Comparison of Machine Learning Algorithms for Classification of Algorithmically Generated Domains

Bc. Frederik Koľbík
Supervisor: Mgr. Jakub Daubner, PhD.

June 16, 2020

Outline

- Malware and domain generation algorithms
- Methodology and data
- Results

Malware

- need to communicate with command-and-control (C&C) servers, botnets especially
- first botnets - hard-coded IP address or domain name of the C&C server - reverse engineering - block communication
- solution - domain generation algorithms (DGAs)

C&C communication

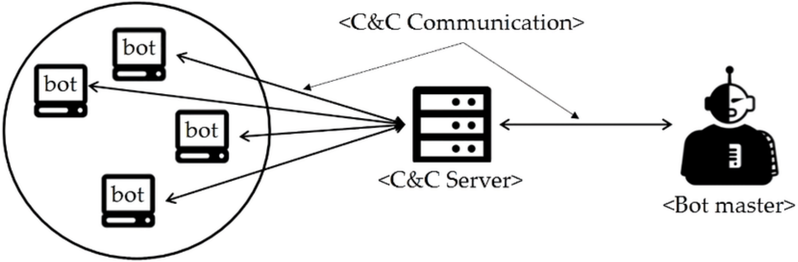

Figure: C&C communication. Jeon, Jaewoo & Cho, Youngho. (2019). Construction and Performance Analysis of Image Steganography-based Botnet in KakaoTalk Openchat. Computers. 8. 61. 10.3390/computers8030061.

DGAs

- dynamically generate a large number of domains
- only a small portion is used in C&C communication
- domain generation based on shared secret (seed)
  - constant
  - current time
  - trending Twitter topics
  - …

DGA types

- arithmetic-based
  - ASCII values (hcfoopojnuqxho.su)
  - offset in arrays of characters (gatyfus.com)
- hash-based (bd9b9c8ca02a67700b45839adb1f37e736.ws)
- wordlist-based (increaseinside.net)
- permutation-based (loreredotntexp.info)

DGA example

```
for i = 0 to 13:
    day = (day >> 15) ^ 16 * (day & 0x1FFF ^ 4 * (seed ^ day))
    year = ((year & 0xFFFFFFF0) << 17) ^ ((year ^ (7 * year)) >> 11)
    month = 14 * (month & 0xFFFFFFFE) ^ ((month ^ (4 * month)) >> 8)
    seed = (seed >> 6) ^ ((day + 8 * seed) << 8) & 0x3FFFF00
    int x = ((day ^ month ^ year) % 25) + 'a'
    domain[i] = x
```

Example 1: Pseudo code of DGA of Ranbyus. Reversed and reimplemented by Johannes Bader [1].

Detecting DGA domains

- machine learning - popular and good results
- various approaches tested - clustering, classification, deep learning...
- side information - none, DNS traffic data, WHOIS

Goal

- which classifiers are the best?
- what features to use?
- comparison of five classifiers:
  - Gaussian Naive Bayes
  - Random Forest
  - Gradient Boosting Classifier
  - Logistic Regression
  - Support Vector Machine
- our focus on supervised classifiers and arithmetic-based and hash-based DGAs

Data

- DGA domains
    - DGArchive [3]
    - almost 50 million domains from previous 3 years
- clean domains
    - TRANCO list [2] - aggregated from Alexa, Cisco Umbrella, Majestic and Quantcast lists
    - one million domains from February 2020

Dataset building

- only malware families with two levels of domains
- domains of 73 malware families used
- from each family - 30,000 domains or all
- all clean domains from TRANCO list
- final dataset - 2,008,828 domains

Experiments

- K-Fold
  - data split into k subsets (folds)
  - k iterations of training and testing
- Leave One Group Out (LOGO)
  - one group of data is left out and used as a testing set
  - in our case - all domains of left out family used as a testing set

Processing results

- Accuracy - ACC $= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$
- True Positive Rate - TPR $= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$
- False Positive Rate - FPR $= \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$

Features

- domain name length
- TLD features
- digits features
- character ratios
- longest character sequences
- n-grams
- other

Features subsets

- all features
- best features from statistical tests (chi-squared test, ANOVA F-test, mutual information test)
- all features except digits features
- all features except n-grams features
- only n-grams features

Results for K-Fold

- best features subsets overall - all features except digits features and all features
- best classifiers overall - Random Forest and Gradient Boosting Classifier
- best result - Random Forest, all features except digits features - 99.2% accuracy, 98.5% TPR and 0.15% FPR
- very low standard deviation in all experiments

Results for LOGO

- best features subsets and classifiers overall - same as before
- best result - Random Forest, all features except digits features
  - mean - 98.9% accuracy, 97.4% TPR, 0.14% FPR
  - median - 99.8% accuracy, 99.6% TPR, 0.14% FPR
- very high standard deviation across all experiments - domains of some malware families are very hard to detect

Hard-to-detect malware families

- 21 hard-to-detect families
- analysis of features of hard-to-detect, easy-to-detect and clean domains
- hard-to-detect domains - short, no digits, small number of unique characters - many features affected
- sometimes DGA design - less random looking domains

Real-world data tests

- real-world data - ESET
  - 1 million random domains
  - 3.2 million NXDomains
  - Authlist - 75,000 clean domains
- results mirror previous tests
- NXDomains - most DGA domains predicted

Implementation and speed

- desktop PC: Intel Core i7-7700 @ 3.6 GHz, 16 GB RAM, Windows 10
- Python: scikit-learn and pandas libraries
- extraction of all features - 6.5 minutes for 1 million domains

Model	**Training**	**Testing**
Gaussian Naive Bayes	0.25 min.	20 s
Gradient Boosting Classifier	64 min.	16 s
Logistic Regression	24 min.	16 s
Random Forest	33 min.	169 s
Support Vector Machine	3 min.	10 s

Table: Training and testing times.

Future work

- better features for hard-to-detect families
- comparison of deep learning methods
- combination of methods for different DGA types

Thank you for your attention

Bibliography

- Johannes Bader.
  The DGA of Ranbyus.
  `https://johannesbader.ch/blog/the-dga-of-ranbyus/`.

- Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen.
  Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation.
  In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, NDSS 2019, February 2019.

- Daniel Plohmann.
  DGArchive.
  URL https://dgarchive. caad. fkie. fraunhofer. de, 2015.