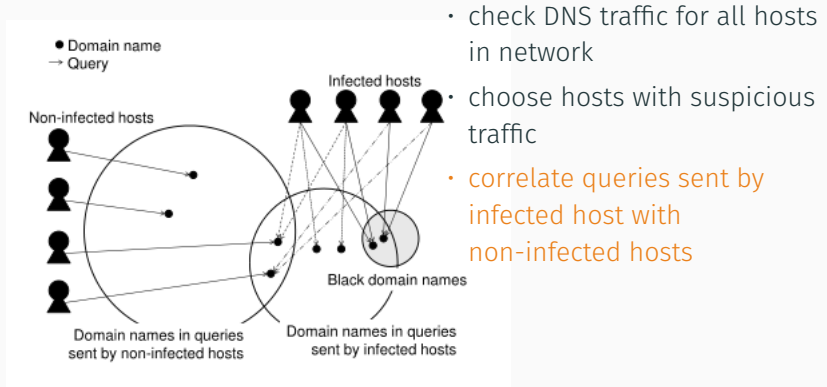


- check DNS traffic for all hosts in network
- choose hosts with suspicious traffic

²Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries

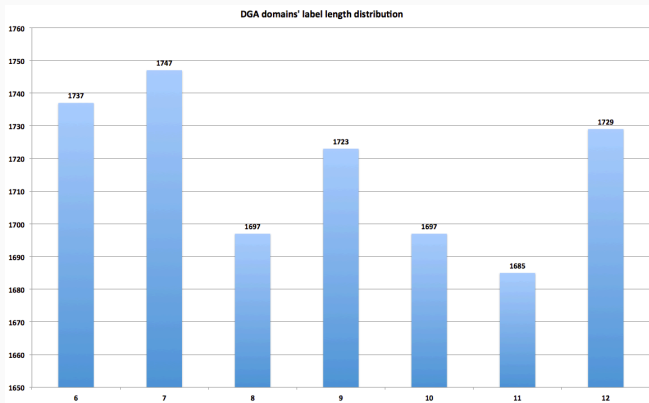
CORRELATION METHOD

Example of method based on correlations between hosts in network is K. Sato solution in Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries, K. Sato.²



²Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries

LEXICAL FEATURES



Labels' length distribution of the 12000+ DGA domains dataset ³

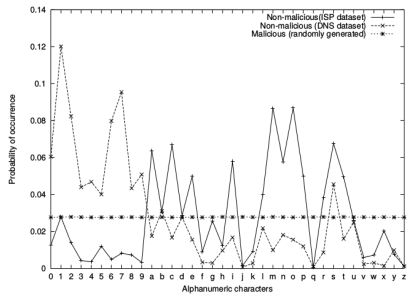
²OpenDNS Security Lab

LEXICAL FEATURES - EXAMPLE

In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.⁴

Features:

- length



(a) Non-malicious and malicious domains.

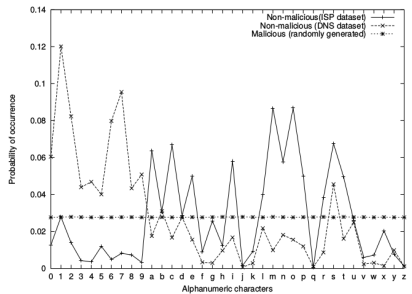
⁴Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names

LEXICAL FEATURES - EXAMPLE

In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.⁴

Features:

- length
- entropy



(a) Non-malicious and malicious domains.

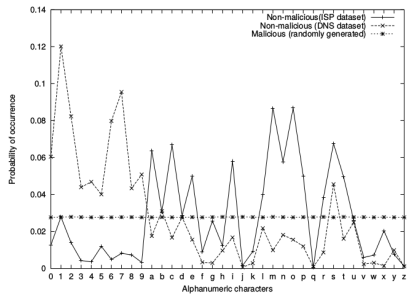
⁴Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names

LEXICAL FEATURES - EXAMPLE

In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.⁴

Features:

- length
- entropy
- K-L divergence



(a) Non-malicious and malicious domains.

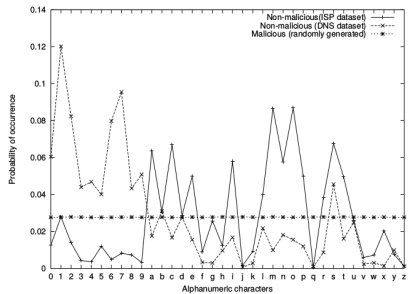
⁴Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names

LEXICAL FEATURES - EXAMPLE

In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.⁴

Features:

- length
- entropy
- K-L divergence
- Jaccard index between bigrams



(a) Non-malicious and malicious domains.

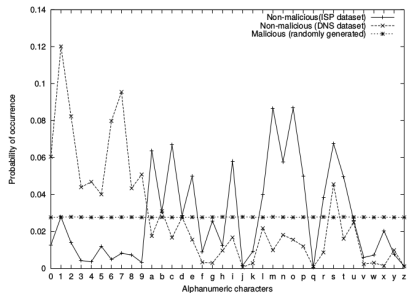
⁴Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names

LEXICAL FEATURES - EXAMPLE

In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.⁴

Features:

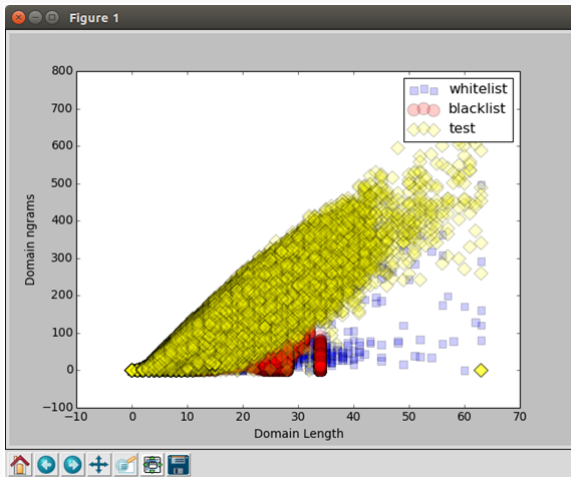
- length
- entropy
- K-L divergence
- Jaccard index between bigrams
- Edit distance



(a) Non-malicious and malicious domains.

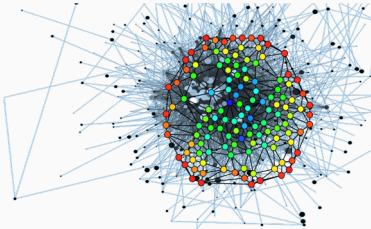
⁴Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names

LEXICAL FEATURES RATIOS COMPARISON



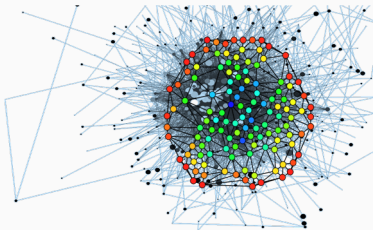
Domains length vs ngrams frequency in PL zone

Regardless of which set of features we chose (lexical or based on DNS traffic), classifier needs to create detection model.



- it is based on ground truth

Regardless of which set of features we chose (lexical or based on DNS traffic), classifier needs to create detection model.



- it is based on ground truth
- most algorithms based on genetic algorithms or decision trees: SVM, Random Forests, J48 etc.

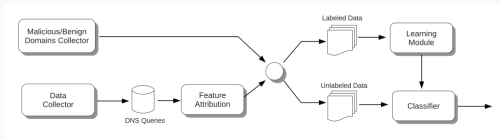
EXAMPLE OF MACHINE LEARNING USE

Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵

⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis

EXAMPLE OF MACHINE LEARNING USE

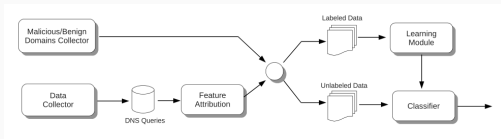
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵



⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis

EXAMPLE OF MACHINE LEARNING USE

Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵



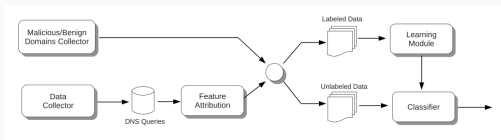
- count features ratios based on ground truth

Feature Set	#	Feature Name
Time-Based Features	1	Short life
	2	Daily similarity
	3	Repeating patterns
	4	Access ratio
DNS Answer-Based Features	5	Number of distinct IP addresses
	6	Number of distinct countries
	7	Number of domains share the IP with
	8	Reverse DNS query results
TTL Value-Based Features	9	Average TTL
	10	Standard Deviation of TTL
	11	Number of distinct TTL values
	12	Number of TTL change
	13	Percentage usage of specific TTL ranges
Domain Name-Based Features	14	% of numerical characters
	15	% of the length of the LMS

⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis

EXAMPLE OF MACHINE LEARNING USE

Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵



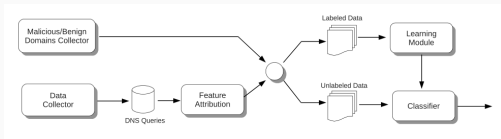
- count features ratios based on ground truth
- create training model (J48)

Feature Set	#	Feature Name
Time-Based Features	1	Short life
	2	Daily similarity
	3	Repeating patterns
	4	Access ratio
DNS Answer-Based Features	5	Number of distinct IP addresses
	6	Number of distinct countries
	7	Number of domains share the IP with
	8	Reverse DNS query results
TTL Value-Based Features	9	Average TTL
	10	Standard Deviation of TTL
	11	Number of distinct TTL values
	12	Number of TTL change
	13	Percentage usage of specific TTL ranges
Domain Name-Based Features	14	% of numerical characters
	15	% of the length of the LMS

⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis

EXAMPLE OF MACHINE LEARNING USE

Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵



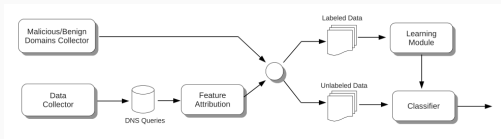
- count features ratios based on ground truth
- create training model (J48)
- count features ratios for host

Feature Set	#	Feature Name
Time-Based Features	1	Short life
	2	Daily similarity
	3	Repeating patterns
	4	Access ratio
DNS Answer-Based Features	5	Number of distinct IP addresses
	6	Number of distinct countries
	7	Number of domains share the IP with
	8	Reverse DNS query results
TTL Value-Based Features	9	Average TTL
	10	Standard Deviation of TTL
	11	Number of distinct TTL values
	12	Number of TTL change
	13	Percentage usage of specific TTL ranges
Domain Name-Based Features	14	% of numerical characters
	15	% of the length of the LMS

⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis

EXAMPLE OF MACHINE LEARNING USE

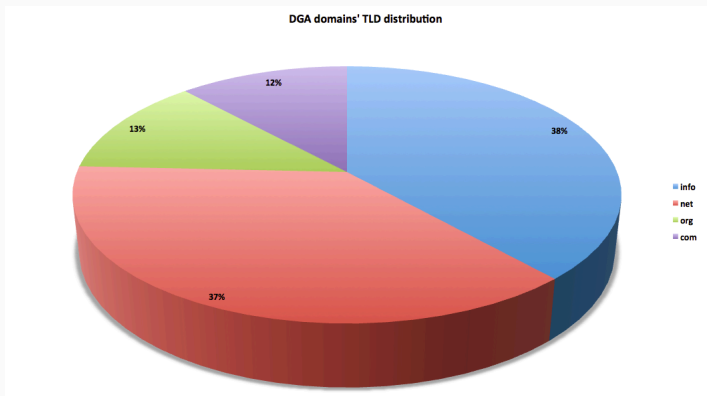
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.⁵



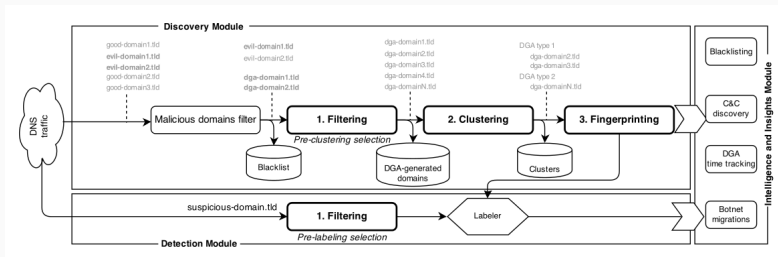
- count features ratios based on ground truth
- create training model (J48)
- count features ratios for host
- classify host by comparing host ratios with training model

Feature Set	#	Feature Name
Time-Based Features	1	Short life
	2	Daily similarity
	3	Repeating patterns
	4	Access ratio
DNS Answer-Based Features	5	Number of distinct IP addresses
	6	Number of distinct countries
	7	Number of domains share the IP with
	8	Reverse DNS query results
TTL Value-Based Features	9	Average TTL
	10	Standard Deviation of TTL
	11	Number of distinct TTL values
	12	Number of TTL change
	13	Percentage usage of specific TTL ranges
Domain Name-Based Features	14	% of numerical characters
	15	% of the length of the LMS

⁵Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis



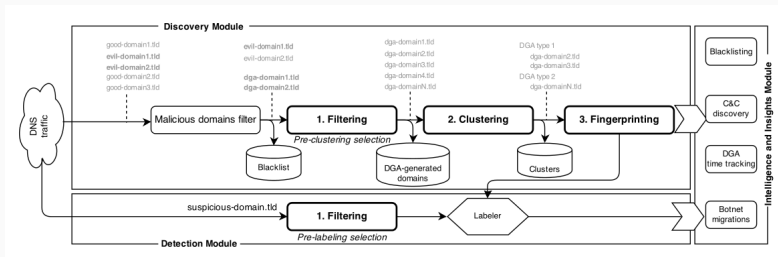
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

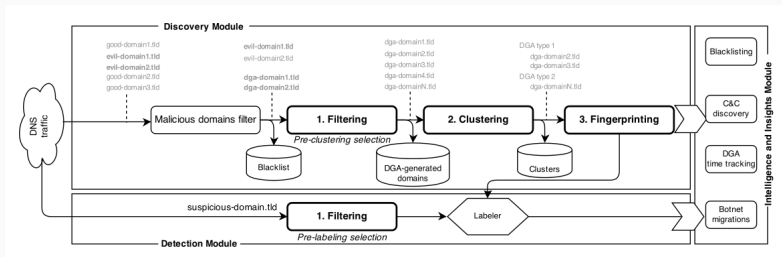
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

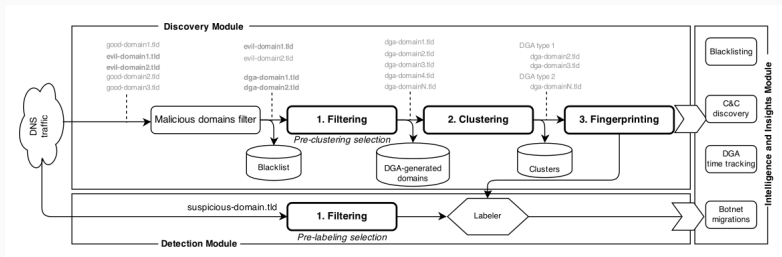
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio
 2. n-gram normality score

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

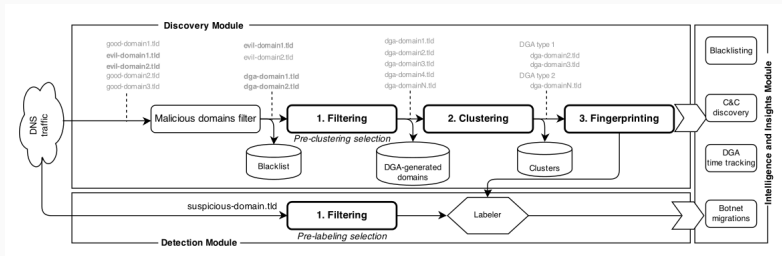
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio
 2. n-gram normality score
 3. statistical linguistic ratios

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

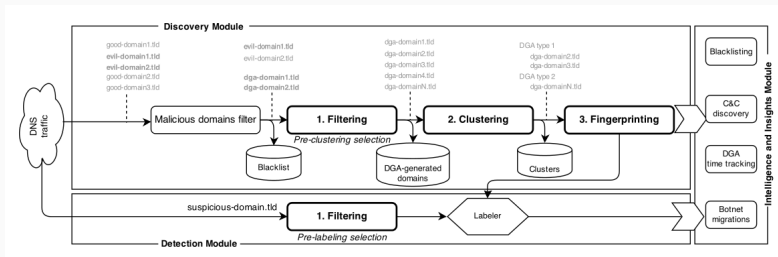
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio
 2. n-gram normality score
 3. statistical linguistic ratios
- Clustering using bipartite graph recursive

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

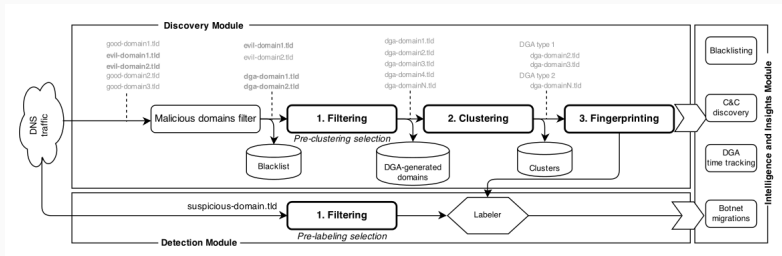
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio
 2. n-gram normality score
 3. statistical linguistic ratios
- Clustering using bipartite graph recursive
 1. IP address features

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

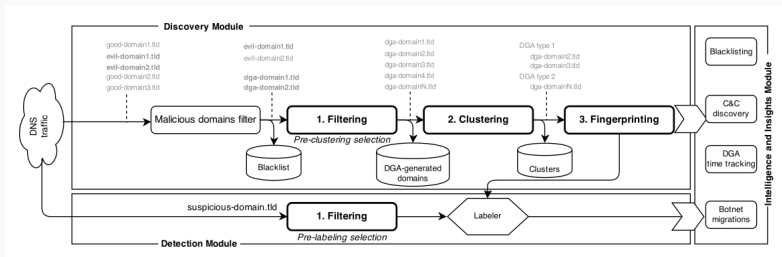
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁶



- Filtering XD domains - lexical features analysis
 1. meaningful characters ratio
 2. n-gram normality score
 3. statistical linguistic ratios
- Clustering using bipartite graph recursive
 1. IP address features
 2. DBSCAN algorithm

⁶Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence

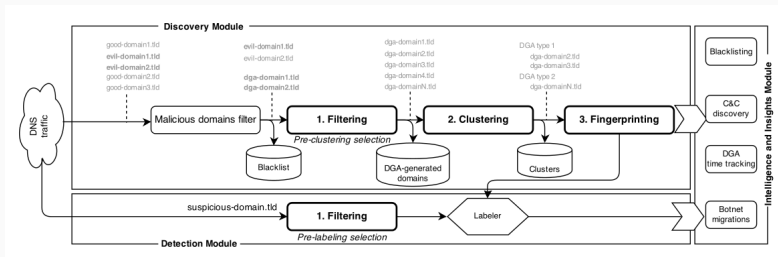
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷



- Fingerprinting

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

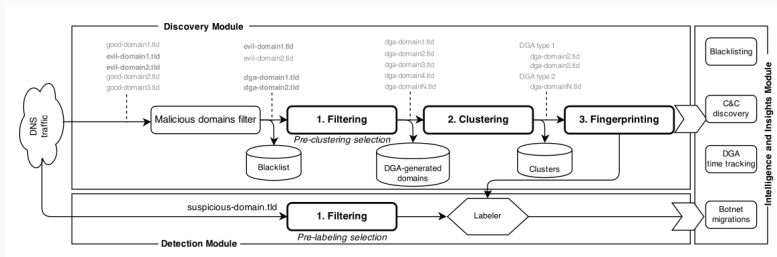
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷



- Fingerprinting
 - 1. C&C servers IP addresses

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷



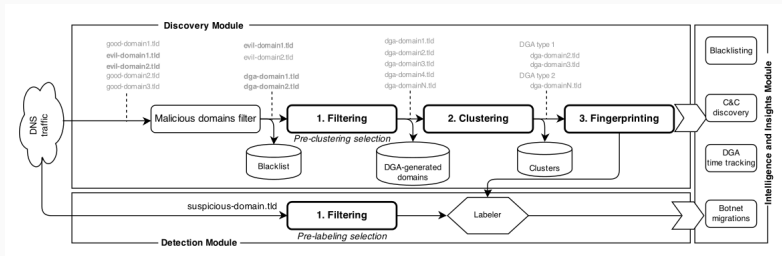
- Fingerprinting

1. C&C servers IP addresses

2. length of the shortest and longest domain name

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷

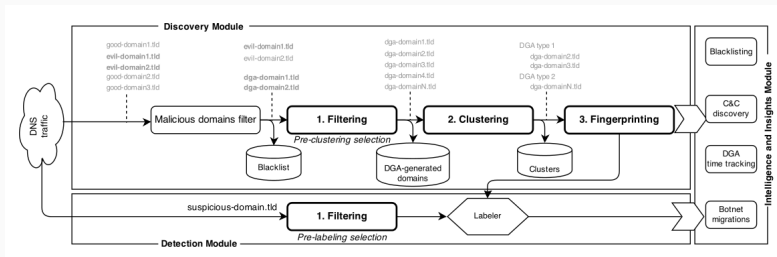


- Fingerprinting

1. C&C servers IP addresses
2. length of the shortest and longest domain name
3. utilized character set

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷

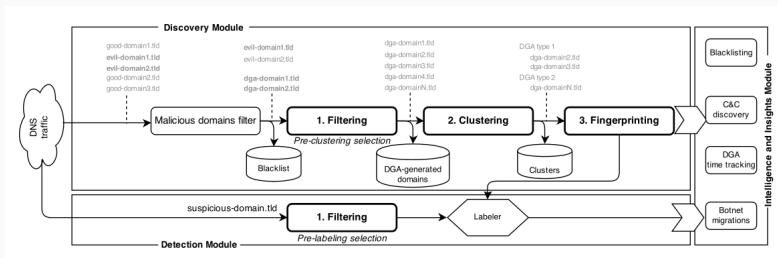


- Fingerprinting

1. C&C servers IP addresses
2. length of the shortest and longest domain name
3. utilized character set
4. number of numerical characters in chosen prefix of domain from cluster

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni⁷



- Fingerprinting

1. C&C servers IP addresses
2. length of the shortest and longest domain name
3. utilized character set
4. number of numerical characters in chosen prefix of domain from cluster
5. set of TLDs used in cluster

⁷Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:

DETECTION RESULTS

Phoenix - hybrid method

FPR - ?

TPR - 81.4-94.8

Lexical method

FPR - 0.3-0.8

TPR - 83.3-100.0

Exposure - machine learning

FPR - 0.3-1.1

TPR - 98.4-99.5

Correlation method

FPR - 0.5

TPR - 80.0

As we see there are two general properties that define each DGA:

As we see there are two general properties that define each DGA:

- predictability

As we see there are two general properties that define each DGA:

- predictability
- time-dependence

As we see there are two general properties that define each DGA:

- predictability
- time-dependence

CURRENT DETECTION AND PROTECTION TECHNIQUES

As we see there are two general properties that define each DGA:

- predictability
- time-dependence

No detection system can predict all generated domains without malware's algorithm and seed, yet.



CHALLENGES AND CONCLUSION

QUESTIONS?