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

Malgorzata Debska
September 22, 2015
CERT Polska
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

- ftvfnzp.info
- uwsnvgnnd.net
- ylrddba.info
- fxpgpidgf.net
- rawftqfzxy.info
- 82.81.120.11
DIFFRENCIES IN GENERATED DOMAINS:

• randomness of characters
DIFFRENCIES IN GENERATED DOMAINS:

- randomness of characters
- characters set
DIFFRENCIES IN GENERATED DOMAINS:

- randomness of characters
- characters set
- distribution of frequency of character usage
DIFFRENCIES IN GENERATED DOMAINS:

- randomness of characters
- characters set
- distribution of frequency of character usage
- length of generated domains
DIFFRENCIES 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
<table>
<thead>
<tr>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>tinba-dga</td>
</tr>
<tr>
<td>jmqvlmbred2e.com</td>
</tr>
<tr>
<td>fg4zstnd3ftwh.net</td>
</tr>
<tr>
<td>qeh2p2u9pd3i1.com</td>
</tr>
<tr>
<td>pttthldqrdrd.net</td>
</tr>
<tr>
<td>dircrypt</td>
</tr>
<tr>
<td>mhrmhuxlcvkxay.com</td>
</tr>
<tr>
<td>ctskthnhq.com</td>
</tr>
<tr>
<td>safkylboxhb.com</td>
</tr>
<tr>
<td>simda</td>
</tr>
<tr>
<td>lykef.eu</td>
</tr>
<tr>
<td>qekol.eu</td>
</tr>
<tr>
<td>puzej.eu</td>
</tr>
<tr>
<td>galin.eu</td>
</tr>
</tbody>
</table>
### EXAMPLES

<table>
<thead>
<tr>
<th>tinba-dga</th>
<th>dyre</th>
</tr>
</thead>
<tbody>
<tr>
<td>jmqvlmmbred2e.com</td>
<td>a3f6e2d182a40304a8874e994a294ec314.cc</td>
</tr>
<tr>
<td>fg4zstnd3ftwh.net</td>
<td>b5191b0ad53da1f1fa66653610e7601856.ws</td>
</tr>
<tr>
<td>qeh2p2u9pd3i1.com</td>
<td>cc466dc54278d8e0fe14bdd2038b927e6f.to</td>
</tr>
<tr>
<td>pttthldqrdt.net</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dircrypt</th>
<th>gameover-zeus</th>
</tr>
</thead>
<tbody>
<tr>
<td>mhrmhuxlcvxxay.com</td>
<td>1g22l018lpt4alpeypioqq24k.com</td>
</tr>
<tr>
<td>ctskthnhqv.com</td>
<td>1yz3uu01yg5zmf1u7goe81sy0xy9.net</td>
</tr>
<tr>
<td>safkylboxhb.com</td>
<td>1fhvdfia1hr7na1gu9vmv6r710j.biz</td>
</tr>
<tr>
<td></td>
<td>5bpzt0njqbkqlbwupc8vi3yt.org</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>simda</th>
<th>banjori</th>
</tr>
</thead>
<tbody>
<tr>
<td>lykef.eu</td>
<td>xjsrrrsensinaix.com</td>
</tr>
<tr>
<td>qekol.eu</td>
<td>hlrfrsensinaix.com</td>
</tr>
<tr>
<td>puzej.eu</td>
<td>antisemitismgavenuteq.com</td>
</tr>
<tr>
<td>galin.eu</td>
<td>bnmtsemitismgavenuteq.com</td>
</tr>
</tbody>
</table>
MALICIOUS USAGE IN BOTNETS
Every second infected host try to connect with hundreds or thousands algorithmically generated domain name

- most of domains return NX response
Every second infected host try to connect with hundreds or thousands algorithmically generated domain name

- most of domains return NX response
- attacker needs to have a couple of registered domains
• DNS communication
- DNS communication
- Algorithm that generates domain names
• DNS communication
• algorithm that generates domain names
• shared seed between botmaster and clients
• 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?
All domains generated algorithmically are dependent on specified seed.
All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker, GameOverZeus)
- currently trending Twitter hashtag (Torpig)
All domains generated algorithmically are dependent on specified seed

- date (CryptoLocker, Conficker,GameOverZeus)
- currently trending Twitter hashtag (Torpig)
- seed hardcoded in infected file (Tinba)
All domains generated algorithmically are dependent on specified seed

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

Figure 1: Ramnit Seeds are globally consistent - victims use the same one at the same time
All domains generated algorithmically are dependent on specified seed

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

Figure 1: Ramnit Seeds are globally consistent - victims use the same one at the same time
All domains generated algorithmically are dependent on specified seed

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

Figure 1: Ramnit
All domains generated algorithmically are dependent on specified seed

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

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
DIFFERENT TECHNIQUES BUT STILL DGA

- domain name contains random alphanumeric characters and words from dictionary
- names are builds from english syllables
BENIGN DGA - FALSE ALARMS IN DETECTION SYSTEMS
### REQUESTS OF AV TOOLS

**Example**

<table>
<thead>
<tr>
<th>IP Address</th>
<th>Domain Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0.0.0.1.0.0.4e.135jg5e1pd7s4735ftrqweufm5.avqs.mcafee.com</td>
<td></td>
</tr>
<tr>
<td>0.0.0.0.1.0.0.4e.13cfus2drmdq3j8cafidezr8l6.avqs.mcafee.com</td>
<td></td>
</tr>
<tr>
<td>0.0.0.0.1.0.0.4e.13kqas3qjj46ttkdhastkrdsv6.avqs.mcafee.com</td>
<td></td>
</tr>
<tr>
<td>0.0.0.0.1.0.0.4e.13pq3hfpunqn1d51pmvbdkk5s6.avqs.mcafee.com</td>
<td></td>
</tr>
<tr>
<td>0.0.0.0.1.0.0.4e.13qh71bf782qb54uzz9uhdz4mq.avqs.mcafee.com</td>
<td></td>
</tr>
</tbody>
</table>

---

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

Example

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0.0.0.1.0.0.4e.135jg5e1pd7s4735ftrqweufm5.avqs.mcafee.com</td>
<td>0.0.0.0.1.0.0.4e.13cfus2drmdq3j8cafidezr8l6.avqs.mcafee.com</td>
<td>0.0.0.0.1.0.0.4e.13kqas3qjj46ttkdhostkrdsv6.avqs.mcafee.com</td>
<td>0.0.0.0.1.0.0.4e.13pq3hfpunqn1d51pmvbdkk5s6.avqs.mcafee.com</td>
<td>0.0.0.0.1.0.0.4e.13qh71bf782qb54uzz9uhdz4mq.avqs.mcafee.com</td>
</tr>
</tbody>
</table>

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.
• IDNs always begin with ‘xn–’ prefix
• IDNs always begin with ‘xn–’ prefix
• Now, IDNs are also used for malicious purposes
• IDNs always begin with 'xn–' prefix
• Now, IDNs are also used for malicious purposes
INTERNATIONALIZED DOMAIN NAME

- IDNs always begin with ‘xn–’ prefix
- Now, IDNs are also used for malicious purposes
CURRENT DETECTION TECHNIQUES - CLASSIFICATION
WHAT KIND OF DATA WE HAVE?

Environment:

• Probe placement (LAN, upper levels of DNS hierarchy)
WHAT KIND OF DATA WE HAVE?

Environment:

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

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

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

Input data always enforce solution
ARCHITECTURE

GROUND TRUTH

WHITELIST

BLACKLIST
## Botnet Detection - Architecture Solutions

Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• analysis of DNS traffic between all hosts at network</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
</tbody>
</table>
**BOTNET DETECTION - ARCHITECTURE SOLUTIONS**

Methods based on:

<table>
<thead>
<tr>
<th>Correlations</th>
<th>DGA Features</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
</tbody>
</table>

21
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
</table>
| • analysis of DNS traffic between all hosts at network | 1. DNS traffic features  
• TTL value-based features  
• DNS answer-based features | hybrid |
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations</td>
<td>Analysis of DNS traffic between all hosts at network</td>
</tr>
<tr>
<td>DGA features</td>
<td>1. DNS traffic features</td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
</tr>
<tr>
<td>Hybrid</td>
<td>2. Lexical features</td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Lexical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• domain length</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td>hybrid</td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
<td></td>
</tr>
<tr>
<td>2. Lexical features</td>
<td>2. Lexical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• domain length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• number of n-grams</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
</table>
| • analysis of DNS traffic between all hosts at network | 1. DNS traffic features  
• TTL value-based features  
• DNS answer-based features  
• SOA record | |
| | 2. Lexical features  
• domain length  
• number of n-grams  
• character’s entropy | |
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
<td></td>
</tr>
<tr>
<td>2. Lexical features</td>
<td>2. Lexical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• domain length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• number of n-grams</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• character’s entropy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• frequency of character usage</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>correlations</th>
<th>DGA features</th>
<th>hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>- analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- SOA record</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Lexical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- domain length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- number of n-grams</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- character’s entropy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- frequency of character usage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- domain level</td>
<td></td>
</tr>
</tbody>
</table>
Methods based on:

<table>
<thead>
<tr>
<th>Correlations</th>
<th>DGA Features</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>• analysis of DNS traffic between all hosts at network</td>
<td>1. DNS traffic features</td>
<td>• mix of correlations and DGA features</td>
</tr>
<tr>
<td></td>
<td>• TTL value-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• DNS answer-based features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SOA record</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Lexical features</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• domain length</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• number of n-grams</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• character’s entropy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• frequency of character usage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• domain level</td>
<td></td>
</tr>
</tbody>
</table>
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.\textsuperscript{2}

\textsuperscript{2}Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries
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
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
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
- correlate queries sent by infected host with non-infected hosts

²Kazumichi Sato, keisuke Ishibashi, Tsuyoshi Toyono, Nobuhisa Miyake, Extending Black Domain Name List by Using Co-occurrence Relation between DNS queries
Labels’ length distribution of the 12000+ DGA domains dataset

\(^2\)OpenDNS Security Lab
In paper Detecting Algorithmically Generated Malicious Domain Names, S.Yadav and A.Reddy described a system based mainly on lexical features.\(^4\)

**Features:**

- length

---

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

Features:

- length
- entropy

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

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.

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

Features:

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

\textsuperscript{4}Sandeep Yadav, Ashwath K.K. Reddy and A.L. Narasimha Reddy, Detecting Algorithmically Generated Malicious Domain Names
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.
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains."
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.\(^5\)

\(^5\)Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.\textsuperscript{5}

- count features ratios based on ground truth

\begin{table}[h]
\centering
\begin{tabular}{|c|c|l|}
\hline
\textbf{Feature Set} & \textbf{#} & \textbf{Feature Name} \\
\hline
Time-Based Features & 1 & Short life \\
& 2 & Daily similarity \\
& 3 & Repeating patterns \\
& 4 & Access ratio \\
\hline
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 \\
\hline
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 \\
\hline
Domain Name-Based Features & 14 & \% of numerical characters \\
& 15 & \% of the length of the LMS \\
\hline
\end{tabular}
\end{table}

\textsuperscript{5}\text{Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis}
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)

---

\(^5\) Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis
Combination of lexical and DNS traffic features was used in Bilge Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains.\textsuperscript{5}

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

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

\(^5\)Leyla Bilge, Engin Kirda, Christopher Kruegel and Marco Balduzzi, EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis
DGA in top level domains

DGA domains' TLD distribution

- com: 38%
- net: 37%
- org: 13%
- info: 12%

5OpenDNS Security Lab
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni

- Filtering XD domains - lexical features analysis

---

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

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

\textsuperscript{6}Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix: DGA-Based Botnet Tracking and Intelligence
PHOENIX: DGA-BASED BOTNET TRACKING AND INTELLIGENCE, SCHIAVONI

HYBRID METHOD

• 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
Hybrid Method

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni

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

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

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni

- Filtering XD domains - lexical features analysis
  1. meaningful characters ratio
  2. n-gram normality score
  3. statistical linguiistic 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, Schiavoni

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

---

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

Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni

• 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, Schiavoni

- Fingerprinting
• Fingerprinting
  1. C&C servers IP addresses
Phoenix: DGA-Based Botnet Tracking and Intelligence, Schiavoni

- 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, Schiavoni

- **Fingerprinting**
  1. C&C servers IP addresses
  2. length of the shortest and longest domain name
  3. utilized character set
• 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, Schiavoni

- **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

---

Stefano Schiavoni, Federico Maggi, Lorenzo Cavallaro, Stefano Zanero, Phoenix:
## Detection Results

<table>
<thead>
<tr>
<th>Method</th>
<th>False Positive Rate (FPR)</th>
<th>True Positive Rate (TPR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phoenix - hybrid method</strong></td>
<td>FPR - ?</td>
<td>TPR - 81.4-94.8</td>
</tr>
<tr>
<td><strong>Lexical method</strong></td>
<td>FPR - 0.3-0.8</td>
<td>TPR - 83.3-100.0</td>
</tr>
<tr>
<td><strong>Exposure - machine learning</strong></td>
<td>FPR - 0.3-1.1</td>
<td>TPR - 98.4-99.5</td>
</tr>
<tr>
<td><strong>Correlation method</strong></td>
<td>FPR - 0.5</td>
<td>TPR - 80.0</td>
</tr>
</tbody>
</table>
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
CURRENT DETECTION AND PROTECTION TECHNIQUES

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
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?