Data-Driven Characterization and Detection of COVID-19 Themed Malicious Websites

Mir Mehedi Ahsan Pritom*, Kristin M. Schweitzer†, Raymond M. Bateman†, Min Xu‡ and Shouhuai Xu*
*Department of Computer Science, University of Texas at San Antonio
†U.S. Army Research Laboratory South - Cyber
‡Mastercard

***Abstract*—COVID-19 has hit hard on the global community, and organizations are working diligently to cope with the new norm of "work from home". However, the volume of remote work is unprecedented and creates opportunities for cyber attackers to penetrate home computers. Attackers have been leveraging websites with COVID-19 related names, dubbed *COVID-19 themed malicious websites*. These websites mostly contain false information, fake forms, fraudulent payments, scams, or malicious payloads to steal sensitive information or infect victims' computers. In this paper, we present a data-driven study on characterizing and detecting COVID-19 themed malicious websites. Our characterization study shows that attackers are agile and are deceptively crafty in designing geolocation targeted websites, often leveraging popular domain registrars and top-level domains. Our detection study shows that the Random Forest classifier can detect COVID-19 themed malicious websites based on the lexical and WHOIS features defined in this paper, achieving a 98% accuracy and 2.7% false-positive rate.**

***Index Terms*—COVID-19 Cyberattacks, Malicious Websites, Detection, Defense**

I. Introduction

The COVID-19 pandemic has incurred many new cyber attack vectors. Many of these cyber attacks incorporate COVID-19 themed factors into phishing, malware, and scamming schemes for various malicious goals (e.g., monetary benefits, stealing credentials, stealing credit card numbers, or identity theft). For example, there is reportedly a 148% increase in ransomware attacks in March 2020 compared with February 2020 [1], where many attacks are initiated by malicious websites abusing victims' trust.

This paper focuses on one emerging attack vector, namely malicious websites leveraging COVID-19 as a theme or *COVID-19 themed malicious websites* [2]. As organizations incorporate the "work from home" policy, the consequences of COVID-19 themed malicious websites can be significantly amplified because home computers are often more vulnerable to attack than work computers. During the COVID-19 pandemic, many people lost their jobs and are affected by mental health issues, which causes excessive pressures. These pressures may make average users even more vulnerable to social engineering attacks waged via COVID-19 themed malicious websites. This increases the motivation of the importance of understanding and defending against COVID-19 themed malicious websites, which is a new problem that has not been studied before in a systematic way.

Our contributions. In this paper, make the following contributions. First, we propose a methodology for characterizing and detecting COVID-19 themed malicious websites through a data-driven approach. To the best of our knowledge, this is the first study on *data-driven* characterization and detection of COVID-19 themed malicious websites. Second, we apply the methodology to specific datasets to draw the following insights: (i) some attackers may be incentivized to use cheaper registrars for registering COVID-19 themed malicious websites; (ii) attackers often abuse popular top-level domains for their COVID-19 themed malicious websites; (iii) attackers are agile in waging the COVID-19 themed malicious website attack; (iv) attackers are crafty in using COVID-19 themed keywords, and geographical information in creating COVID-19 themed malicious website domain names; (v) the small degree of data imbalance does not have any significant impact in the effectiveness of detecting COVID-19 themed malicious websites; and (vi) COVID-19 themed malicious website detectors must consider WHOIS features and Random Forest performs better than K-nearest neighbor, decision tree, logistic regression, and support vector machine.

Paper outline. The rest of the paper is organized as follows. Section II explores the related work. Section III explores the research questions which guide us to characterize and detect COVID-19 themed malicious websites. Section IV reports the experiments and results. Section V discuss our weakness and future research opportunities. Section VI concludes the paper.

II. Related Works

Although the problem of COVID-19 themed malicious websites has not been investigated until now, the problem of malicious websites has been studied in the literature prior to the COVID-19 pandemic. The problem of detecting malicious URLs generated by domain generating algorithms has been investigated in [3]. The problem of detecting phishing websites has been addressed via various approaches, including: the descriptive features-based model [4], the lexical and HTML features-based model [5], the HTML and URL features-based model [6], and the natural language processing and word vector features-based model [7]. The problem of detecting malicious websites has been addressed via the following approaches: leveraging application and network layers information [8], leveraging image recognition [9], leveraging generic URL features [10], [11], leveraging character-level embedding

or keyword-based recurrent neural networks [12]–[14], the notion of adversarial malicious website detection [15]. However, these studies do not consider features pertinent to the COVID-19 pandemic, which are we leverage. Nevertheless, the present study fall under the umbrella of cybersecurity data analytics [16]–[20], which in turn belong to the Cybersecurity Dynamics framework [21]–[25].

III. Methodology

Our methodology for *data-driven* characterization and detection of COVID-19 themed malicious websites is centered at answering a range of research questions.

A. Characterization Methodology

In order to characterize COVID-19 themed malicious websites, we address 4 Research Questions (RQs):

- RQ1: Which WHOIS registrars are most abused to launch COVID-19 themed malicious websites?
- RQ2: Which Top Level Domains (TLDs) are most abused by COVID-19 themed malicious websites?
- RQ3: What trends are exhibited by COVID-19 themed malicious websites?
- RQ4: Which theme keywords are mostly abused by attackers, and how?

We consider WHOIS information because it has shown to be useful in the era prior to the COVID-19 pandemic [8], [15]. Answering the preceding questions will deepen our understanding of COVID-19 themed malicious website attacks.

B. Detection Methodology

We propose leveraging machine learning to detect COVID-19 themed malicious websites and answer:

- RQ5: Which classifier is competent in detecting COVID-19 themed malicious websites?
- RQ6: What is the impact of WHOIS features on the classifier's effectiveness?

In order to answer these questions, we need to train detectors. Figure 1 highlights the methodology for detecting COVID-19 themed malicious websites. The methodology can be decomposed into the following modules: data collection, feature definition and extraction, data pre-processing, classifier training, and classifier test.

Data about websites need to be collected from reliable sources. The collected data may need enrichment to provide more information, as what will be illustrated in our case study. Then, features may be defined to describe these websites. In the case of using deep learning (which requires much larger datasets), features may be automatically learned. One may consider a range of classifiers, which are generically called C_i's in Figure 1. As shown in Figure 1, one can use classifiers individually or an ensemble of them (e.g., via a desired voting scheme, such as weighted vs. unweighted majority voting). In the simple form of unweighted majority voting, a website is classified as malicious if majority of the classifiers predict it as malicious; otherwise, it is classified as benign.

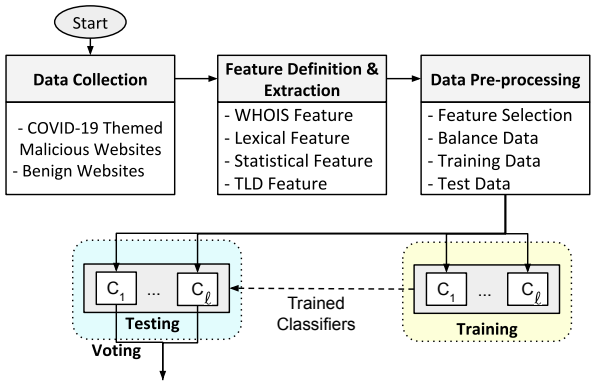

Fig. 1. Methodology for detecting COVID-19 themed malicious websites

In order to evaluate the effectiveness of the trained classifiers, we propose adopting the standard metrics, including: accuracy (ACC), false-positive rates (FPR), false-negative rates (FNR), and $F1$-score. Specifically, let TP be the number of true positives, TN be the number of true negatives, FP be the number of false positives, and FN be the number of false negatives. Then, we have ACC $= \frac{TP+TN}{TP+TN+FP+FN}$, FPR $= \frac{FP}{FP+TN}$, FNR $= \frac{FN}{FN+TP}$, and $F1$-score $= \frac{2TP}{2TP+FP+FN}$.

IV. Case Study

Our case study applies the methodology to specific datasets.

A. Data Collection

Our dataset of COVID-19 malicious website examples are obtained from what was published between 2/1/2020 and 5/15/2020 by two sources: (i) CheckPhish [26], which contains 131,761 malicious websites waging scamming attacks related to COVID-19; and (ii) DomainTools [27], which contains 157,579 malicious websites waging malware, phishing, and spamming attacks related to COVID-19. The union of these two sets leads to a total of 221,921 malicious websites, denoted by $D_{malicious}$, owing to the fact that 67,419 websites belong to both sets. For obtaining benign websites, we use the top 250,000 websites from Cisco's Umbrella 1 million websites dataset [28] on 05/16/2020, denoted by D_{benign}, which is a source of reputable websites. We compile a merged dataset denoted by $D_{initial} = D_{malicious} \cup D_{benign}$.

In order to collect WHOIS information of a website, we use the python library `whois 0.9.7` to query the WHOIS database on 8/7/2020. We observe that 42,540 (or 19.17%) out of the 221,921 malicious websites have no WHOIS information available, and 93,082 (or 37.2%) out of the 250,000 benign websites have no WHOIS information available. This means that the presence/absence of WHOIS information does not indicate that a website is malicious or not.

B. Characterization Case Study

1) Answering RQ1: Identifying the WHOIS registrars that are most abused to launch COVID-19 themed malicious websites: For this purpose, we use a subset of $D_{malicious}$ set,

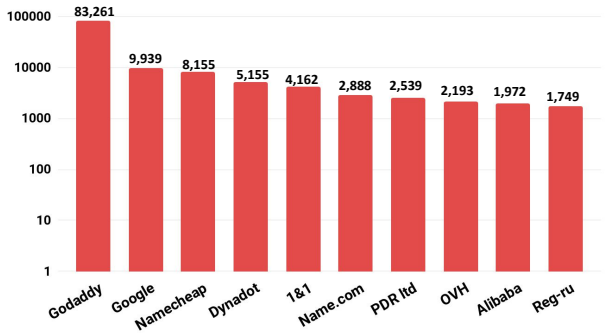

Fig. 2. Top 10 abused WHOIS registrars of COVID-19 themed malicious websites (the y-axis is in the log-scale).

denoted by $D'_{malicious}$, which contains 171,901 malicious websites with WHOIS *registrar_name* information available.

Figure 2 depicts the top 10 abused registrars, which are ranked according to the absolute number of COVID-19 themed websites in $D'_{malicious}$ that are respectively registered by them. We observe that `Godaddy` is the most frequently abused registrar, followed by `Google` and `Namecheap`. This finding inspires us to analyze if there is any financial incentive behind the use of a specific registrar. The cost registering a `.com` domain in the first year, is: `Godaddy` for $11.99, `Google` for $9, `Namecheap` for $8.88, `Dynadot` for $8.99, `1&1` for $1, `name.com` for $8.99, `PDR Ltd` for $35, `OVH` for $8.28, `Alibaba` for $7.99, `Reg-ru` for $28. This suggests that some attackers might have considered registrar `1&1` because it is the cheapest, while some attackers use reputed registrars.

***Insight** 1:* Some attackers may be incentivized to use cheaper registrars but some of the other don't.

2) Answering RQ2: Which Top Level Domains (TLDs) are most abused by COVID-19 themed malicious websites?: In order to answer this question, we use the original dataset $D_{malicious}$, which contains 221,921 COVID-19 themed malicious websites with corresponding TLD information.

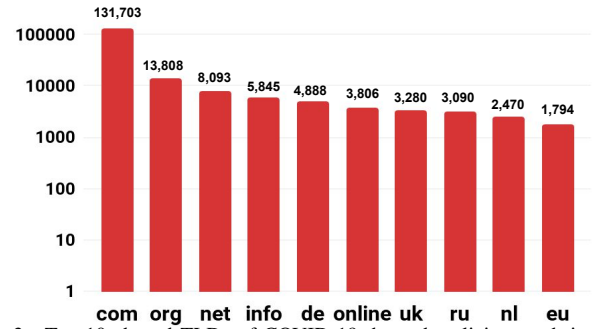

Fig. 3. Top 10 abused TLDs of COVID-19 themed malicious websites (the y-axis is in the log-scale).

Figure 3 depicts the top 10 abused TLDs, which are ranked according to the absolute number TLDs for COVID-19 themed malicious websites. We make the following observations. First, `.com` hosts the highest number of malicious websites, followed by `.org` and `.net`. Second, 5 of the top 10 abused TLDs correspond to country-level `ccTLDs`, including `.de`, `.uk`, `.ru`, `.nl` and `.eu`.

***Insight** 2:* Attackers often abuse popular TLDs.

3) Answering RQ3: What trends are exhibited by COVID-19 themed malicious websites?: In order to answer this question, we use the dataset $D_{malicious}$ mentioned above. Figure 4 depicts the trend of malicious websites, leading to two observations. First, there is a discrepancy between the daily numbers of websites that are reported by the two sources. According to CheckPhish, the number of COVID-19 themed malicious websites reaches the peak on 03/25/2020, with 18,495 malicious websites; according to DomainTools, the number of COVID-19 themed malicious websites reaches a peak on 03/20/2020, with 3,981 malicious websites. This data indicates that there are reporting inconsistencies among sources and many COVID-19 themed malicious websites are created at the early stage of the pandemic when *uncertainties* are maximum. Second, the number of COVID-19 themed malicious websites, by and large, has been decreasing since the last week of March 2020 (i.e., two weeks after the pandemic declaration), leading to about 1,000 websites per day during the first week of May 2020 (i.e., about two months after pandemic declaration). However, there is still oscillation. One possible cause is that the attackers have been waiting to create new COVID-19 themed malicious websites based on the pandemic's new developments (e.g., vaccine).

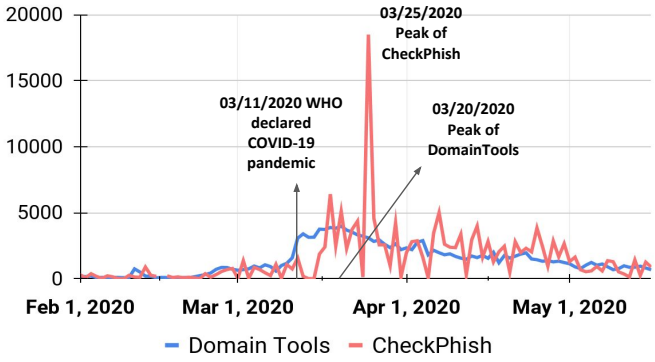

Fig. 4. Trends of COVID-19 themed malicious website.

***Insight** 3:* Inconsistencies in reporting mechanisms, attackers are agile in creating COVID-19 themed malicious websites.

4) Answering RQ4: Which theme keywords are mostly abused by attackers, and how?: In order to answer this question, we analyze the dataset $D_{malicious}$ mentioned above. We use the python library `wordninja` with English Wikipedia language model [29] to split domain name strings and extract COVID-19 themed keywords. We observe that 4 keywords (i.e., *covid*, *corona*, *covid19*, and *coronavirus*) are most widely used as expected; they are followed by *mask*, *quarantine*, *virus*, *test*, *facemask*, *pandemic*, and *vaccine*. We extract more than 19,000 keywords. A further analysis of the domain names reveals that attackers create COVID-19 themed malicious websites with names containing geographical attributes. For example, `coronavirusprevention sanantonio.com`, `coronavirusprecentionhouston.com`, and `coronavirusprevention dallas.com` use a

combination of city name and a COVID-19 themed keyword. Moreover, we observe the existence of COVID-19 themed "parking" websites, which have no content at the present time but might be used for upcoming COVID-19 themes.

***Insight** 4:* Attackers are crafty in using COVID-19 themed keywords and geographical information in creating COVID-19 themed malicious website domain names.

C. Detection Case Study

Given $D_{initial}$, the detection case study proceeds as follows.

1) Feature Definition and Extraction: We define features according to the following aspects of websites: WHOIS (F1-F4), domain name lexical information (F5-F9), statistical information (F10), and Top-Level Domain or TLD (F11).

- Current WHOIS registration lifetime (F1): This is the number of days that has passed since a website's registration, with respect to the date when this feature's value is extracted (e.g., 08/07/2020 in our case).
- Remaining WHOIS expiration lifetime (F2): This is the number of remaining days before a website's WHOIS registration expires, with respect to the date when this feature's value is extracted (e.g., 08/07/2020 in our case).
- Number of days since last WHOIS update (F3): This is the number of days elapsed since a website's last update with respect to the date when this feature's value is extracted (e.g., 08/07/2020 in our case).
- WHOIS registrar reputation (F4): We propose measuring a WHOIS registrar's reputation as $\frac{n}{|D_{benign}|}$, where n is the number of benign websites in D_{benign} that are registered by this particular registrar and $|D_{benign}|$ is the size of set D_{benign}.
- Number of dots in domain name (F5): This is the number of dots (character '.') in the domain name. For example, domain `any.com` has 1 dot.
- Domain hyphen count (F6): This is the number of hyphens ('-') in a domain name.
- Domain vowel count (F7): This is the number of vowels (i.e., *a*, *e*, *i*, *o*, *u*) in a domain name.
- Domain digits percentage (F8): This is the ratio of the number of digits (0-9) in a domain name to the number of characters including digits.
- Domain unique alphabetic-numeric characters count (F9): This is the total number of unique alphabetic and numeric characters (i.e., a-z, A-Z, 0-9) in a domain name.
- Domain entropy (F10): This is the Shannon entropy [30] of the domain name (i.e., a kind of statistical information), which is computed based on the frequency of characters in the domain name.
- TLD Reputation (F11): We propose measuring a TLD's reputation as $\frac{m}{|D_{benign}|}$, where m is the number of websites in D_{benign} that contain this particular TLD.

2) Data Pre-Processing: Given that some websites may not have information for the features, it is important to consider different scenarios. In our example, we propose considering two datasets that can be derived from $D_{initial}$ because some websites do not have information for the WHOIS features.

- Dataset $D_1 \subset D_{initial}$ consists of websites for which WHOIS information is available (i.e., features F1-F4 are available). D_1 contains 21,749 websites in total, including 16,411 COVID-19 themed malicious websites and 5,338 benign websites.
- Dataset $D_2 \subset D_{initial}$, where $D_1 \cap D_2 = \emptyset$, consists of websites for which WHOIS information is absent (i.e., features F1-F4 are entirely missing). D_2 contains 135,621 websites, including 42,540 malicious websites and 93,081 benign websites. For each website belonging to D_2, only values of the 7 features (i.e., F5-F11) are available.

TABLE I
RELATIVE IMPORTANCE OF FEATURES IN D_1 WITH RESPECT TO THE RANDOM FOREST METHOD.

Feature	Importance	Feature	Importance
F1	0.429	F7	0.080
F2	0.094	F8	0.009
F3	0.131	F9	0.028
F4	0.065	F10	0.029
F5	0.065	F11	0.068
F6	0.003		

Since only D_1 contains all WHOIS information, We use it for feature selection study. For this purpose, we use the *random forest classification feature importance* method [31] (with the 80-20 splitting of training-test data) to find the important features. Table I depicts the relative importance of the features in D_1. We observe that F6 and F8 have a very small relative importance (i.e., < 0.01) when compared to the others, suggesting that hyphens and digits are equally used in malicious or benign domain names. Hence, we will eliminate F6 and F8 in the rest study of D_1.

In order to see whether or not the feature selection result is impacted by the data imbalance of D_1 (with the malicious:benign ratio being 3.1:1), we explore two widely-used methods: (i) *oversampling* the minority class to replicate some random examples; and (ii) *undersampling* the majority class to remove some random examples. At first, we do the 80-20 splitting of training-test data, and then change the malicious:benign ratio in the training set, while keeping the test set intact. We wish to identify the ratio that achieves the highest $F1$-score. In what follows we only report the results of Random Forest because it outperforms the other classifiers for the original dataset D_1.

Table II shows the impacts of the malicious:benign ratio in the training set. We observe that the oversampling-incurred ratio 1.67:1 leads to the highest $F1$-score (and the second best FPR and lowest FNR), while undersampling never performs better than the original data ratio in terms of accuracy and $F1$-score. This can be explained by the fact that the latter eliminates useful information. This prompts us to use oversampling to achieve the 1.67:1 ratio when training classifiers, which turns D_1 into D_1' (i.e., the training set is augmented).

Figure 5 further highlights the *confusion matrix* of the experiment one the same test set but corresponding to D_1

TABLE II
IMPACT OF THE MALICIOUS:BENIGN RATIO ON THE EFFECTIVENESS OF THE RANDOM FOREST CLASSIFIER WITH *Oversampling* AND *Undersampling*, WHERE D_1 WITH RATIO 3.1:1 IS THE ORIGINAL D_1.

Dataset	Method	*Ratio*	ACC	FPR	FNR	$F1$-score
D_1	(none)	3.1:1	0.980	0.030	0.017	0.987
D_1	Oversample	2:1	0.980	0.030	0.018	0.986
D_1	Oversample	1.67:1	0.980	0.027	0.017	0.988
D_1	Oversample	1.43:1	0.979	0.028	0.019	0.986
D_1	Oversample	1.25:1	0.979	0.028	0.018	0.986
D_1	Oversample	1.11:1	0.979	0.027	0.019	0.986
D_1	Oversample	1:1	0.979	0.026	0.019	0.986
D_1	Undersample	2:1	0.977	0.023	0.022	0.985
D_1	Undersample	1.67:1	0.976	0.023	0.025	0.984
D_1	Undersample	1.43:1	0.975	0.023	0.025	0.984
D_1	Undersample	1.25:1	0.972	0.020	0.031	0.981

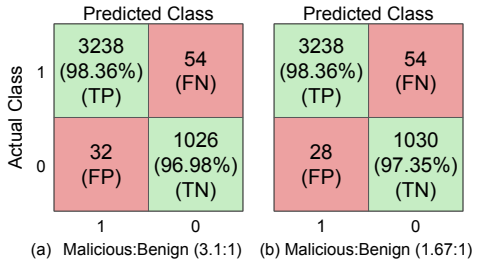

Fig. 5. Confusion matrix for (a) D_1 with 3.1:1 malicious:benign ratio in the training data and (b) D_1' with 1.67:1 ratio in the training data.

and D_1', which shows a slight improvement in detection when augmenting the training set with oversampling.

***Insight** 5:* The data imbalance issue does not affect the model performance significantly in this case, perhaps because the degree of imbalance is not severe enough.

3) Training and Test: Having addressed the issue of feature selection and data imbalance, we consider the following classifiers: Random Forest (RF), Decision Tree (DT), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Specifically, we use the python `sklearn` module to import the following classifier algorithms: (i) Random Forest or RF with parameters `n_estimator`=100 (i.e., 100 trees in a forest) and `criterion`=*'entropy'* (i.e., entropy is used to measure information gain); (ii) K-Nearest Neighbor or KNN, with parameters `n_neighbors`=8 (i.e., 8 of neighbors are considered), `metric`=*'minkowski'* with $p = 2$ (i.e., the Minkowski metric with $p = 2$ measures the distance between two feature vectors), and the rest parameters are the default values; (iii) Decision Tree or DT with default parameters; (iv) Logistic Regression or LR with default parameters; (v) Support Vector Machine or SVM with `linear` kernel and other default parameters. For voting the outputs of the five classifiers mentioned above, we use the `VotingClassifier()` function and set `voting`=*'hard'* (i.e., majority voting). We always considering the 80-20 splitting of the scaled training-test data.

4) Answering RQ5 and RQ6: In order to answer RQ5 and RQ6, we conduct the following experiments, where we use the 80-20 train-test splitting of D_1 and then augmenting the training set as mentioned above. Our experiments are conducted on a virtual machine on https://www.chameleoncloud.org/, running CentOS 7 on a machine of an x86_64 processor with 48 cores and CPU frequency 3.1 GHz.

- Experiment (Exp.) 1: Use the lexical, statistical, and TLD features (i.e., F5, F7, F9-F11) only, while ignoring the WHOIS features. (This experiment is equally applicable to D_2, which is not reported owing to space limitation.)
- Experiment (Exp.) 2: Use the WHOIS features (i.e., F1-F4), while ignoring all other features.
- Experiment (Exp.) 3: Use both lexical and WHOIS features (i.e., F1-F5, F7, F9-F11).

TABLE III
EXPERIMENTAL RESULTS ON DATASET D_1' WITH A RANGE OF CLASSIFIERS (WITH OVERSAMPLING), THEIR TOTAL CPU TIMES FOR TRAINING AND TEST: EXP. 1 USES LEXICAL FEATURES ONLY; EXP.2 USES WHOIS FEATURES ONLY; EXP. 3 USES BOTH LEXICAL AND WHOIS FEATURES.

Exp.	Classifier	ACC	FPR	FNR	$F1$-score	Execution Time(s)
1	RF	0.924	0.150	0.052	0.950	0.48
2	RF	0.977	0.025	0.023	0.985	0.59
3	RF	0.980	0.027	0.017	0.988	0.64
1	KNN	0.887	0.199	0.086	0.925	0.40
2	KNN	0.949	0.034	0.056	0.966	0.25
3	KNN	0.947	0.031	0.060	0.964	0.30
1	DT	0.917	0.151	0.061	0.945	0.07
2	DT	0.973	0.045	0.022	0.982	0.08
3	DT	0.974	0.051	0.019	0.983	0.14
1	LR	0.885	0.216	0.082	0.924	20.30
2	LR	0.883	0.362	0.038	0.926	23.03
3	LR	0.918	0.178	0.051	0.946	44.40
1	SVM	0.888	0.220	0.078	0.925	1.69
2	SVM	0.881	0.373	0.038	0.924	1.68
3	SVM	0.920	0.164	0.054	0.946	2.38
1	Ensemble	0.916	0.171	0.056	0.945	21.40
2	Ensemble	0.962	0. 031	0.041	0.974	24.75
3	Ensemble	0.970	0.035	0.028	0.980	45.70

Table III summarizes the experimental results with a range of classifiers and the actual time spent on training a model and classifying the entire test set. We make several observations. First, for a specific classifier, using WHOIS features alone (Exp. 2) almost always leads to significantly higher effectiveness than using lexical features alone (Exp. 1), except for Logistic Regression. Second, for a fixed classifier, using both lexical and WHOIS features together (i.e., Exp. 3) always performs better than using lexical or WHOIS features alone. Third, among the classifiers considered, Random Forest performs the best in every metric in each experiment. In particular, Random Forest (i.e., non-linear classifier) achieves a better performance than the Ensemble method because there are classifiers (e.g., Logistic Regression and SVM) that are substantially less accurate than the other classifiers and therefore "hurt" the voting results. Fourth, Decision Tree has the fastest execution time, followed by KNN and Random Forest, while Logistic Regression is the slowest and causes a delay for the voting ensemble. To understand the generalizability, when conducting Exp. 1 on the augmented D_2'

with the benign:malicious ratio at 1.25:1, we observe that Random Forest outperforms other models by achieving a 0.947 accuracy, a 0.066 FPR, a 0.041 FNR, and a 0.947 F1-score.

***Insight** 6:* COVID-19 themed malicious website detectors must consider WHOIS features; and Random Forest performs the best among the classifiers that are considered.

V. Discussion

The present study has several limitations, which should be addressed in future studies. First, we use a heuristic method to determine the ground truth. This heuristic method can only approximate the ground truth because the data sources (i.e., CheckPhish and DomainTools feeds in this case) may contain some errors. Second, we could not avoid the data imbalance problem, meaning that the resulting detectors or classifiers may be slightly biased towards the majority class even after the oversampling. Third, we only considered the WHOIS and URL lexical features, but not the website contents or the network layer features, Fourth, we only considered five WHOIS features because most of the other kinds of WHOIS information are largely missing, which means that WHOIS registrars need to collect more detailed information than what is presented at the moment of writing. Fifth, application of deep learning models or explainable ML are left to future research. Sixth, we observe that the python library `wordninja` can make bad splits at times (e.g., when a domain name is seemingly in English characters but actually in another languages).

VI. Conclusion

We have presented the first systematic study on *data-driven* characterization, and detection of COVID-19 themed malicious websites. We presented a methodology and applied it to a specific dataset. Our experiments led to several insights, highlighting that attackers are *agile*, *crafty*, *economically incentivized* in waging COVID-19 themed malicious websites attacks. Our experiments show that Random Forest can serve as an effective detector against these attacks, especially when WHOIS information about websites in question is available. This highlights the importance of domain registrars to collect more information when registering domains in future.

Acknowledgement. We thank the reviewers for their useful comments. This work was supported in part by ARO Grant #W911NF-17-1-0566, ARL Grant #W911NF-17-2-0127, and the NSA OnRamp II program.

References

[1] J. T. PATRICK UPATHAM, "Amid covid-19, global orgs see a 148% spike in ransomware attacks; finance industry heavily targeted," https://www.carbonblack.com/2020/04/15/amid-covid-19-global-orgs-see-a-148-spike-in-ransomware-attacks-finance-industry-heavily-targeted/, 2020, accessed on 10 June, 2020.

[2] Z. Zorz, "Spotting and blacklisting malicious covid-19-themed sites," https://www.helpnetsecurity.com/2020/04/07/covid-19-malicious-sites/, 2020, accessed on 12 August, 2020.

[3] Y. Liang and X. Yan, "Using deep learning to detect malicious urls," in *Proc. IEEE ICEI*, 2019, pp. 487–492.

[4] O. Christou, N. Pitropakis, P. Papadopoulos, S. McKeown, and W. J. Buchanan, "Phishing url detection through top-level domain analysis: A descriptive approach," in *ICISSP*, 2020.

[5] M. Chatterjee and A. Namin, "Detecting phishing websites through deep reinforcement learning," in *Proc. IEEE COMPSAC*, 2019, pp. 227–232.

[6] Y. Li, Z. Yang, X. Chen, H. Yuan, and W. Liu, "A stacking model using url and html features for phishing webpage detection," *Future Gener. Comput. Syst.*, vol. 94, pp. 27–39, 2019.

[7] O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from urls," *Expert Systems with Applications*, vol. 117, pp. 345 – 357, 2019.

[8] L. Xu, Z. Zhan, S. Xu, and K. Ye, "Cross-layer detection of malicious websites," in *Third ACM Conference on Data and Application Security and Privacy (CODASPY'13)*, 2013, pp. 141–152.

[9] D. Liu, J. Lee, W. Wang, and Y. Wang, "Malicious websites detection via cnn based screenshot recognition*," in *International Conf. on Intelligent Computing and its Emerging Applications*, 2019, pp. 115–119.

[10] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Learning to detect malicious urls," *ACM TIST*, vol. 2, no. 3, pp. 30:1–30:24, 2011.

[11] H. M. Junaid Khan, Q. Niyaz, V. K. Devabhaktuni, S. Guo, and U. Shaikh, "Identifying generic features for malicious url detection system," in *Proc. IEEE UEMCON*, 2019, pp. 0347–0352.

[12] R. Verma and A. Das, "What's in a url: Fast feature extraction and malicious url detection," in *Proc. ACM IWSPA'17*, 2017, p. 55–63.

[13] F. D. Abdi and L. Wenjuan, "MALICIOUS URL DETECTION USING CONVOLUTIONAL NEURAL NETWORK," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1155304

[14] W. Yang, W. Zuo, and B. Cui, "Detecting malicious urls via a keyword-based convolutional gated-recurrent-unit neural network," *IEEE Access*, vol. 7, pp. 29 891–29 900, 2019.

[15] L. Xu, Z. Zhan, S. Xu, and K. Ye, "An evasion and counter-evasion study in malicious websites detection," in *Proc. IEEE CNS*, 2014, pp. 265–273.

[16] J. Mireles, E. Ficke, J. Cho, P. Hurley, and S. Xu, "Metrics towards measuring cyber agility," *IEEE T-IFS*, vol. 14, no. 12, pp. 3217–3232, 2019.

[17] Z. Zhan, M. Xu, and S. Xu, "Characterizing honeypot-captured cyber attacks: Statistical framework and case study," *IEEE T-IFS*, vol. 8, no. 11, 2013.

[18] ——, "Predicting cyber attack rates with extreme values," *IEEE T-IFS*, vol. 10, no. 8, pp. 1666–1677, 2015.

[19] Y. Chen, Z. Huang, S. Xu, and Y. Lai, "Spatiotemporal patterns and predictability of cyberattacks," *PLoS One*, vol. 10, no. 5, p. e0124472, 05 2015.

[20] M. Xu, K. M. Schweitzer, R. M. Bateman, and S. Xu, "Modeling and predicting cyber hacking breaches," *IEEE T-IFS*, vol. 13, no. 11, pp. 2856–2871, 2018.

[21] S. Xu, "Cybersecurity dynamics: A foundation for the science of cybersecurity," in *Proactive and Dynamic Network Defense*, 2019, pp. 1–31.

[22] R. Zheng, W. Lu, and S. Xu, "Preventive and reactive cyber defense dynamics is globally stable," *IEEE TNSE*, vol. 5, no. 2, pp. 156–170, 2018.

[23] H. Chen, J. Cho, and S. Xu, "Quantifying the security effectiveness of firewalls and dmzs," in *Proc. HoTSoS'2018*, 2018, pp. 9:1–9:11.

[24] M. Pendleton, R. Garcia-Lebron, J. Cho, and S. Xu, "A survey on systems security metrics," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 62:1–62:35, 2016.

[25] H. Chen, J. Cho, and S. Xu, "Quantifying the security effectiveness of network diversity," in *Proc. HoTSoS'2018*, 2018, p. 24:1.

[26] CheckPhish, "Covid-19 (coronavirus) phishing scam tracker," https://checkphish.ai/coronavirus-scams-tracker, 2020, accessed on 15 May, 2020.

[27] DomainTools, "Free covid-19 threat list - domain risk assessments for coronavirus threats," https://www.domaintools.com/resources/blog/free-covid-19-threat-list-domain-risk-assessments-for-coronavirus-threats, 2020, accessed on 14 May, 2020.

[28] D. Hubbard, "Cisco umbrella 1 million," https://umbrella.cisco.com/blog/cisco-umbrella-1-million, 2016, accessed on 13 June, 2020.

[29] D. Anderson, "wordninja 2.0.0," https://pypi.org/project/wordninja/, 2019, accessed on 12 June, 2020.

[30] Wikipedia, "Shannon entropy," https://en.wiktionary.org/wiki/Shannon-entropy, 2020, accessed on 3 June, 2020.

[31] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.