Augmenting phishing squatting detection with GANs

Original
Augmenting phishing squatting detection with GANs / Valentim, Rodolfo; Drago, Idilio; Trevisan, Martino; Cerutti, Federico; Mellia, Marco. - ELETTRONICO. - (2021), pp. 3-4. (Intervento presentato al convegno ACM CoNEXT 2021 - International Conference on emerging Networking EXperiments and Technologies tenutosi a Virtual Event Germany nel 7 December 2021) [10.1145/3488658.3493787].

Availability:
This version is available at: 11583/2943633 since: 2021-12-08T18:16:48Z

Publisher:
ACM

Published
DOI:10.1145/3488658.3493787

Augmenting Phishing Squatting Detection with GANs

Rodolfo Valentim
Politecnico di Torino
Torino, IT
rodolfo.vieira@polito.it

Idilio Drago
Università degli Studi di Torino
Torino, IT
idilio.drago@unito.it

Martino Trevisan
Politecnico di Torino
Torino, IT
martino.trevisan@polito.it

Federico Cerutti
Università degli Studi di Brescia
Brescia, IT
federico.cerutti@unibs.it

Marco Mellia
Politecnico di Torino
Torino, IT
marco.mellia@polito.it

ABSTRACT

Current solutions to tackle phishing employ blocklists that are built from user reports or automatic approaches. They, however, fall short in detecting zero-day phishing attacks. We propose the use of Generative Adversarial Networks (GANs) to automate the generation of new squatting candidates starting from a list of benign URLs. The candidates can be either manually verified or become part of a training set for existing machine learning models. Our results show that GANs can produce squatting candidates, some of which are previously unknown existing phishing domains.

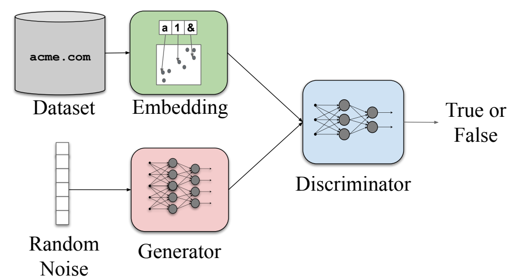

Figure 1: Solution overview showing GAN training.

1 INTRODUCTION AND MOTIVATION

Phishing is a cyber attack in which the attacker tries to convince the victim to reveal personal information through fraudulent messages. When it is coupled with *Cybersquatting*, the attackers register Internet domain names – in the following, *domain* for short – similar to legitimate services to fool the victims in a phishing attempt. The typical defense consists of blocklists, composed of domains, whose timely update and collation are key to block the latest attacks. The management of these lists is time-consuming and often based on human interaction, posing scalability and economic issues. Moreover, it is inefficient against zero-day attacks.

In [3], the authors search and detect squatting phishing domains in the wild. The search process starts by creating a list of squatting candidates. Their method of building such list extends existing tools for generating squatting domains by adding new generation algorithms. The authors generated 657, 663 squatting domain candidates for 702 target brands using 5 different typo-squatting techniques. After verification, 1, 175 of the domain candidates were confirmed to be used for phishing. More alarming, 90% of the found phishing squatting domains evaded blocklists for more than one month, which corroborates the hypothesis that these lists fall show on zero-day attacks.

Machine Learning (ML) comes to the rescue in this picture, allowing generalizing predictions for previously unseen URLs. However, ML models require a considerable amount of labeled training data, often scarce or cumbersome to obtain. Generative Adversarial Networks (GANs) [1] emerged as a tool to generate new samples given a limited set of training examples. In a GAN, a *generator* model generates candidates, and a *discriminator* model evaluates them as real or false. Adversarial training allows the generator to produce realistic samples and the discriminator to generalize the model.

In this preliminary work, we propose the use of GANs to augment domain squatting training datasets. We focus on typo-squatting to start. Our rationale is that GANs can learn data distribution and generate new samples from a target domain, reproducing some or all typo-squatting techniques. We argue that data augmentation could anticipate new cybersquatting attacks and increase classifiers' robustness. We introduce an embedding layer in our GANs, capable of controlling the desired variations in the data generation. These variations allow us to control the desired typo-squatting candidates.

We are not the first to apply generative models to augment data used in cybersquatting identification systems. Current solutions are built on image generation where the domain is converted to an image, and a GAN generates new images that eventually are used to train homograph phishing identifier systems [2]. Our technique instead relies on text and on embeddings, which can be used to guide the generation process introducing similarities and dissimilarities between characters. We illustrate the process with particular use cases.

2 SOLUTION OVERVIEW AND METHODOLOGY

Figure 1 shows an overview of our proposal. Given a dataset of URLs belonging to a given class (e.g., URLs of the website *acme.com*), we train a GAN where the generator learns to generate new samples, and the discriminator shall distinguish real from generated ones. We train a GAN instance for each class of URLs. We use an embedding layer to guide trends in the generation, converting all URLs in its vector space before feeding them to the GAN. Our rationale is to use the embedding so that vectors that represent similar characters are

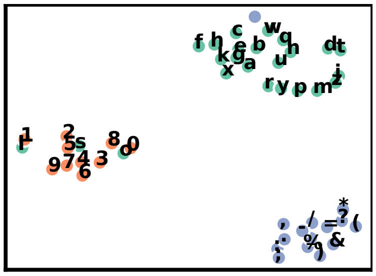

Figure 2: Manual embed tSNE projection for typo-squatting.

Table 1: Samples extracted from generate data.

Examples	Unique samples
ap1.accuveather.c0m api.accuweathet.com	21
micro50ft.com m1cros0ft.com	19
dovnload.w1ndow5update.c0m downl0ad.w1ndow5update.c0m	489

closer to each other, i.e., making those characters that are typically abused for cybersquatting look like “alias”. For instance, an 'o' will be close to a '0'. The vector representation can be learned from data, but, in this preliminary study, we define it manually by making those vectors that represent similar characters to be close. This process can be automatized to include errors humans make with captcha or in textual documents. Exploring such approaches will be the center of our future work. Figure 2 shows a two-dimensional projection of our embedding layer using tSNE. Note the pairs $(o, 0)$, $(5, s)$, $(1, l)$ are closer compared to other characters.

3 PRELIMINARY RESULTS

Our GAN uses Long Short-Term Memory (LSTM) units in the generator and a Convolutional Neural Network (CNN) as the discriminator. In our preliminary experiments, we focus on three domains to generate possible squatting candidates. We evaluate whether the generated domains are possible cybersquatting candidates based upon manual inspection and some automatic checks, such as regular expressions and the edit distance value between the original domain and the generated candidates.

Table 1 reports some examples of cybersquatting domains the GAN generates, while Figure 3 details the frequency of changes across the various characters of the domains. The GAN forms a limited number of unique domains. Among these candidates, we have found 2 registered domains for `api.accuweather.com`, one of which likely to be phishing (determined by manual inspection). We found 4 existing domains for `microsoft.com` of which 3 appear to be phishing in the wild.

In general, the generation abilities of the GAN allows us to obtain likely squatting domain with little domain knowledge and at a low cost when compared to deterministic or image-based solutions. These experiments illustrate the possibility of using a GAN to produce typo-squatting candidates from target brands and testing them even before abuses.

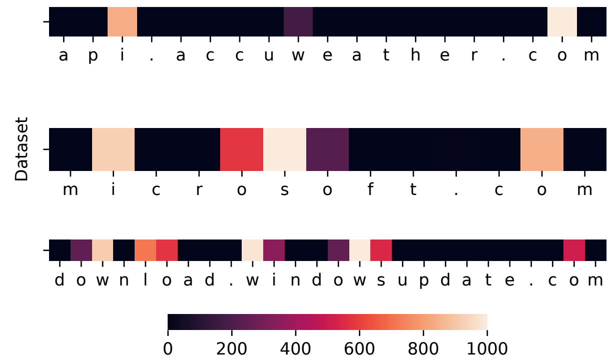

Figure 3: Characters frequently modified by the generator.

4 FUTURE DIRECTIONS

Our work gives solid suggestions and guidelines for GAN-powered automation in dataset generation for cybersecurity. The usage of GAN changes the reactive nature of the blocklists to a proactive search, anticipating new attacks. These results are preliminary and future work will focus on exploring other cybersquatting techniques and automatically learning embeddings that capture behaviors of other (and possibly yet unknown) cybersquatting techniques.

ACKNOWLEDGMENTS

The research leading to these results has been funded by the Huawei R&D Center (France) and the SmartData@PoliTO center for Big Data technologies.

REFERENCES

[1] Ian Goodfellow et al. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
[2] Lee Joon Sern et al. 2020. PhishGAN: Data Augmentation and Identification of Homoglyph Attacks. In *CCCI 2020*. 1–6. https://doi.org/10.1109/CCCI49893.2020.9256804
[3] Ke Tian et al. 2018. Needle in a haystack: Tracking down elite phishing domains in the wild. *Proceedings of the ACM SIGCOMM IMC* (2018), 429–442. https://doi.org/10.1145/3278532.3278569