PharmaGuard: Automatic Identification of Illegal Search-Indexed Online Pharmacies

Igino Corona*, Matteo Contini*, Davide Ariu*, Giorgio Giacinto*, Fabio Roli†, Michael Lund† and Giorgio Marinelli†

*DIEE, University of Cagliari, Italy †Danish Institute of Fire and Security Technology, Denmark

Abstract—We present PharmaGuard, a novel system for the automatic discovery of illegal online pharmacies, aimed at assisting law-enforcement toward their early identification, blacklisting and shutdown. Given a previously labelled set of examples, the system is able to learn a profile of (illegal) pharmacies, and then exploit it to discover never-before-seen instances indexed by popular web search engines. Our experiments, performed on webpages found in the wild, indicate that our approach is lightweight, allows for high accuracy and can substantially complement state-of-the-art blacklists. We also present a report on the detected online pharmacies that better highlights the relevance of this threat for Internet users.

Index Terms—Detection of Illegal Pharmacies; Search Engines; Pattern Classification; Human-Machine Interaction

I. INTRODUCTION

The advent of the world wide web deeply transformed the way we live and interact with other people. Unfortunately, it also introduced new –easy– ways for cyber-criminals and miscreants to make money without qualms about other people’s rights and health. A clear example of this is the increasingly rising number of websites selling pharmaceutical products whose prescription would require a strict medical assistance, to anyone in the world is willing to pay for. In addition, such products are typically sold without any guarantee about their actual composition and quality. This can cause severe damages to the health, and even lead to the death of people who irresponsibly –or ignorantly– buy and use these products [1].

The US Food and Drug Administration (FDA), but also international organizations such as Interpol, are actively fighting such cyber-criminal activities [2], [3]. In June 2013, FDA, though the collaboration of various law enforcement agencies, seized and shut down about 1,700 illegal websites [4]. In May 2014, more than 11,800 illegal online pharmacies were disrupted through the collaboration of law enforcement agencies in more than one hundred countries. The operation was coordinated by Interpol and resulted in more than two hundred arrests worldwide and the seizure of potentially dangerous pharmaceuticals worth more than 32 million US dollars [2].

In spite of these efforts, websites that illegally sell prescription drugs still constitute a concrete, persistent threat. First of all, this phenomenon is relatively new, and for this reason, many countries still miss a clear legislation in this regard. Moreover, laws across different countries are typically heterogeneous. This allows cyber-criminals to persistently deploy such malicious websites, with little effort. These websites might be even hosted on countries where legislation explicitly forbids them (e.g., US), since law enforcement authorities are often ill-equipped to find and investigate them. An important point is that new victim users may easily reach such sites through simple queries on web search engines, that apparently provide little to no protection against this kind of threat. Notably, the search giant Google has been even prosecuted for illegal profits from those activities [5].

In this paper, we present a novel system, named PharmaGuard, capable to assist law enforcement operators toward a quick identification, blacklisting and seizure of illegal online pharmacies. The system learns autonomously the profile of illegitimate online pharmacies given a set of examples, and then detects with high accuracy never-before-seen instances appearing from search engine results. Such profile can be updated through human interaction, as soon as new webpages are labelled by an operator as either legitimate, illegitimate pharmacies, or other webpages. PharmaGuard is fruit of a collaboration between the University of Cagliari and Danish Institute of Fire and Security Technology (DBI), an independent, non-profit enterprise, comprising more than 3,000 employers, which is actively fighting illegal pharmacies for the Danish government.1

The paper is organized as follows. Section II presents a summary of related work, and better highlights our contribution with respect to the state of the art. The architecture of PharmaGuard is described in Section III. Our experimental evaluation of PharmaGuard is presented Section IV. We discuss about the limitations of our proposal and future research work in Section V. Section VI presents our conclusions and closes the paper.

II. RELATED WORK

Most of related work focuses on the investigation of specific spam campaigns, or advertisement techniques, where pharmacy scams may appear as just one of a plethora of supported illicit activities. To enable a better understanding of the infrastructure and the economic incentives behind spam campaigns, several researchers infiltrated real botnets, such as Storm (see, e.g., [6]), but also used active probing and

1Thus, throughout the paper, the term illegal is used to refer to activities that do not comply with the European legislation concerning the online market of drugs and pharmaceuticals. Please note, however, that albeit all these websites are indeed related to pharmacy scams, in some countries might not be recognized as illegal, typically because of a missing/weak legislation on this matter.
Some studies have also investigated the illegal market of drugs in the deep web [8] right before the seizure of Silk Road [9], as well as its relationship with some scam campaigns in the web [10]. A well-known advertisement technique employed by miscreants to support pharmacy scams is search engine poisoning that exploits the popularity of legitimate (but vulnerable) websites [11], [12]. The so-called “Made for AdSense” (MFA) websites constitute another way to publicize pharmacy scams. Such websites are built with the sole goal of making revenue through the association of popular “trending” terms with a set of advertised links. The goal is to exploit web search engines to attract new victim users that click on such links [13].

In two related studies [14], [15], link-based and content-based algorithms originally devised for web-spam detection were employed towards the discrimination between legitimate and fake medical websites (including fake online pharmacies). However, whereas such works obtained good results in terms of discrimination between legitimate and fake pharmacy webpages, it is not clear whether they are able to discriminate between online pharmacies and any other site that may appear, e.g., in search engine results. Moreover, due to their computational complexity, such solutions may not be suitable for practical deployment, i.e., the analysis of a large number of websites per day.

**Contributions** Compared to previous work, we explicitly focus on the automatic identification of novel pharmacy scams indexed by web search engines, or linked to known (previously classified) illegal online pharmacies, regardless the employed spam campaigns or advertisement techniques. As we will see in §IV, this allows for detecting instances that do not appear in state-of-the-art blacklists. The approach adopted by our crawler is similar to that used in [16] to find candidate pages to match against drive-by-download attacks. However, our classification engine is explicitly tailored to the detection of illegal online pharmacies. Moreover, our system supports by design the interaction with a human expert: that allows for enhancing detection accuracy as soon as new webpages are validated.

**III. Architecture**

Cyber-criminals need to advertise their scams to attract new victims and increase their revenues. For this reason, the first module of PharmaGuard is a focused web crawler: its goal is to scan the web for the identification and download of websites that are potentially related to online pharmacies, focusing on those indexed by web search engines, that can be easily reached by million users (candidate websites, see Section III-A).

To determine the actual activity behind a candidate website identified in the previous stage, a fine-grain analysis is clearly necessary. To this end, a second module of the infrastructure contains: (a) the illegal pharmacy detector (Section III-B); (b) the advertisement detector (see Section III-C). The first detector aims at providing law enforcement with a prioritized list of websites to be further investigated, i.e., most likely related to illegal online pharmacies. The second detector aims to detect novel advertisement websites, i.e., websites that are actively employed by miscreants to publicize, rank and suggest one or more illegal pharmacy websites, providing the related URLs. While these websites might not actually sell pharmacy products, they constitute an important source of information for the discovery of novel illegal websites.

Figure 1 shows a full picture of the PharmaGuard infrastructure. Law enforcement operators may provide their feedback, updating: (a) the set of legitimate/illegitimate online pharmacies; (b) the set of search keywords to be employed by the web crawler; (c) the set of known advertisement websites.

**A. Web Crawler**

The objective of this module is to find a set of URLs to be inspected by the analysis engine. In the current implementation, we collect URLs from web search engines, known advertisement and illegal pharmacies. In particular, the module actively queries web search engines for indexed pages corresponding to ad-hoc search queries that characterize the typical content of the known illegal pharmacies. For instance, search queries may look for known products sold (illegally) online, or to webpages which point to known advertisement websites. The web crawler also downloads each webpage that corresponds to the search results and automatically extracts all displayed textual content, as well as all embedded links. To this end, we exploited Selenium [17], a browser automation framework. Please note that this is a relevant step, because some content, e.g., generated/downloaded through JavaScript or Flash code, Cascade Style Sheets (CSS), HTML frames, might not be visible without using a real browser.

**B. Illegal Pharmacy Detector**

To approach the problem complexity, we split this classification task into two subtasks in cascade (see Figure 2): (a) identification of online pharmacies vs other websites (Pharma vs Other), and (b) identification of legitimate vs illegitimate pharmacies (Pharma vs Pharma). The illegitimate activities identified by the latter module are prioritized and sent to a human operator for further inspection. The operator may validate or correct the predicted label, and thus update the ground truth available to the system.

Pharma vs Other Online pharmacies can be characterized by some typical, distinguishing features with respect to other, generic websites. In particular, such websites may be characterized by the presence/advertisement of (a) pharmacy products (b) shopping cart/basket; (c) payment methods; (d) customer care; (e) shipping methods. In our investigation, we found that these features can be described through the presence of special keywords, that can be determined through a statistical analysis of textual content related to either online pharmacy or other webpages (see Section III-B1). Afterwards, a supervised machine learning algorithm can use these keywords to perform an accurate detection of online pharmacies, in a fully automated way. Upon human supervision, such
keywords can be also exploited by the web crawler (see Section III-A) to build search queries in order to detect novel online pharmacies throughout the web.

Pharma vs Pharma All webpages identified in the previous stage as online pharmacies (see Section III-B) undergo a further analysis in order to understand whether they can be considered as legitimate or not. From a conceptual point of view, illegitimate webpages are characterized by the presence of pharmacy products that either (a) cannot be sold online [10], or (b) would require a medical prescription, but are sold skipping this requirement. Albeit this concept may be rather easy to understand, it is not straightforward for a human operator to apply it in practice for the detection of illegitimate webpages. Even more difficult is codify it within a program, in a general way. We would require to know a full list of products that fall within category (a) or (b) and detect their presence within a webpage, but this process is tedious and error prone. Instead, as in the detection of online pharmacies, we apply a fully automated machine learning algorithm based on the analysis of textual content within either illegitimate or legitimate online pharmacies, as explained in Section III-B1. The set of extracted keywords allowed us to somewhat describe features (a) and (b), and thus obtain very accurate results.

1) Classification algorithm: For both detectors in Figure 2, we resort to a fully automated analysis based on Term Frequency-Inverse Document Frequency (TF-IDF). This is a well-known technique for information retrieval tasks and document classification. In particular, we firstly strip all stop words from within each webpage $p$. Each term (word) $t$ within page $p$ receives a weight $w_{t,p}$, computed as follows:

$$w_{t,p} = \begin{cases} 
(1 + \log(tf_{t,p})) \cdot \log\left(\frac{N}{pf_{t}}\right) & \text{if } tf_{t,p} > 0 \\
0 & \text{otherwise} 
\end{cases} \quad (1)$$

where $tf_{t,p}$ is the number of times term $t$ appears in page $p$; $pf_{t}$ is the number of pages that $t$ occurs in; $N$ is the number of webpages in the collection. For each webpage $p$ we then reshape its weights $w_{t,p}$ using cosine ($l_2$) normalization, i.e., such that $\sum_{t\in p} w_{t,p}^2 = 1$. This normalization has been found very effective combined with the weights computation formula in Eq. 1, in many well-known document classification tasks [18].

At this point, each webpage $p$ is described by a (sparse) vector of weights $w_{t,p}$. We use this vector to learn a linear classifier using samples related to either online pharmacies or other websites (first detector), and either legitimate or illegitimate online pharmacies (second detector). Please note that, in this application, the use of linear classifiers is somewhat a must, since the number of features might be very high (larger than the number of training samples), and more complex classifiers are likely to overfit training data [19].

C. Advertisement Detector

The advertisement detector aims to detect any website that advertises other websites that sell pharmacy products online (“advertiser”). Its current implementation is simple, but effective. We rank each website according to the number of links associated to (previously validated) illegal online pharmacies. The first websites in this rank are considered as candidate advertisement websites and made available to a human operator for further inspection. Ad-hoc parsing mechanisms may be implemented to allow the web crawler (see Section III-A) to automatically parse the detected “advertisers” and obtain new URLs to be inspected.

IV. EXPERIMENTAL EVALUATION

We experimentally evaluated PharmaGuard to assess: (a) its accuracy when discriminating illegitimate online pharmacies from other kind of websites, including legitimate online pharmacies, (b) its learning time and throughput (c) its complementarity with respect to state-of-the-art tools. To this end, we were able to build and then employ a dataset of more than 1,200 websites (see §IV-A and §IV-B). Our experiments aim also to give to the reader more insights into the characteristics

---

Fig. 1. Main architecture of PharmaGuard.

Fig. 2. A logical scheme of the Illegal Pharmacy detector of PharmaGuard.
of illegal online pharmacies in the wild, as well as of the threat posed drugs sold online. To this end, in Section IV-D we investigate the detected illegal online pharmacies from three main viewpoints: their network source, popularity, and products sold online. Throughout our discussion, we give concrete examples of illegal websites that we identified in the wild and that are still active at the time of writing.

A. Dataset

Obtaining a representative dataset of illegitimate as well as legitimate online pharmacies is a hard, expensive task, due to a lack of ground truth, and the problem complexity. In fact, as we will further see in Section IV-D, many illegitimate websites are very similar to legitimate ones and their classification requires a thorough investigation by a human expert. For instance, many of them (a) warn their users against fake pharmacies (ironically), (b) identify themselves as “approved by FDA” or other reputable organizations, (c) provide detailed information about (phantasmic) laboratory tests and pharmacy statistics, (d) allow users to chat with each other, (e) are characterized by a nice-looking graphical interface, (f) are very popular, just like many legitimate web pharmacies.

For this reason, we needed about 240 men-hours (one month worth) for the collection and labelling of data. To this end, we exploited the expertise of DBI investigators actively working in the field, and “bootstrapped” the infrastructure of PharmaGuard to download and analyze the webpages. In particular, PharmaGuard was bootstrapped (a) using manually-defined keywords related to known illegally sold products for searching the web and analyze the content of webpages, (b) extracting links from two known advertisement websites, i.e., musclegurus.com, eroids.com. At the end of this phase we were able to obtain three datasets:

- \( L \): set of 172 legitimate online pharmacies. This set was mainly built using LegitScript (www.legitscript.com) as reference. LegitScript is an US organization that provides a list of websites that legally sell pharmacy products on-line and whose quality/safety is certified. All such pharmacies may be considered as legal also for EU citizens.

- \( I \): set of 446 illegitimate online pharmacies. This set was built using both webpages known as illegal to our research partner, or pages that we were able to verify as illegal through the employment of PharmaGuard in the bootstrap setting, and the advice and expertise of our research partner.

- \( O \): set of 647 other webpages. This set was built incrementally, during the bootstrap phase: we added to this set every webpage that were downloaded by the web crawler, but that were not related to online pharmacies.

B. Detection of Illegal Online Pharmacies

In order to evaluate the effectiveness of the classification algorithm in Section III-B1 for Pharma vs Other and Pharma vs Pharma, we experimented with \( D_1 = L \cup I \cup O \), and \( D_2 = L \cup I \) datasets, respectively. Both datasets were randomly split into two disjunct portions: the first one (70%) was used to train the classification algorithm (training split), the remaining one (30%) was used for evaluating its accuracy under operation (testing split). In particular, this process was repeated for \( N=5 \) runs, to estimate the average accuracy of the classifier, as well as its standard deviation. Such procedure is widely adopted in the pattern recognition field to evaluate the capability of a classifier (detector) to correctly label never-before-seen samples (i.e., samples that were never seen in training data) [19].

We employed the classification algorithm presented in Section III-B1 to automatically select and weight a set of textual keywords for distinguishing between online pharmacy (\( L \cup I \)) and other (\( O \)) websites (Pharma vs Other) and then distinguishing between illegal (\( I \)) and legal online pharmacy (\( L \)) (Pharma vs Pharma). A linear classifier was built to actually perform the detection task, using the selected set of features. We compared several different learning techniques, based on different notions of loss function\(^2\), that may work better or worse depending on the application. For each technique, we optimized their parameters to maximize the Area Under the ROC curve (AUC), using grid-search and a three-fold cross-validation technique. In our experiments, we found the Stochastic Gradient Descent (SGD) as the most accurate technique. An interesting aspect of SGD learning is also that it can be performed incrementally, i.e., it is a perfect choice for our application, because new webpages validated by a human operator can be easily added to training data.

In this application, it is desirable to detect as much as possible online pharmacies, in order to further evaluate their legitimacy. For this reason we are particularly interested in the operating points for which a maximum detection rate is attained. For a true positive rate (TPR) of 100%, the Pharma vs Other and Pharma vs Pharma classifiers were able to achieve a false positive rate (FPR) of 1.64% (± 0.6%) and 1.09% (± 1.2%), respectively.

An important point, however, is that false positives generated by the Pharma vs Other classifier may cause additional false positives on the Pharma vs Pharma classifier. Thus, we evaluated the fraction of misclassified “other” webpages that would be also misclassified as illegitimate online pharmacy by the Pharma vs Pharma classifier. These false positives were the 0.41% (± 0.38%) of the total “other” webpages. In order to evaluate the total mean and standard deviation of false positives, either related to legitimate pharmacy or other webpages, we can resort to the formula in [20]. Table I shows a summary of the accuracy of the two classifiers, as well as the result of such computation, that leads to an overall false positive rate of 0.59% (± 0.57%). This value is particularly low and thus allows for an accurate detection of illegitimate pharmacy activities over a large number of websites.

Learning and Classification time On a commodity computer\(^3\), using a single thread, the whole learning process...
takes about 40 (Pharma vs Other) and 18 (Pharma vs Pharma) seconds in average. Updating the detection model through the addition of a new sample to training data takes a negligible amount of time (less than one second). The classification process requires a negligible amount of time (5 msec). Please note however that parsing a webpage requires 0.5 seconds in average, and this is the average amount of time needed to actually analyze a webpage. This translates into an average throughput of 172,800 webpages per day (per thread on a commodity computer).

C. Comparison with other tools

We matched our results against 15 reputable tools, namely, DNS-BH, DShield, Feodo Tracker, Google SafeBrowsing, Malc0de, Malwarebytes PhishHosts, MalwareDomainList, OpenPhish, PhishTank, Spam404, Spamhaus DBL, SURBL, Yandex SafeBrowsing, Zeus Tracker, providing domain and URL blacklists for a wide range of scams, including phishing, malware, spam. Only 5.24% of the 446 malicious websites detected by means of PharmaGuard were also listed in such blacklists. In particular, Google Safebrowsing, that is used as a default protection for many popular web browsers such as Firefox, Safari and Chrome, was able to detect only 0.06% of such illegal pharmacies. This basically means that so many web users are currently unprotected against this kind of threat.

D. Illegal Online Pharmacies in the wild

1) Network Source: According to our data, a single domain name associated to illicit pharmacy typically resolves to a unique (network of) IP addresses, and thus is typically associated to one or few Autonomous Systems (ASs), just like legitimate online pharmacies. Given this behavior, in Table II we show the hosting organizations / ASs that are mostly involved in these illegitimate activities. Notably, such organizations are not malicious per se, but may be characterized by permissive hosting/control policies, or special features (such as the CloudFlare’s IP anonymity) that ease the deployment of such illicit websites.

2) Domain Names and Popularity: Another relevant aspect of illegal online pharmacies is how much they are popular. That is, how many users could potentially fall victim of such scams. To evaluate this aspect, we ranked each domain name according to Alexa (www.alexa.com) (see Table III). Interestingly, but also scaringly, there are many illegal websites with a high popularity and lifetime, in the web. For each domain name we also listed its rank within the advertisement communities of musclegurus.com and eroids.com. Each domain name in Table III is active at the time we are writing this paper.

3) Top Illegally-sold Prescription Drugs: As mentioned in the introduction, there is no guarantee about the actual composition of the products bought in illegal pharmacies. Only this aspect may lead to severe adverse effects: think for instance to allergenic, injectable substances. But even if miscreants sell exactly what they claim, some drugs may be lethal if used without strict medical advice and control. Among the top drugs illegally-sold online we identified: Peptides, Sustanon, Stanozolol, Prolixin Enanthate, Trenbolone, Clenbuterol, Human Growth Hormone (HGH). All such products have been identified by looking to the most discriminating keywords extracted automatically by the classification algorithm described in Section III-B1, while discriminating between legitimate and illegitimate online pharmacies (Pharma vs Pharma classifier). The adverse effects of these products are
very wide and sometimes largely unknown. Among them, we highlight: death, strokes and heart attacks, irreversible tardive dyskinesia, irreversible citoomedaly, oligospermia, urinary obstruction, priapism, edema. These effects can be verified by looking for the above substances into the reputable website http://www.drugs.com.

V. LIMITATIONS AND FUTURE WORK

As discussed in §III-A, textual content is able to accurately reflect the activities behind a web page, since online pharmacies (both legal and illegal) need to advertise their products and services by means of words and sentences. This is the main reason why our approach is so effective. Nevertheless, our analysis may be attacked by a skilled adversary who wants to stay under the radar. For instance, cyber-criminals might hide or add some displayed text in the attempt to cause the algorithm in §IV-B to misclassify an illegitimate pharmacy as an “other” website (evasion attack). Please note, however, that displayed text is considered by search engines as well: this behavior may thus lead to poorly indexed illegal pharmacies, i.e., it may be counterproductive for the adversary. Moreover, hidden text can still be detected using image analysis techniques [23].

Whereas in our evaluations we employed a fairly large set of illegal online pharmacies (please recall that all webpages have been thoroughly validated by a human), it is still limited compared to those that are actually available in the web. Thus, we plan to extend our dataset, as long as PharmaGuard is used by our research partners at DBI, and law enforcement operators during their investigations.

VI. CONCLUSION

The pervasiveness of illegal online pharmacies as well as the adverse effects of the substances they sell online, constitute a real, important threat. In this work, we presented a novel architecture, PharmaGuard, that can automatically discover such illicit activities advertised throughout the web and indexed by popular web search engines. Our experimental results show that our system is accurate, lightweight and may substantially complement current blacklists. PharmaGuard is able to significantly reduce the burden of the analysis task by law enforcement operators, letting them to focus only on webpages that are most likely related to illegal online pharmacies. In this paper, we also reported on the main characteristics of the detected pharmacy scams, highlighting the related risks.

ACKNOWLEDGMENTS

The authors gratefully thank Guido Mureddu, who compared the detection results of PharmaGuard with state-of-the-art blacklists and tools. This research has been funded with support from the European Commission (Grant Agreement HOME/2012/ISEC/AG/4000004360). This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES