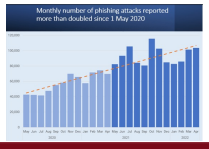
PHISHWEB – a Progressive, Multi-Layered System for Phishing Websites Detection

Lucas Torrealba A. (1 & 2), P. Casas (1), Javier Bustos-Jiménez(2)
(1) AIT Austrian Institute of Technology, Vienna, Austria
(2) NIC Chile Research Labs, University of Chile, Chile

MOTIVATION & CHALLENGE

¿Why phishing – malicious domain detection is important?

- **Compromises the security** of users.
- By the Internet Crime Complaint Center (IC3) Report *"Almost **2 billion lost** to BEC scams in 2020"*
- By the phishing landscape the number of phishing attacks and domain names reported for **phishing trended up** over the yearly period.



TYPES OF MALICIOUS DOMAINS

The domain names used for phishing could be (1) Hacked Domain Names or (2) Malicious Domain Names. The second can be classified into:

- **Typo:** Domains that are registered with a name that is an intentionally misspelled version of other sites. *e.g. facebook-policyform442.tk*
- **Homoglyph**: Domains in which international characters or symbols are included that look like a character in the original domain. *e.g. amazon.com*
- **AGD**: Domains that are generated by DGA, which can dynamically generate several random domain names with a random seed. These domains are mainly used by botnets and malware to avoid detection and blocking. *e.g. wmncfyxcwc.com*

DATASET

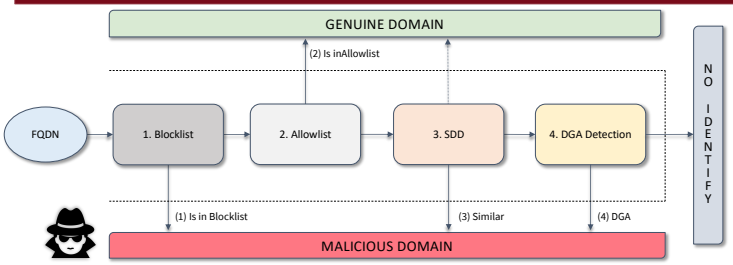
- **PhishStorm:** Dataset used in [1] for phishing detection. 96.018 URLs. 50% benign URLs and 50% phishing URLs.
- **Phishing-Benign:** Dataset used in [2] for phishing detection. 38.800 URLs. 50% benign - 50% phishing.
- **DGA:** Dataset used in [3] for algorithmically generated malicious domain detection. 675.000 domain names. 50% non-AGD and 50% AGD.

The data in 2 first datasets correspond to URLs. This is why a data processing is performed where only both the protocol and the path are removed, leaving only the domain name. This process causes loss of information as we can see:

xosothudo.com.vn/**paypal.co.il**/**paypal.co.il**cgi-bin.websc… → PROCESSING → xosothudo.com.vn

[1] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel. PHishStorm: Detecting phishing with streaming analytics. IEEE Transactions on Network and Service Management, 2014.
[2] Peya Mowar and Mini Jain. Fishing out the phishing websites. In 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2021.
[3] Alessandro Cucchiarelli, Christian Morbidoni, Luca Spalazzi, and Marco Baldi. Algorithmically generated malicious domain names detection based on n-grams features. Expert Systems with Applications, 2021.

DESIGN SYSTEM



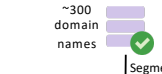
1. BLOCKLIST & 2.WHITELIST

The first filter corresponds to the usage of the popular blacklist PhishTank that on 18th August contained around 18000 different domain names actives. The domain names are ordered lexicographically and binary search is used to check if the FQDN is in the blacklist.

The second filter corresponds to the usage of a whitelist with well known domains. For the construction of the whitelist, we collected the data of one year from Cisco Umbrella 1 Million Top Domains and for each domain name we calculate the median ranking over the year. After we order the domain names by ranking and select the first 100.000 domain names as a whitelist.

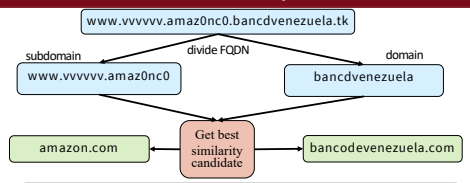
3. SDD (SIMILAR DOMAIN DETECTION)

Hash Table Creation



Segment each domain into 3-gram and adding into the list in the hash table

Key	Value
ban	[bancochile.cl, bankofamerica.com,…]
pay	[paypal.com, amazonpay.in, …]
boo	[facebook.com, outlook.com, …]
ter	[twitter.com, interbank.pe, …]



To get the best candidate we segment the (sub-)domain into 3-gram and for each 3-gram we collected the domain in the list associated in the hash table as possibility. For each possibility we calculate:

1. Edit Distance: We use the Levenshtein Distance.
2. Percent contained: $\frac{\#3\text{-grams in both domains}}{\#3\text{-grams in the possibility domain}}$

If ED==1: This is the candidate !
Else: We select as candidate the domain with highest PC

Domain

Once obtained the best candidate for the domain, we calculate the Cosine similarity and the Jaccard index between the 3-grams set:

A: bancdvenezuela
B: bancodevenezuela

1. Cosine Similarity $\frac{\sum_{i=1}^{m} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{m} A_i^2} \cdot \sqrt{\sum_{i=1}^{m} B_i^2}}$ → 0.69
2. Jaccard Index $\frac{|A \cap B|}{|A \cup B|}$ → 0.53

- If ED == 1 || PC > λ_0
- If PC > λ_1 && CS > λ_2
- If PC > λ_3 && JI > λ_4
- If new ED == 1

Subdomain

Once obtained the best candidate for the subdomain, we calculate the same metrics as in the domain and a new edit distance for the subdomain is calculated.

4. DGA DETECTION

Weighted Dictionary Creation

e.g. facebook.com

Top 100k domains Alexa

For each domain name
Segment into n-grams and adding into a list

3-gram: [fac,ace,ceb,ebo,boo,ook,…]
4-gram: [face,aceb,cebo,eboo,book,…]
5-gram: [faceb,acebo,ceboo,ebook,…]
6-gram: [facebo,aceboo,cebook,…]
7-gram: [faceboo,acebook,…]

Counting the number of occurrence of each n-gram

3-gram Weighted Dictionary

ban: 7.42
fac: 6.31
…
ci5: 0

Applying $log_2(^{c_i}/_n)$

ban: 551
fac: 238
…
ci5: 1

c_i : number of occurrences of the i-th substring
n : n in the n-gram

Repeat for n in {3, 4, 5, 6, 7}

Calculation Reputation Value PHISHWEB

aumqdksbdntltjuj.xyz
domain length: L = 28

Segment into **n**-grams and adding the associated weighted values

aum: 3.54
umq: 0.00
…
tju: 0.00
juj: 1.22

Sum values → 17.33 → Applying reductor $^{n}/_{L}$ → 3.06

Repeat for **n** in {3, 4, 5, 6, 7} and sum values to obtain the Reputation Value PHISHWEB (RVP)

RVP=3.06 + 0.12 + 0 + 0 + 0 = 3.18

Modification of the work by Zhao et. al. [4]

If $RVP(d) < \lambda_{dga}$

$$RVP(d) = \sum_{n=3}^{7} RVP_n(d) = \sum_{n=3}^{7} \frac{n}{L} \times \sum_{i=1}^{L-n+1} W_i = \sum_{n=3}^{7} \frac{n}{L} \times \sum_{i=1}^{L-n+1} log_2(^{c_i}/_n)$$

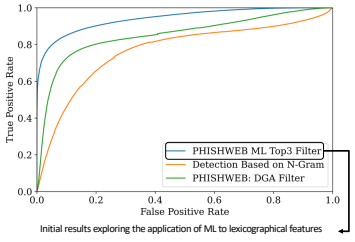
[4]Hong Zhao, Zhaobin Chang, Guangbin Bao, Xiangyun Zeng, and Saman S. Chaeikar. Malicious domain names detection algorithm based on n-gram. J. Comput. Netw. Commun., 2019.

RESULTS & FUTURE WORK

Example of Classification

DOMAIN	FILTER DETECTOR	MATCH	CLASSIFICATION
habbocretidosparati.blogspot.com	Blocklist	-	Malicious
www.citibank.co.jp	Whitelist	-	Genuine
www.bancoestadochile.net	SDD	bancoestado.cl	Malicious
bancodevenezla.co	SDD	bancodevenezuela.com	Malicious
facebeok.net.ms	SDD	facebook.com	Malicious
paypail.cgi-bin.webscr.cmd…	SDD	paypal.com	Malicious
yyhxkjpx.com	DGA	-	Malicious
www.biohidrica.cl	-	-	Unidenfity

Performance DGA Filter



Initial results exploring the application of ML to lexicographical features

Performance PHISHWEB System

DATASET	# DOMAINS	% CLASSIFIED	% ACC	% PREC
PhishStorm	46.000	21.2%	95.0%	95.3%
Phishing-Benign	19.500	21.6%	89.4%	89.6%
DGA	675.000	20.0%	93.5%	92.5%

Future Work

- Remove diacritics for SDD analysis
- Replace by similar character (0 → o)
- Add redirection checker in SDD
- Add new features (ASN, DNS, content-html, etc.)

¡¡ questions or comments → lucas@niclabs.cl !!