Detection of malicious domains through lexical analysis

International Conference On Cyber Security And Protection Of Digital Services

June 11, 2018

Egon Kidmose, Matija Stevanovic and Jens Myrup Pedersen
egk@es.aau.dk, mst@es.aau.dk, jens@es.aau.dk

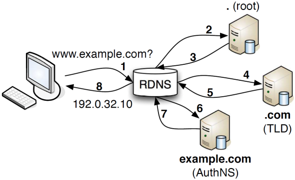
Department of Electronic Systems
Aalborg University
Denmark

AALBORG UNIVERSITY
DENMARK

Domain Names and DNS

DNS: Domain Names System

- Used everywhere, by everybody on the Internet
- ... also criminals!
- Technology/service choke-point
- Only small fraction of Internet traffic



Source: Antonakakis et al. 2011.

Example domains

- `www.example.com`
- `goggle.com`
- `yf32d9ac7f0a9f463e8da4736b12d7044a.tk`

Abuse and Malicious Domains

- Command and control
- Phishing
- Spamming
- Typo-squatting
- Blending in, tunnelling
- Fast-flux, Double-flux
- Domain-flux (DGA)

Source: Haymarket Media, Inc. [1]

Malicious domain: Any domain that is used for criminal, malicious or otherwise nefarious activities.

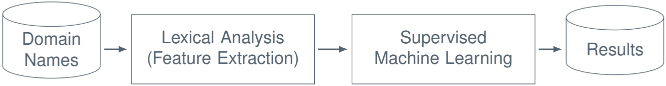
[1]: https://www.scmagazineuk.com/cyber-criminals-becoming-increasingly-professional/article/531709/

Assumption and Method

Assumption:

"Malicious domain names can be detected by lexical features."

Method:



Machine Learning: Supervised, 10-fold cross validation, 10 repeats, Random Forrest, Python, Scikit-learn.

Data

Table I: Data sets of malicious domain names.

Data sets		Number of domains	
		Non-DGA	DGA
abuse.ch	ZeuS Tracker	437	-
	Palevo Tracker	14	-
	Ransomware Tracker	1248	-
malware-\ domains.com	Just domains	13073	-
	Zeus Gameover	-	190033
	Conficker	-	106101
	Pushdo	-	10951
	GOZ	-	7348
	Microsoft Botnet	22036	-
host-file.net	Ad/tracking servers (ATS)	47960	-
	Malware distribution (EMD)	137237	-
	Exploits sites (EXP)	17282	-
	Fraud sites (FSA)	134501	-
	Spamming sites (GRM)	674	-
	Spamming sites (HFS)	573	-
	Hijack sites (HJK)	74	-
	Misleading marketing (MMT)	5533	-
	Pharmacy activities (PHA)	23143	-
	Phishing sites (PSH)	133913	-
	Warez distribution (WRZ)	3231	-
datadriven\ security.info	Cryptolocker	-	34319
	Goz	-	7347
	New Goz	-	10999
malware\ domainlist.com	Malware-related domains	1253	-
malc0de.com	Malware-related domains	208	-
malwarepatrol.net	Malicious URLs	35518	-
phistank.com	Phishing URLs	14807	-
AAU-STAR	Domains from malw.testing	27778	-
Total		620493	367098

Table II: Data sets of benign domain names.

Data sets		Number of domains	
		Non-DGA	DGA
alexa.org	Most popular domains	971424	-
datadriven\ security.info	Legitimate DGA domains	-	133927

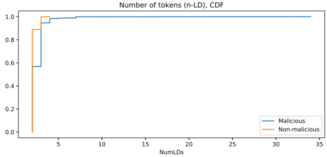
Features

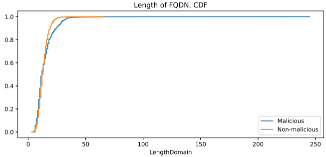
1. Basic Domain Features
    - Count, Categorical
    - TLD and FQDN
2. Simple Lexical Features
    - Length, count, ratio
    - 2LD, character classes
3. Advanced Lexical Features
    - Approximating language, recognising words
    - Entropy of letter distribution
    - N-gram analysis (`alexa.org`, English)

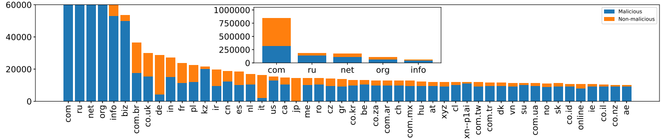
Top-Level Domain (TLD): www.example**.com**.
Second-Level Domain (2LD): www.**example**.com.
Fully Qualified Domain Name: **www.example.com**

Feature Analysis

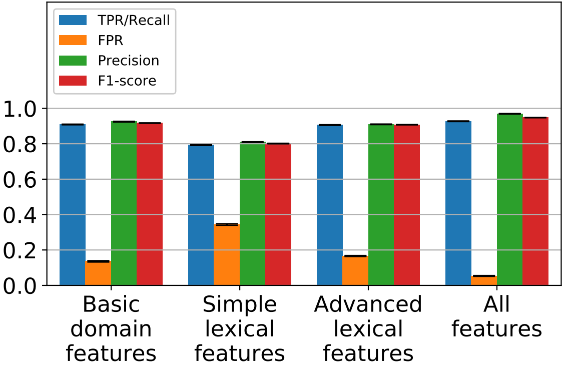
Basic Domain Features





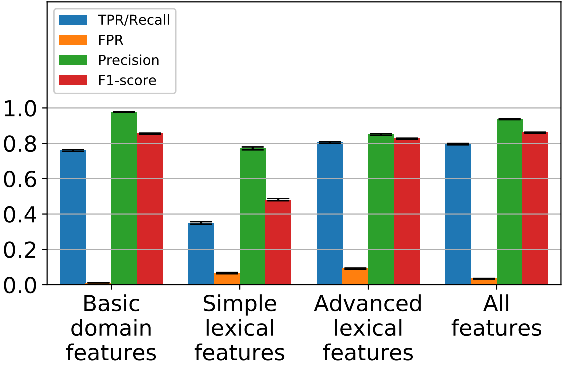


Detection: Scenario I - Full Data Set

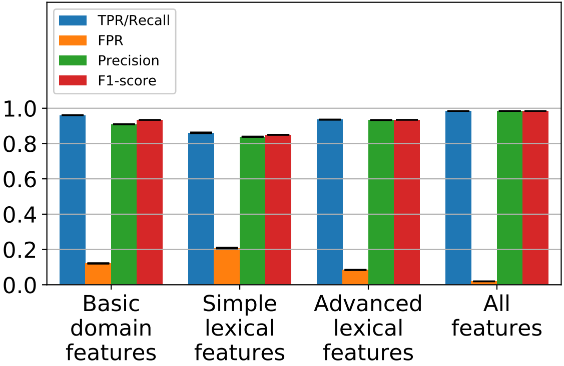


Dept. of Electronic Systems
Aalborg University
Denmark

Detection: Scenario II - Non-DGA



Detection: Scenario III - DGA



Scenario III includes the benign non-DGA domains.

Conclusion

Results

Malicious domains can be detected by using lexical features only

- Detection of DGA-domains is particular promising
  - Precision: 0.984. Recall: 0.984. F1-score: 0.948.
- Detection of Non-DGA-domains requires further work
  - Precision: 0.937. Recall: 0.795. F1-score: 0.860.
- Using all features outperforms the individual sets

General

- Vague distinction between DGA and Non-DGA
- Labelled data of high quality is hard to come by

Conclusion

Future work

- Explore other lexical features
- Explore effect of combining with non-lexical features
- Improve performance for non-DGA domains
- More substantial data sets
- Evaluate in a practical, online setting
- Apply unsupervised machine learning

Backup slides

Dept. of Electronic Systems
Aalborg University
Denmark

Existing Solutions

- Reputation
- Traffic and activity (DNS, e-mail, ...)
- Resilience/anonymity techniques

Dept. of Electronic Systems
Aalborg University
Denmark

Basic Domain Features

1. Number of domain levels (n-LD).
2. Top level domain (TLD).
3. Length of Fully Qualified Domain Name (FQDN).

Simple Lexical Features

1. Length of 2nd Level Domain (2-LD).
2. Ratio of consonants in the 2-LD.
3. Number of vowels in 2-LD.
4. Number of numeric characters in 2-LD.
5. Number of special characters in 2-LD.
6. Ratio of special characters in 2-LD.

Features

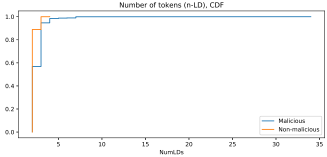
Advanced Lexical Features

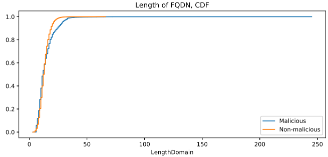
1. Language indicator (`langid.py`)
2. Number of English words in 2-LD.
3. Entropy of 2-LD.
4. N-gram analysis of 2-LD (www.alexa.org) [1] .
5. N-gram analysis of 2-LD (English dictionary).

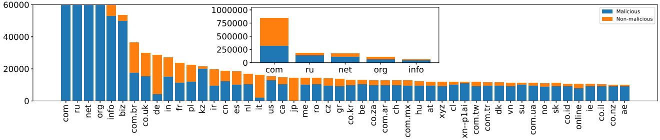
[1] {3,4,5}-grams,

Feature Analysis

Basic Domain Features

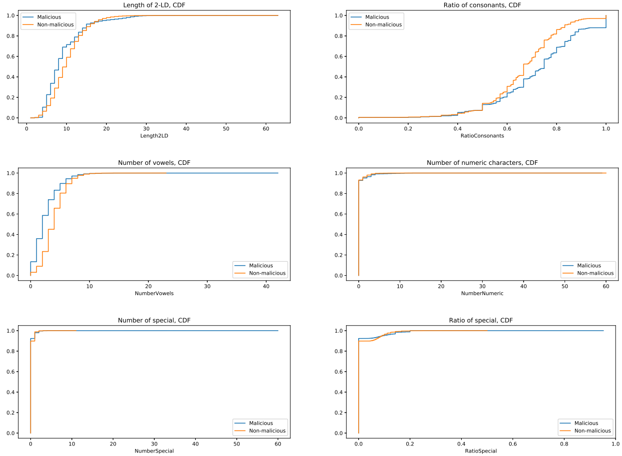






Feature Analysis

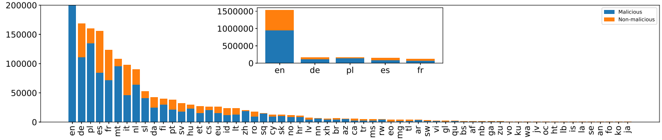
Simple Lexical Features

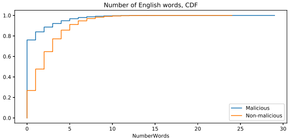


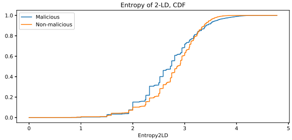
Dept. of Electronic Systems
Aalborg University
Denmark

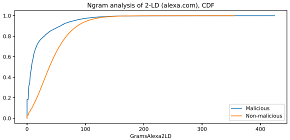
Feature Analysis

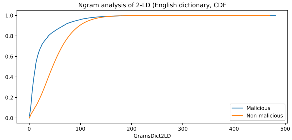
Advanced Lexical Features











Detection: Mean performance

Scenario	Features	TPR/Recall	FPR	Precision	F1-score
I (Full)	1	9.09e-01	1.36e-01	9.25e-01	9.17e-01
	2	7.93e-01	3.44e-01	8.10e-01	8.01e-01
	3	9.06e-01	1.66e-01	9.10e-01	9.08e-01
	All	**9.27e-01**	**5.35e-02**	**9.70e-01**	**9.48e-01**
II (Non-DGA)	1	7.59e-01	**1.10e-02**	**9.78e-01**	8.55e-01
	2	3.50e-01	6.69e-02	7.69e-01	4.81e-01
	3	**8.05e-01**	9.16e-02	8.49e-01	8.26e-01
	All	7.95e-01	3.41e-02	9.37e-01	**8.60e-01**
III (DGA)	1	9.60e-01	1.21e-01	9.09e-01	9.34e-01
	2	8.61e-01	2.08e-01	8.38e-01	8.50e-01
	3	9.35e-01	8.40e-02	9.33e-01	9.34e-01
	All	**9.84e-01**	**1.98e-02**	**9.84e-01**	**9.84e-01**