Building a Better Botnet DGA Mousetrap

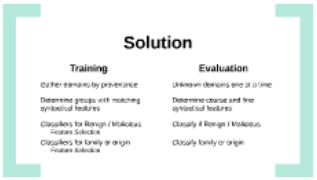
Separating Rats, Mice and Cheese in DNS Data

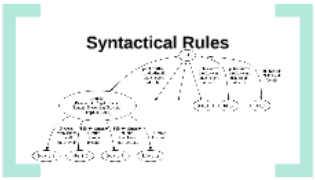
Josiah Hagen

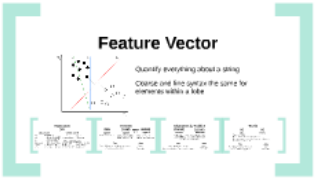
Hewlett Packard Enterprise TippingPoint

Miranda Mowbray & Prasad Rao

Hewlett Packard Labs







Building a Better Botnet DGA Mousetrap

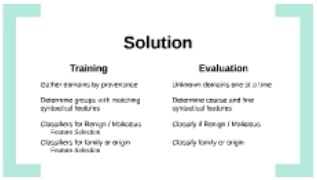
Separating Rats, Mice and Cheese in DNS Data

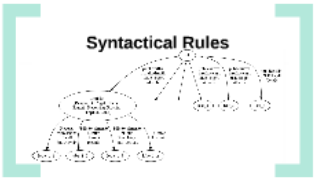
Josiah Hagen

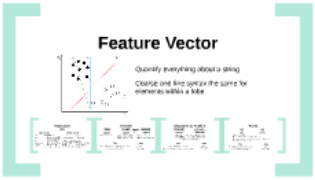
Hewlett Packard Enterprise TippingPoint

Miranda Mowbray & Prasad Rao

Hewlett Packard Labs







Goals

Classify domain names as benign and malicious

Minimize false positive classifications of malicious

Classify malicious domains according to DGA family

Not a Goal

Determine which hosts are infected with which malware

Solution

Training

Gather domains by provenance

Determine groups with matching syntactical features

Classifiers for Benign / Malicious
- Feature Selection

Classifiers for family or origin
- Feature Selection

Evaluation

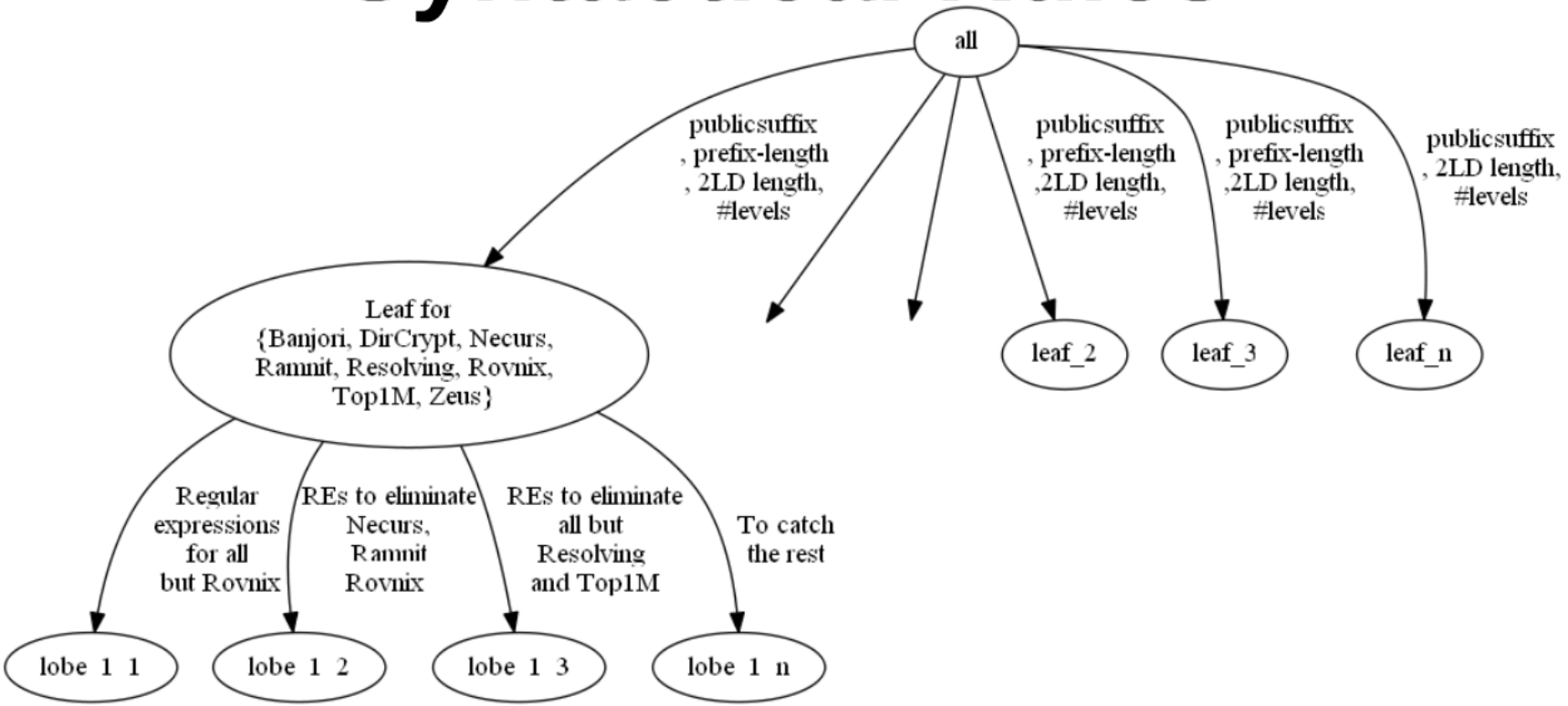
Unknown domains one at a time

Determine coarse and fine syntactical features

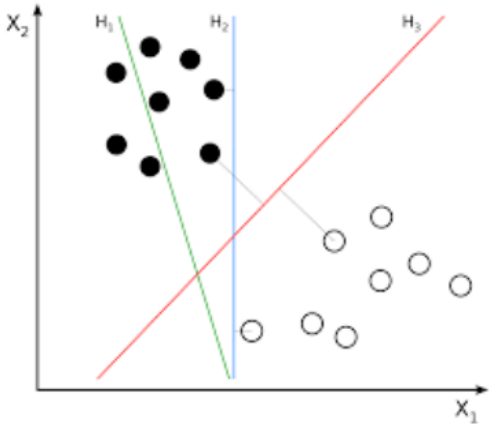
Classify if Benign / Malicious

Classify family or origin

Syntactical Rules



Feature Vector



Quantify everything about a string

Coarse and fine syntax the same for elements within a lobe

Aggregates
(22)

2LD Counts

A-F	consonants
G-Z	vowels
a-f	digits
g-z	

ISO-8859-1
- uppercase foreign
- lowercase foreign
- other printable
- other non-printable

doubles 'aa' - 'zz'
non-linguistic bigrams

Overall Counts

Dots Dashes
Dots, Dashes, & Underscores

RFC 1034 Violations

length > 254	labels not .*[a-z0-9]
labels > 63	empty labels '..'
labels not [a-z].*	invalid chars

N-Grams

(256) **1-grams** — Counts of characters

(1600) **2-grams** — Counts of character pairs: 'aa' 'ab'... 'zz'

Reduce to 40

(64000) **3-grams** — Counts of character triples: 'aaa', 'aab'...'zzz'

Pros	Cons
Separate linguistic / non-linguistic elements	Space
Catch bias in DGA PRNG	Overfitting

Characters by Position
(10240) **Forward**

(10240) **Backward**

Boolean slots for whether a given character occurs within a given position indexed from beginning and end of domain

Pros	Cons
Classifying fixed substrings	Space
Banjori, Bankpatch, Caphaw, Web Services	Overfitting

Words
(4) **2LD**

(4) **Prefix**

Counts of words
Max count of non-overlapping words
Max percentage of characters comprised of words
Length of the longest word

Pros	Cons
Classifying Benign vs Malicious	Time
Matsnu, Rovnix, Suppobox	Overfitting

Aggregates (22)

2LD Counts

A-F
G-Z
a-f
g-z
consonants
vowels
digits

ISO-8859-1

- uppercase foreign
- lowercase foreign
- other printable
- other non-printable

doubles 'aa' - 'zz'
non-linguistic bigrams

Overall Counts

Dots
Dashes
Dots, Dashes, & Underscores

RFC 1034 Violations

length > 254
labels > 63
labels not [a-z].*
labels not .*[a-z0-9]
empty labels '..'
invalid chars

x_1

N-Grams

(256)

1-grams

Counts of characters

(1600)

2-grams

Counts of character pairs: 'aa' 'ab'... 'zz'

Reduce to 40

(64000)

3-grams

Counts of character triples: 'aaa', 'aab'...'zzz'

Pros

Separate linguistic / non-linguistic elements

Catch bias in DGA PRNG

Cons

Space

Overfitting

Characters by Position

(10240) (10240)

Forward Backward

Boolean slots for whether a given character occurs within a given position indexed from beginning and end of domain

Pros	Cons
Classifying fixed substrings	Space
Banjori, Bankpatch, Caphaw, Web Services	Overfitting

Words

(4)

2LD

(4)

Prefix

Counts of words
Max count of non-overlapping words
Max percentage of characters comprised of words
Length of the longest word

Pros

Classifying Benign vs Malicious

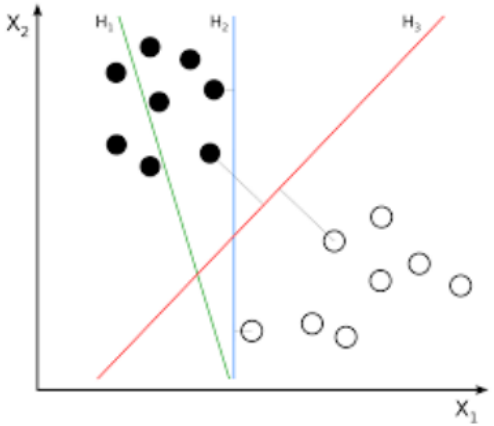
Matsnu, Rovnix, Suppobox

Cons

Time

Overfitting

Feature Vector



Quantify everything about a string

Coarse and fine syntax the same for elements within a lobe

Aggregates
(22)

2LD Counts

A-F	consonants
G-Z	vowels
a-f	digits
g-z	

ISO-8859-1
- uppercase foreign
- lowercase foreign
- other printable
- other non-printable

doubles 'aa' - 'zz'
non-linguistic bigrams

Overall Counts

Dots Dashes
Dots, Dashes, & Underscores

RFC 1034 Violations

length > 254	labels not .*[a-z0-9]
labels > 63	empty labels '..'
labels not [a-z].*	invalid chars

N-Grams

(256) **1-grams** — Counts of characters

(1600) **2-grams** — Counts of character pairs: 'aa' 'ab'... 'zz'

Reduce to 40

(64000) **3-grams** — Counts of character triples: 'aaa', 'aab'...'zzz'

Pros	Cons
Separate linguistic / non-linguistic elements	Space
Catch bias in DGA PRNG	Overfitting

Characters by Position
(10240) **Forward**

(10240) **Backward**

Boolean slots for whether a given character occurs within a given position indexed from beginning and end of domain

Pros	Cons
Classifying fixed substrings	Space
Banjori, Bankpatch, Caphaw, Web Services	Overfitting

Words
(4) **2LD**

(4) **Prefix**

Counts of words
Max count of non-overlapping words
Max percentage of characters comprised of words
Length of the longest word

Pros	Cons
Classifying Benign vs Malicious	Time
Matsnu, Rovnix, Suppobox	Overfitting

Conclusions

Syntactical rules help

Unbalanced data hurts

Results worse on real data

Especially word based DGA FPs

Some features are good for classifiers

- Aggregates
- Linguistic or not (bigrams)
- Word based

Hash words to compress dictionary to reduce FP

Not a standalone solution

Build classifers for infected hosts

C

Syntactical rules help

Unbalanced data hurts

Results worse on real da

Conclusions

Syntactical rules help

Unbalanced data hurts

Results worse on real data

Especially word based DGA FPs

Some features are good for classifiers

- Aggregates
- Linguistic or not (bigrams)
- Word based

Hash words to compress dictionary to reduce FP

Not a standalone solution

Build classifers for infected hosts

onclusions

Some features are good for classifiers

- Aggregates
- Linguistic or not (bigrams)
- Word based

A FPs

Hash words to compress dictionary to reduce FP

t a standalone solution

classifers for infected hosts

Conclusions

Syntactical rules help

Unbalanced data hurts

Results worse on real data

Especially word based DGA FPs

Some features are good for classifiers

- Aggregates
- Linguistic or not (bigrams)
- Word based

Hash words to compress dictionary to reduce FP

Not a standalone solution

Build classifers for infected hosts

Building a Better Botnet DGA Mousetrap

Separating Rats, Mice and Cheese in DNS Data

Josiah Hagen

Hewlett Packard Enterprise TippingPoint

Miranda Mowbray & Prasad Rao

Hewlett Packard Labs