Panning for gold.com: Understanding the Dynamics of Domain Dropcatching

Najmeh Miramirkhani, Timothy Barron, Michael Ferdman, Nick Nikiforakis
Stony Brook University
[nmiramirkhani,tbarron,mferdman,nick]@cs.stonybrook.edu

ABSTRACT

An event that is rarely considered by technical users and laymen alike is that of a domain name expiration. The massive growth in the registration of domain names is matched by massive numbers of domain expirations, after which domains are made available for registration again. While the vast majority of expiring domains are of no value, among the hundreds of thousands of daily expirations, there exist domains that are clearly valuable, either because of their lexical composition, or because of their residual trust.

In this paper, we investigate the dynamics of domain dropcatching where companies, on behalf of users, compete to register the most desirable domains as soon as they are made available and then auction them off to the highest bidder. Using a data-driven approach, we monitor the expiration of 28 million domains over the period of nine months, collecting domain features, WHOIS records, and crawling the registered domains on a regular basis to uncover the purpose for which they were re-registered (caught). Among others, we find that on average, only 10% of the expired (dropped) domains are caught with the vast majority of the re-registrations happening on the day they are released. We investigate the features that make some domains more likely to be caught than others and discover that a domain that was malicious at the time of its expiration is twice as likely to be caught than the average domain. Moreover, previously-malicious domains are significantly more likely to be reused for malicious purposes than previously benign domains. We identify three types of users who are interested in purchasing dropped domains, ranging from freelancers who purchase one or two domains to professionals who invest more than $115K purchasing dropped domains in only three months. Finally, we observe that less than 11% were used to host web content with the remaining domains used either by speculators, or by malicious actors.

1 INTRODUCTION

The Domain Name System (DNS) is the defacto identity management system on the web, providing human readable IDs called domain names that can be translated to routable IP addresses. These domains, however, are not permanent. An owner pays to register their domain name for a certain period of time after which it will expire unless the owner pays to renew the domain for another period. When a domain is allowed to expire, it gets deleted and then referred to as "dropped." After dropping, the domain name is made available for registration again on a first-come first-served basis. New registrants can race to re-register, or "catch" the domain name and the winner gains full control of the domain name. This is the foundation of the dropcatch industry.

This system may be considered fair, yet ruthless. Such a model can endanger businesses which build their brand and services around their domain name, but forget to renew it as was the case for both Foursquare and the Dallas Cowboys in 2010 [28, 34]. Even more importantly it begets a broad and severe range of security threats. JavaScript libraries, software/operating system updates, and many other services and security protocols depend on domain names. When the associated domain name expires, the new registrant inherits the residual trust of the domain name and can take over its previous clients, visitors, and dependent resources. As Lauinger et al. showed in concurrent work, people are aware of the value of these domain names [24]. Registrars spend millions of dollars supporting the infrastructure to catch valuable domains at the exact moment they become available. Recent work by Lever et al. studied the consequences of residual trust using their system which detects domain ownership changes [26]. There are, however, many aspects of the dropcatch ecosystem which have yet to be studied. In this paper, we analyze the operations of different parties in order to gain a better understanding of the security implications of domain dropcatching. We frame our main contributions and findings as follows:

- **Large-scale data collection system for dropped domains**. We develop DomainPanner to harvest zone files via a distributed search engine, aggregate daily dropped domains, identify caught domains, obtain their WHOIS records and blacklist status, crawl, and characterize them. Our tool collected over 20 TB of data over the duration of our study.
- **Analysis of caught domains with negative residual trust:** We analyze the impact of negative residual trust on domain registration. We discover that re-registration rates are higher among previously blacklisted domains and that these domains are also more likely to become malicious again, serving malware in 94% of cases.
- **Analysis of domain selection strategies:** We study the strategies used for selecting domain names and show how they differ between normal and malicious registrants, across different demographics, and between registrants categorized by scale (Freelancers, Domainers, and Dropcatchers).
- **Study of registrants' intentions:** We cluster the web contents of caught domains in a large-scale study in order to understand the catchers' intentions. We find that 69% of the domains are registered by speculators, exposing users to potentially unwanted content.

WWW 2018, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5639-8/18/04.
https://doi.org/10.1145/3178876.3186092

2 DATA COLLECTION

In this section, we describe the various data sources that we use for our analysis of domain dropcatching and present architecture and implementation details of our tool which utilizes the following sources to extract as much information as possible about dropped and subsequently caught domains.

2.1 Data Sources

- **Zone files:** We collected `.com`, `.net` and `.org` zone files on a daily basis starting from Jan. 10th, 2017 and continuing for nine months until Oct. 10th, 2017. These zone files include the list of registered and active domain names on any given day and, because of their popularity, have a combined file size of multiple Gigabytes for each single day of domain names.
- **Lists of dropped domains:** For nine months starting from Jan. 10th, 2017, we gathered the daily lists of dropping domains from the following drop catch services: SnapNames, DropCatch, Pool, Namejet and Dynadot. Each service provides a list of domain names that will be dropped in the next five days in its own format at varying degrees of completeness. We combined all the individual daily lists and built the daily dropping domains list by performing a majority vote on the drop date reported by different services for each domain. There is more than 90% overlap between the lists of drop catch services for our studied TLDs and we are therefore confident that our aggregate list covers the vast majority of domains in the pending-delete stage (last five days of their lifetime). Through this daily aggregation process, we collected a total of 28,401,974 pending-delete domain names across the `.com`, `.net`, and `.org` TLDs.
- **Domain blacklists:** We used two sources to compile our database of blacklisted domains: Google Safe Browsing (GSB) and VirusTotal (VT). For each domain, we queried the safe browsing API before the dropping, the first day after its registration, and in repeated intervals for the whole duration of our study. While GSB provides the status of a domain name at the time of query, it does not provide any information about the history of the domain's malicious activity. We therefore queried VT which aggregates historical data from a wide range of antivirus products and online scan engines about domains. Given that VT applies strict rate limits, we queried it for a limited period of one month as well as for any supplementary queries that were needed in other parts of our study.
- **WHOIS records:** We collected WHOIS records of all the dropping domains before their drop date to be able to analyze their previous registrations. Moreover, we obtained fresh WHOIS records for all the newly registered domains for 3 months to capture the information of new registrants.
- **Domain features:** Finally, for all dropping domain names, we collected statistics related to their historical traffic, search volume, lexical features, and previous content hosted on them (obtained from the Internet Archive).

2.2 Implementation of DomainPanner

In order to process the aforementioned dropping-related sources on a daily basis, we developed DomainPanner, a system that tracks the re-registrations of deleted domains, crawls the newly registered domains, and extracts features from each domain. In addition to the data sources described in the previous section that are collected and compiled from various online sources, DomainPanner's pipeline (Figure 1) consists of the following components:

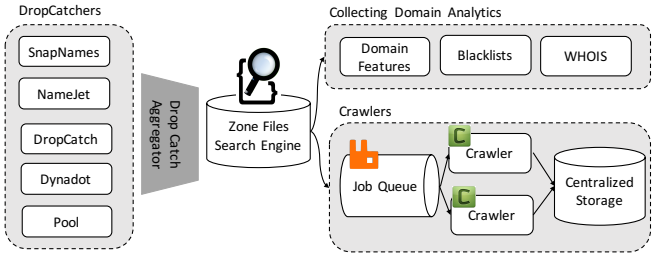

Figure 1: High-level view of DomainPanner.

Zone file search engine: By the time a domain name enters the pending-delete state it is no longer present in the zone file, but if it is re-registered, it is added back into the zone file with new records. We therefore utilize the re-appearance of a dropping domain in the zone files to indicate that the domain was caught. To be able to search through Terabytes of zone files efficiently for the purpose of our longitudinal study, we make use of Elasticsearch. Since most zone file entries remain unchanged between consecutive days, we calculate the delta between zone files of two sequential days. This delta, which contains information about domains that were registered, de-registered, or otherwise altered, is stored in Elasticsearch. The workload of each query is distributed over 22 computing nodes which we tuned to minimize response time. The aggregated list of dropping domains is queried, on a daily basis, against our search engine to generate the daily list of caught domains.

Web crawler: The caught domains are fed to DomainPanner's distributed web crawler which is responsible for visiting the domains and collecting their HTML code, final URL, nested iframes, redirections, and a screenshot of the final page. It also stores DNS records of the domain names, including A, NS, and SOA, and follows CNAME chains. Our distributed web crawler performs job management using Celery [3], and RabbitMQ [7] as its message broker. Crawling jobs are picked up by celery workers which visit the pages in a Chrome browser using the Selenium web driver and stay for 40 seconds before closing the window. They also collect DNS records and network information for each domain and finally send all the data to a centralized CouchDB.

3 REGISTRATION OF DROPPING DOMAINS

In this section, we study the registration of dropped domains and investigate whether the negative residual trust of a domain name (i.e. the domain was part of a blacklist at the time that it was dropped) affects its registration prospects. Additionally, we examine various features of the domains to understand which features make domains attractive and whether these features are different between regular and malicious domains. Using DomainPanner, we monitored the registration and usage of 28M dropping domains for a period of 9 months between Jan. 10th and Oct. 10th, 2017.

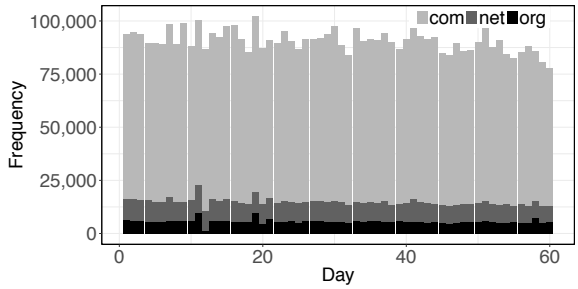

Figure 2: Number of Daily Dropped Domains

3.1 Rate of Registrations

Figure 2 shows the number of dropped domains per TLD for a period of two months starting from July 1st, 2017. Note that the vast majority of dropped domains belong to the `.com` TLD and there was no day where fewer than 75K domains were returned to the pool of available domains.

Figure 3 shows the cumulative percentage of dropped domains which were caught during a period of 30 days starting from February 1st, 2017. Specifically, DomainPanner tracks the registration status of each domain from the day that it was dropped until 30 days later. The solid line depicts the interpolations of medians of registration rates on each day for all the lists of domains. For all lists of dropping domains, we observe that the rate of registration is highest on the first day. Depending on the dropping list, the day-one rate ranges from 5% to 15%, but the median is approximately 11%. Afterwards, the registration rate decreases such that, in the remaining 29 days of that first month, the rate increases by a mere 5%.

3.2 Registration of Malicious Domains

To understand how the negative residual trust of certain domains affects their registration prospects, we examine the relationship between previous malicious activity of domain names and their registration probability. We make use of Google Safe Browsing (GSB) and VirusTotal (VT) to determine a domain's malicious history.

To capture the registration rate of malicious domains, we queried GSB for all dropping domains for a period of 30 days starting from April 15th. We then monitored the registration of all the malicious domains for a period of 5 months. Figure 4 shows the interpolated function of cumulative registration percentage of different lists over 5 months for both previously malicious as well as previously benign domain names. Even though both sets of domains exhibit the highest registration rate on their dropping date and then taper-off, we see marked differences in terms of their rates of registration. Namely, the rate of registering previously-malicious domains names is twice that of previously benign ones. As Lever et al. pointed out in their study of the residual trust of domain names [26], attackers can choose to register previously malicious domains, either because these domains can be used to reanimate malicious infrastructure (e.g., registering the C&C domain of a dormant botnet) or because attackers want to hide a more severe attack behind a less severe label (e.g., abusing a domain that was labeled as delivering PUPs and using it as a drop-server for a highly targeted attack).

To shed light on how a domain's prior malicious history affects its registration prospects, we utilized VT to obtain the latest date a domain was marked as malicious. Given VT's strict API limits, we restrict our analysis for a period of one month. We label a domain as malicious if it had ever engaged in any malicious activity. Figure 5 shows the CDF of the latest date of abuse for the 65K domains that VT labeled as malicious during our one-month observation period. We observe that as long as a domain's activity was malicious less than two years ago, its age does not affect its registration prospects. At the same time, we observe a departure from that trend for domains with malicious activity more than two years prior, which may be due to the fact that these malicious domains are too old to be useful for reanimating dormant malicious infrastructure [26] and therefore less desirable for re-registration.

3.3 How are the domains chosen?

There is an intense competition among dropcatchers to be the first to register the most valuable domains, to the extent that these companies invest millions of dollars to purchase multiple registrar licenses and increase their chances of catching a dropping domain [24]. Our results (which are in line with the recent results of Lauinger et al. [24]) show that only 10% of the daily dropping domain names are caught. In this section, we aim to understand what entices registrants to buy a given domain. Moreover, we investigate malicious domains separately, to understand why they are registered twice as often as regular domains, despite their negative reputation.

Domain Features. The desirability of a domain name depends on many factors including keywords, trends, length, language, demographics, previous traffic and indexing in search engines. At the same time, given that humans (in addition to automated bots) make these decisions, no model can perfectly predict all the desirable domains.

The people who create large portfolios of domains with the intent of selling them later for a profit are typically called domainers [20]. Domainers have their own strategies for identifying valuable domains. Some try to predict a trend and register the related domains, while others look into the characteristics of the dropping domains and use appraisal services, such as Estibot [5] and GoDaddy [6], which train machine learning models with hundreds of features on their private databases [9] to provide an estimate of a domain's worth.

To quantify the features of dropped domains that are of interest to domainers, and to understand their different domain-selection strategies, we collect a set of 12 features which are inspired by industry reports on domain selection [2, 23].

Intrinsic value: Due to their lexical compositions, domain names carry a value. For example, domain names such as business.com and sex.com have sold for millions of dollars [8]. We consider the following features that reflect the intrinsic value of a domain name: length (number of characters), number of meaningful words, having a hyphen, containing a number, including adult keywords, targeting a trademark through domain squatting [10, 19, 21, 22, 30, 35], and the number of other TLDs (.com, .net, .org, .info, .biz, and .us) where the domain is already registered.

Traffic: We capture the residual traffic of a domain and how organic it is using the following features: Alexa rank of the domain before dropping (if the domain is missing Alexa rank we give it

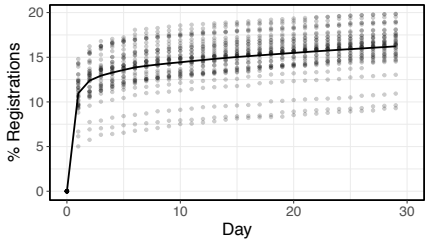

Figure 3: Percentage of re-registrations on daily bases for the domains dropped over one month and their re-registration tracked for a month

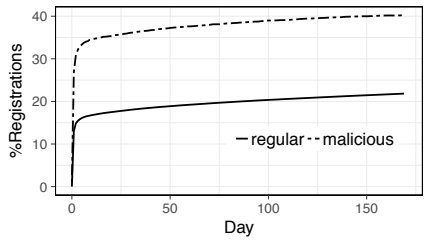

Figure 4: Percentage of re-registrations of regular and malicious domains on daily bases

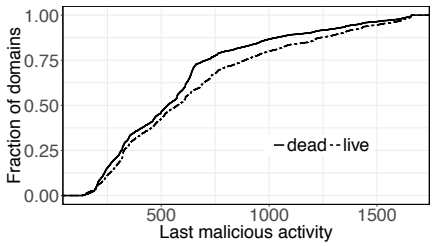

Figure 5: Last date of malicious activity before re-registration

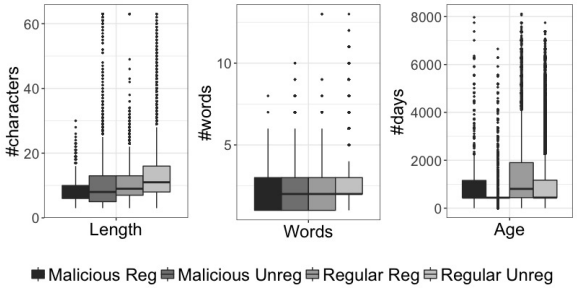

Figure 6: Distribution of domain features with normalized values

Table 1: Feature distribution of caught/uncaught domains.

Features	Caught	Uncaught
Age	49m ± 49m	32m ± 3m
#Words	2±0.9	2.5 ± 1.2
Domain length	10 ±4	12 ± 5
Search results	215K ± 3M	98K ± 2.2M
Search volume	833 ± 4K	130 ± 7K
other TLDs	0.3 ± 0.74	0.23 ± 0.62
Brand	2.7%	3.8%
Adult	0.86%	0.95%
Numbers	12.98%	16.9%
Hyphen	5.5%	10.3%
Having Wayback records	10.8%	3.2 %
In Alexa top 1M	3.3%	0.6%

the maximum value from our database), search volume for the domain keywords from the database of a commercial service, and number of search results for domain keywords in popular search engines. Search volume is a proxy for keyword popularity and can be translated to high ranking in organic search.

Registration and Usage History: We extract a domain's age from its WHOIS creation date, and determine whether it pointed to real content based on the existence of a record in Wayback Machine. These are indicators of how established and developed the domain is. Both of these features make domains more likely to have incoming links and be indexed by search engines.

Characteristics of Registered Domains. To compare the values of each feature across registered and unregistered domains, we randomly select ten daily lists of dropping domains and compile the lists of domains caught immediately (98,817 domains), and the domains that went uncaught in the first two months after dropping (873,854 domains). For each feature we perform a t-test to determine if there is a statistically significant difference between caught and uncaught domains. For all features, the p-value is much less than 0.05 (the largest p-value is 10^{-7}). The low p-value of the tests indicates that the features originate from different distributions.

Table 2: Binary features for malicious and overall population.

Features	All domains	Malicious
Brand	7.8%	9.9%
Adult	10%	13%
Numbers	8.3%	31%
Hyphen	6%	11%
Other TLDs	12%	31%

Table 1 summarizes the average values for each pair of features. Caught domains are, on average, 17 months older than uncaught domains and they are more likely to have records in the Wayback machine. This tells us that registrants favor established and developed domains. Interestingly, not all old domains were caught. Specifically, we found 2,834 uncaught domains which were more than 15 years old. For example, the oldest uncaught domain was *'wwwsexsites.com'*, with a creation date of 1997.

Moreover, caught domains tend to be shorter, contain a smaller number of words, and the probability that the same domains are already registered in other TLDs (.com, .net, .org, .info, .biz, and .us) is higher. They are less likely to have numbers or hyphens. Hyphens can improve the readability of highly keyword stuffed domains, but they may also harm branding (as they are not usually pronounced). Expectedly, the search volume, search results for domain keywords, and their residual traffic, have higher values for the caught domains.

We use the same set of features to compare malicious domains and regular domains. Our malicious list is based on the domains detected by Google Safe Browsing in 140 days, and we use the same method that we used for all domains, to compile the list of caught (5,641 domains) and uncaught (23,420 domains) malicious domains.

Table 2 highlights the key differences for the binary features. From these, we see that malicious domains are more accepting of the features that are generally considered negative. We observe that 31% of the domains containing numbers appear in the malicious caught set compared to only 8.3% in general. A reasonable explanation is that many desirable malicious domains may be generated by malware DGA algorithms. We also find that malicious caught

domains are more likely to be registered in other TLDs as well which suggests that these are used for squatting domain names present on different TLDs.

Similarly, Figure 6 uses some of the top features to compare malicious and non-malicious domains that are caught and uncaught. We find that registrants of malicious domains value shorter domain lengths, a trait which is generally favorable, but they also allow a shorter average age than regular domains, which is generally unfavorable. We argue that this behavior is likely the result of two different domain-selection strategies: i) malicious domain names that have favorable attributes can be chosen *despite* their negative past and ii) malicious domains that have less favorable attributes can be chosen because their negative residual trust is still of use to an attacker (e.g. can be used to reanimate a dormant botnet).

3.4 Clustering Registrants

In this section, we aim to obtain a better understanding of the users who utilize dropcatching services. While Lauinger et al. reported on various classes of registrars with a variety of sizes and success rates in catching dropped domains [24], our focus is on the end users who drive this market.

For this purpose, we track the changes of WHOIS records for three months starting from March 23, 2017 for all dropped domains, recording the contact information of the new registrants. In total, we collected 1,069,420 records, 6% of which do not include a registrant name. We cluster the domains based on the registrants email address and considering Levenshtein distance with an empirically chosen threshold as the similarity metric. This "fuzzy matching" allows us to group together email addresses such as `john.doe@gmail.com` and `john.doe.1@gmail.com`.

After removing clusters of WHOIS privacy protection services, we obtain a list of 31,731 clusters. We informally identify three types of clusters based on the number of purchased domains; Freelancers (individuals who bought fewer than 100 domains), Professional Domainers (small businesses with 100 to 10,000 domains), and Dropcatchers (Services that registered more than 10K domains). Even though these cut-off points are arbitrary and are only informed by our domain experience, they help in discretizing the continuum of domain registrations. The majority of the clusters (98.4%) are Freelancers, which in total registered 12% of the dropped domain names. Professional Domainers (1.5% of the clusters), registered 27% of the domains, and Dropcatchers themselves (0.03% of clusters) registered the majority (60%) of the domains. Note that the above results are based on a snapshot of the WHOIS records on the drop date.

We focus on the Freelancers and Domainers classes (we exclude the Dropcatchers class since it will, by definition, include a large number of domains which will be transferred to Freelancers/Domainers at a later date) and perform statistical tests to quantify to what extent different classes of registrants focus on different domain features. Specifically, we extract the features described in Section 3.3 for the two sets of domains caught by the registrants at the tail of each tier (top domainers registering more than 3K domains and the individuals registering a single domain). We then perform t-tests and calculate Cohens' d effect size to find the most distinctive features. As shown in Table 3, the domains selected by these two parties are significantly different in terms of age, domain length, Alexa rank, and the number of domains taken from other TLDs.

As Table 3 shows, the two populations of registrants have significantly different selection strategies. The domains registered by freelancers are, on average, longer, have a worse Alexa rank, and there is a low probability that the same domain is taken from other TLDs. The only feature which is more in-line with common wisdom [2, 23] and the strategies of Domainers is the age of a domain.

Table 4 shows the top domainers. A registrant with the email address of *80010864@qq.com* has caught more than 11K domains. Such a portfolio can only be amassed by investing at least $115k in the dropcatch market (assuming 69.6 Yuan/$10.47 per domain, the lowest price from their most used registrar).

Another top domainer (*yaomaiyumingzhaowo@126.com*) is associated with a coin-mining campaign [1]. Currently, this account has registered more than 247K domains [4]; therefore we cannot be certain whether all of these domains are acting maliciously or just some of them have been compromised.

The strategies of the registrants also vary by their demographics. Registrants belong to 145 different countries yet just the US and China account for 89% of all domain catching in this time period. Table 5 shows the breakdown of the domain registrations for the top five countries. We perform a t-test on the set of caught domains from China and the US to find out if registrants belonging to these countries choose domains in different ways. The Chinese domain names have a completely different distribution of domain length compared to the US domains. We find that, on average, Chinese domains are much shorter (7.5 ± 3 characters) than the US domains (11.5 ± 4 characters), and they are more likely to use numbers (25%), while the US-registered domains tend to avoid numbers (3%). These differences are likely rooted in the Chinese language and the fact that numbers have symbolic meaning. Chinese domains are also younger (32 ± 31 months) than the US domains (55 ± 51 months), and are less likely to be registered on other TLDs (0.1 ± 0.4 versus 0.4 ± 0.8 different TLDs).

3.5 Domain Deletion

According to the domain-name life cycle, a domain may enter the pending-delete phase either when it is not renewed, or its owner intentionally deletes its. To quantify the fraction of the domains which dropped because they organically expired, we extracted and analyzed the creation dates of dropped domains. We chose a domain's creation date instead of its expiration date since the expiration date may change during the auto-renew period and therefore cannot be used to reliably gauge the status of a domain name. Overall we extracted the creation date from the WHOIS records of 6,637,389 domains dropped in two consecutive months.

Figure 7 shows how many days before the drop date each domain was created (we limit the duration to 10 years). Most domains were created 445 days before their dropping date which means they were registered for one year (365 days), expired, and went through the auto renew phase (45 days), redemption period (30 days), and pending delete phase (5 days). The pattern of significant bursts continues for yearly intervals. The domains that do not follow this pattern are due to slightly different registrar policies and the domains that were prematurely deleted by their owners. Interestingly, most malicious domains exhibit the same patterns

Features	Freelancer	Domainers
Age	54m ± 50m	33m ± 41m
Domain length	11.6± 4.7	9.9±3.9
Alexa rank	13M ± 7M	106K ± 1M
Other TLDs	0.3±0.9	0.8±1.3

Table 3: Different characteristics of freelancers and professional domainers

Cluster Size	Email domain
11,325	80010864@qq.com
6,616	godaddy2018@qq.com
5,170	pub144@hotmail.com
4,562	dt0598@outlook.com
3,306	8648240@qq.com
3,209	yaomaiyumingzhaowo@126.com

Table 4: Email addresses of top domain name registrants

Country	# Registrations
USA	427,001
China	280,236
Japan	34,378
Hong Kong	15,984
Singapore	4,518

Table 5: Registrations by country

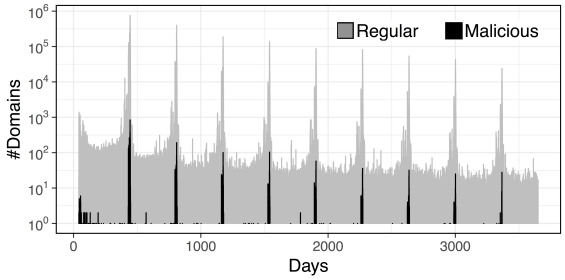

Figure 7: Age of domains before dropping showing bursts at typical yearly expiration intervals

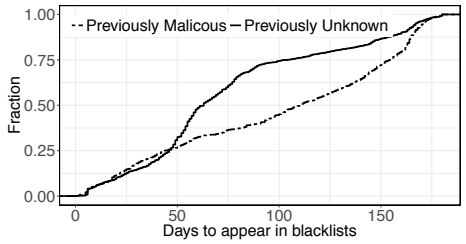

Figure 9: Number of days it takes for a malicious domain to appear in blacklists.

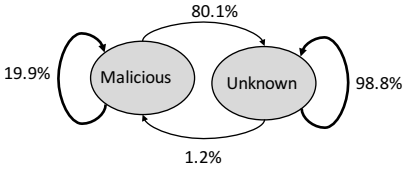

Figure 8: Transitions between unknown and malicious for the caught domains

which suggests that they are allowed to remain registered even after they have been detected as abusive.

4 POST-REGISTRATION USAGE

In this section, we describe how domains are used post-registration. We study the characteristics of domains that host malicious content prior to dropping, after they are caught, or both. Then, we explore the usage of non-malicious domains to gain insights into the intentions of their buyers.

4.1 Domains Tainted by Malicious Activity

We begin our analysis by considering domains that were, at some point in time, known to host malicious content. For this purpose, we closely tracked the registration and status of 1,802,813 domains dropped in a 10-day period, which we found to be a sufficiently representative sample. After each domain was dropped, we tracked its registration for the following 10 days. Of the dropped domains, we found that 145,087 (8%) were caught. We queried VirusTotal for each of the caught domains six months after their registration date to check if they became malicious. We also use the responses of VirusTotal to check the history of the domains in terms of having previously served malicious content. If a domain was ever reported as malicious, we consider it to be a malicious domain. Otherwise, we label it as "unknown," because we are unsure of its status and will investigate its activity in Section 4.3. Figure 8 presents the state transition diagram for these domains.

We find that 3,893 of the caught domains had a history of malicious activity and 19.9% of them continued serving malicious content even after deletion and re-registration. Additionally, 1.2% of the unknown domains that were not previously present in blacklists, also started to serve malicious content. These results highlight the fact that domains that were malicious in the past are much more likely to be caught for malicious purposes compared to non-malicious domains being caught to serve malicious content.

We also investigate the time frame in which caught domains begin serving malicious content. Figure 9 plots the number of days since registration that it takes for a domain to appear in a blacklist. We find that more than 60% of the domains appear in blacklists less than 80 days after being caught. Considering the delay between serving malicious content, being detected as malicious, and being listed in a blacklist, these results indicate that more than half of the domains started their malicious activity soon (less than two months) after registration. Notably, domains that were not known to be previously malicious enter blacklists quicker than if they were already marked as malicious prior to being dropped.

4.2 Subversion of Non-malicious Domains

We conducted an in-depth investigation of the domains that became malicious after being caught to gain insights about the responsible parties. We leveraged Google Safe Browsing (GSB) to track domains over the course of 80 days, checking them for malicious activity (as indicated by GSB) from the day prior to their re-registration. We only consider domains that were not present in GSB on the

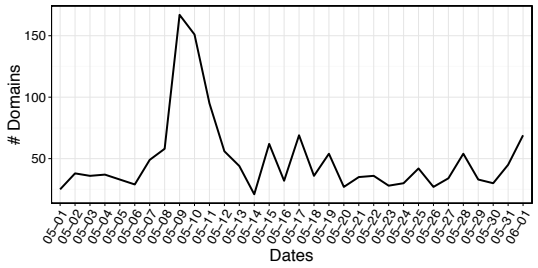

Figure 10: Number of domains used for malicious purposes

day before their registration, but entered the list at some point during the 4 months after their re-registration. This purposefully constrains our analysis to subverted domains, i.e., those that used to be benign but were caught for malicious purposes.

Figure 10 shows the daily number of registrations of these domains over the course of the first month of our experiment. In total, we observe 6,838 domains that became malicious after being caught. Of these, 6,449 (94%) serve malware, 351 (5%) are used for social engineering, 34 host potentially unwanted programs (PUPs), and 4 domains launch multiple attacks.

We use the WHOIS records to cluster these domains based on the registrant email addresses following the methodology described in Section 3.4. In total, we find 901 registrants, of which 76 caught more than 10 domains, with the biggest cluster, *dt0598@outlook.com* (*OUTLOOK*), registering 385 domains during the 80-day period. The domains that *OUTLOOK* turned malicious were caught throughout the course of the 80-day period, ranging from one or two, to as many as 117 registrations per day.

Overall, we find that many malicious registrants conduct their registration campaigns in bulk, registering many domains in a single day. For example, the spike on May 9th in Figure 10 is the result of bulk registrations mainly by *ok358w@qq.com* (*QQ*) who registered 88 domains and *aosornntyr@sina.com* (*SINA*) who registered 22 domains. Domains registered by *QQ* were in the format of <6 digits>.net, but those registered by *SINA* contained meaningful words and no digits. This behavior suggests that many malicious campaigns are conducted sporadically and manually, with malicious actors curating the domain list by hand on arbitrary days, rather than algorithmically seeking out the most advantageous dropped domains on any given day.

Studying the origin of the registrants reveals another interesting factor. Subverted domains are caught by actors from 64 different countries, but 5,048 (80%) of these domains are registered from China. The next highest country of registrants is the United States, with only 6% of the registrations that become malicious. These statistics are in stark contrast to the general trends of caught domains, where registrants from China account for only 35% of caught domains and registrants from the US are responsible for more than half (53%).

Subverting domains can serve as a good indicator of maliciousness on behalf of the registrant. Using the list of subverted domains, we consider all registrant clusters responsible for catching these domains as malicious. We then conservatively remove accounts of the dropcatchers, domain aftermarkets, and privacy protection services to avoid false positives.

Table 6: Contents of registered dropped domains crawled one month after drop date

Category	Frequency
Malicious domains	0.2%
Affiliate abuse	0.3%
Parked/Ads	69.2%
Error pages	18.1%
Ecosystem Total	89.6%
Real web content	<10.4%

As a result, we are left with 812 registrants who are likely to be malicious. Of the 1,059,050 domains caught during the 80-day period, these 812 registrants are responsible for 105,112 (10%), giving us a lower bound on the percentage of the dropcatching activity that is maliciously motivated.

4.3 Contents of Re-registered Domains

A major goal of our study is to understand the participants in the dropcatch ecosystem and the market forces behind them. To this end, we collected a 25-day dataset consisting of all domains that were dropped and caught. We then used a distributed crawler to explore these domains, and undertook a labeling effort to categorize all 375,537 of them. Notably, caught domains go through a series of temporary states before they are transferred to the final registrant. For example, following a backorder at dropcatch.com, the user is given four days to pay the fee. During this time, the domain registration indicates *"This domain was caught by DropCatch.com"* and, if the registrant does not pay, the domain is listed for sale at *hugedomain.com*. Because of this, we perform our crawl with a one month delay after a domain is caught, to ensure that sufficient time has elapsed for the new owner to take control of the domain and put it into service.

Content-Based Classification. We took a multi-stage approach to labeling the caught websites. We briefly summarize the labeling process here, and include a complete in-depth explanation of all labeling steps and interesting observations made along the way in the following paragraphs.

(1) Eliminate domains unreachable via HTTP
(2) Label as "malicious" if they are blacklisted
(3) Label as "malicious" if they include malicious content
(4) Identify "affiliate abuse"
(5) Identify "Parking/Ads" based on DNS records
(6) Identify "Parking/Ads" based on redirects
(7) Cluster and label visually or structurally similar pages
(8) Label a random sample of remaining domains

The results of our clustering effort are presented in Table 6. We observe that, although a notable portion of dropped domains are caught for malicious use, they currently form a small fraction of overall caught domains. The majority of the domains serve advertisements for online casinos or serve parking pages, which frequently expose visitors to social engineering, adult content, scams, or malware [37]. Less than 10.4% of the domains were used to provide real web content. In summary, an overwhelming majority of the thriving multi-million dollar [24] ecosystem revolves around capitalizing on the residual traffic and trust of dropped domains, predominantly through means that are considered detrimental from the perspective of web users and security experts.

Table 7: Malware category served by Injected URLs

Malware	Frequency
Trojan.HTML.Ramnit.A	72%
W32.Malware.Gen	14%
JS.eIframeHlNMe.F841	2%
Win32.Trojan.Raasmd.Auto	1.5%

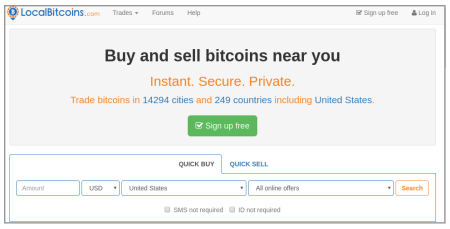

Figure 11: Screenshot of the deleted domain *loacalbitcoins[.]com* which used to perform phishing against (*localbitcoins[.]com*), a service for trading local currency and bitcoins.

Content Clustering Methodology. We now provide the details of our clustering methodology, separating the process into the eight steps summarized in the previous section.

(1) One month after re-registration, 12% of the 375,537 domains that we studied did not resolve to an IP address and 5.8% did not listen on port 80.

(2&3) We first identified 2,594 malicious domains by checking the domains against the Google Safe Browsing (GSB) service. We then extracted the JavaScript and iframes included in all the landing pages, yielding a set of 588,104 unique URLs of which 5,474 were found to be malicious. These malicious URLs were included on 3,311 crawled domains, so in total we labeled 5,905 domains as malicious which were either detected by GSB directly or included a malicious iframe or JavaScript script. We further investigated the type of malware served by malicious URLs by downloading the most recent associated samples from VirusTotal. For 17% of the URLs, we downloaded at least one example and used a majority vote between labels reported by the available AV engines. Table 7 presents the most frequent malware among the URLs. In total, 30 different malware labels were detected.The most popular malware was *Trojan.HTML.Ramnit.A*, which steals cookies and login credentials, hijacks sessions, and performs man-in-the-browser attacks.

Overall, the malicious domains were used for a range of unwanted activities, including dropping malware, social engineering attacks, unwanted software (PUPs) (example shown in Figure 12), and phishing attacks against financial services. Figure 11 shows an example of phishing attacks utilizing the deleted domain *loacalbitcoins[.]com*. This domain is a typosquatting version of (*localbitcoins[.]com*), which is a legitimate service for trading local currency for bitcoins.

(4) We identified affiliate abuse [12] by following redirection chains to find the landing page of the domains. If the final domain was among the Alexa top 10K and a tracking parameter was passed to it, we marked the domain as performing affiliate abuse. To avoid false positives, we manually checked the redirections and excluded non-affiliate services. For example, we excluded redirections to

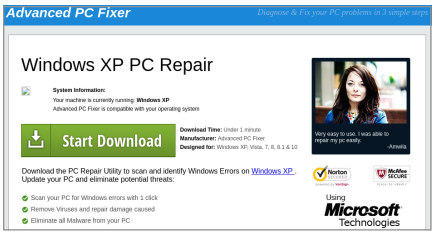

Figure 12: A re-registered deleted domain (*granny-daily[.]com*) asks user to download malware.

domain aftermarket services such as hugedomains.com (80,995 redirections) and uniregistry.com (1,307 redirections). Similarly, we excluded domains that redirected to popular hosting providers and content delivery networks such as hostgator.com and rackcdn.com, and popular parking services such as thewhizmarketing.com. In total, we found only 692 domains participating in affiliate abuse. The top targets of affiliate abuse are shopping websites such as Amazon, Edmunds, Ebay, and HomeDepot. We also found a number of domains that redirect to social media sites such as Facebook (where the traffic redirects to specific pages by passing a campaign ID), and search engines such as Yahoo and Google, where the user is taken directly to the search results page for specific keywords.

(5 & 6) We used the domains' nameservers and techniques by Vissers et al. [37] which we extended with more parking-operated nameservers to improve detection. To account for smaller players, we also manually examined redirections that were not affiliate-abuse related and discovered parking-related redirections, such as those to hugedomains.com, the aforementioned domain aftermarket.

(7) To label the rest of domains, we applied automated clustering and manually labeled each cluster based on a sample of five screenshots of that cluster. Pages were clustered based on their visual similarity using perceptual hash and structural similarity using simhash. To facilitate the labeling process, we implemented a web-based cluster labeling application that presents the cluster screenshots and allows the user to label the cluster as parking, error page, or real web content.

We first clustered visually similar pages by calculating the perceptual hash of the screenshots and considering an empirically selected threshold. To select a threshold that results in few false positives, we examined different thresholds and manually verified the results by looking into samples of the top 500 clusters. The domains grouped into 52,162 different clusters. By examining the clusters that had more than 20 domains, we were able to label 35,284 domains.

Next, we used simhash on the DOM tree of the pages to capture their structural similarities and label together the pages that are syntactically similar. For example, parked pages that all include an iframe, or websites developed using the same web design templates may be visually different and have different perceptual hash values, but contain similar structural elements. After tuning the algorithm thresholds, we identified 18,302 different clusters. Again we examined the clusters that had more than 20 domains, which enabled us to label an additional 36,895 domains.

(8) The above methods resulted in labeling 55% of the domain names. In order to characterize the remainder, we chose a sample of 10% from the unlabeled domains and used their perceptual hash to label them.

5 RELATED WORK

Prior work by Halvorson et al. has investigated malicious intent behind domain registrations [14–16]. Many others have sought to discover/predict malicious activity based on domain names and DNS data [11, 13, 17, 18, 27].

Several works have explored the potential to abuse residual trust in domain names when their owners allow them to expire. Moore and Clayton investigated expired domain names from failed financial institutions [29]. They found cases where these domains were re-registered to abuse their residual trust for SEO and malware spreading. Two years later, Lever et al. further explored residual trust abuse for six years of domains [26]. They developed an algorithmic approach to detect domain ownership changes and found that 8.7% of domains in public blacklists are re-registered after expiration. Compared to this work, our work focuses on the dropcatching ecosystem and examines registrants motivations and uses of domains beyond abuse of residual trust. Moreover, we characterize the use of caught domains by crawling 375,537 pages, finding that less than 11% serve real web content. In 2012, Nikiforakis et al. showed that remote JavaScript inclusions pointing to expired domains can be re-registered and used for code injection attacks [31]. Vissers et al. showed that expired domains can lead to hijacking of large numbers of domains through their nameservers [36]. Schlamp et al. identified hundreds of expired domains in databases of Regional Internet Registrars which could be abused to hijack entire networks and Autonomous Systems [32, 33]. All of these cases begin with an expiring domain. Our paper sheds light on this stage by exploring the patterns and motivations behind dropcatch registrations.

In a concurrent work, Lauinger et al. studied the processes of domain expirations and re-registrations [24, 25]. The authors explored how long it takes for domain names to be re-registered in the largest TLDs and illuminate the competitive process between registrars to re-register desirable domain names. Compared to this work, our paper investigates features of caught domain names in greater depth, particularly with respect to malicious history and use after registration.

6 DISCUSSION & CONCLUDING REMARKS

In this paper we presented evidence of the high levels of activity in the dropnd that, on a daily basis, more than a hundred thousand domains excatching ecosystem, an online ecosystem that few have heard of. We foupire and as they expire, dedicated registrars called dropcatchers rush to be the first to register the most valuable 10% of these domains. By extracting tens of features from each domain name, we noticed that even though there exist features that are, in general, desirable (such as the length and age of a domain name) not everyone requires these features to be present for each domain that they register. Specifically, we were surprised to find that previously malicious domains are twice as likely to be caught as benign domains. We presented evidence showing the existence of professional registrants from China and the US amassing portfolios with thousands of previously dropped domain names and identified the parties behind the domains that turn malicious after re-registration. Finally, we performed a large-scale crawl of 375,537 dropped domains finding that the majority of domains become parked and, next to serving malware, phishing pages, scareware, and PUPs, less than 11% of the caught domains are put to use for showing web content.

Overall, our findings demonstrate that dropcatching creates an unfortunate environment that results in pages filled with ads (domain parking), allows attackers to abuse an expired domain's residual trust (e.g. their incoming links), and exposes users to a wide range of malicious content. We recommend that the curators of popular blacklists take into account the phenomenon of dropcatching and be extra vigilant about domains that are re-registered. From the side of dropcatching services, we recommend that they integrate blacklists into their service, such as Google Safe Browsing, and scrutinize the registrants that exhibit an interest in re-registering previously malicious domains.

Acknowledgments: We thank the reviewers for their valuable feedback. Moreover, we thank Ivan Rasskazov and Luc Lezon from Intelium for giving us access to their domain analytics platforms. This research was supported by the Office of Naval Research (ONR) under grant N00014-16-1-2264 as well as the National Science Foundation under grants CNS-1617902 and CNS-1617593. Some of our experiments were conducted with equipment purchased through NSF CISE Research Infrastructure Grant No. 1405641.

REFERENCES

[1] 2016. Resurrection of the Evil Miner. https://www.fireeye.com/blog/threat-research/2016/06/resurrection-of-the-evil-miner.html. (2016).

[2] 2017. Assessing the Value of a Domain. In *MarkMonitor White Paper*.

[3] 2017. Celery: Distributed Task Queue. http://www.celeryproject.org. (2017).

[4] 2017. DoaminIQ (a Domain Intelligence Service). https://www.domainiq.com/. (2017).

[5] 2017. Estibot appriasal tool. http://www.estibot.com. (2017).

[6] 2017. GoDaddy appraisal tool. http://www.godaddy.com/domain-value-appraisal. (2017).

[7] 2017. RabbitMQ: The most widely deployed open source message broker. https://www.rabbitmq.com. (2017).

[8] 2017. Sex.com Domain Sale Entered into Guinness Book of World Records. http://www.domainnamenews.com/news/sexcom-domain-sale-entered-guinness-book-world-records/8822. (2017).

[9] 2017. Using Deep Learning for Domain Name Evaluation. http://engineering.godaddy.com/using-deep-learning-domain-name-valuation/. (2017).

[10] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. 2015. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *Proceedings of the ISOC Network and Distributed System Security Symposium (NDSS 15)*.

[11] Davide Canali, Marco Cova, Giovanni Vigna, and Christopher Kruegel. 2011. Prophiler: a fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th international conference on World wide web*. ACM, 197–206.

[12] Neha Chachra, Stefan Savage, and Geoffrey M. Voelker. 2015. Affiliate Crookies: Characterizing Affiliate Marketing Abuse. In *Proceedings of the 2015 Internet Measurement Conference (IMC '15)*. 41–47.

[13] Mark Felegyhazi, Christian Kreibich, and Vern Paxson. 2010. On the Potential of Proactive Domain Blacklisting. *LEET* 10 (2010), 6–6.

[14] Tristan Halvorson, Matthew F Der, Ian Foster, Stefan Savage, Lawrence K Saul, and Geoffrey M Voelker. 2015. From. academy to. zone: An analysis of the new TLD land rush. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*. ACM, 381–394.

[15] Tristan Halvorson, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. 2014. Xxxtortion?: inferring registration intent in the. xxx tld. In *Proceedings of the 23rd international conference on World wide web*. ACM, 901–912.

[16] Tristan Halvorson, Janos Szurdi, Gregor Maier, Mark Felegyhazi, Christian Kreibich, Nicholas Weaver, Kirill Levchenko, and Vern Paxson. 2012. The BIZ top-level domain: ten years later. In *International Conference on Passive and Active Network Measurement*. Springer, 221–230.

[17] Shuang Hao, Alex Kantchelian, Brad Miller, Vern Paxson, and Nick Feamster. 2016. PREDATOR: proactive recognition and elimination of domain abuse at time-of-registration. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1568–1579.

[18] Shuang Hao, Matthew Thomas, Vern Paxson, Nick Feamster, Christian Kreibich, Chris Grier, and Scott Hollenbeck. 2013. Understanding the domain registration behavior of spammers. In *Proceedings of the 2013 conference on Internet measurement conference*. ACM, 63–76.

[19] Tobias Holgers, David E Watson, and Steven D Gribble. 2006. Cutting through the Confusion: A Measurement Study of Homograph Attacks.. In *USENIX Annual Technical Conference, General Track*. 261–266.

[20] David Kesmodel. 2008. *The Domain Game: How People Get Rich from Internet Domain Names*. Xlibris Corporation. http://books.google.be/books?id=PvAzPAAACAAJ

[21] Mohammad Taha Khan, Xiang Huo, Zhou Li, and Chris Kanich. 2015. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting. In *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, 135–150.

[22] Panagiotis Kintis, Najmeh Miramirkhani, Charles Lever, Yizheng Chen, Rosa Romero-Gomez, Nikolaos Pitropakis, Nick Nikiforakis, and Manos Antonakakis. 2017. Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse. In *Proceedings of 24th ACM Conference on Computer and Communications Security (CCS)*.

[23] Kelsey LaBelle, Kyle Wilhoit, Mark Kendrick, Steve Butt, Tim Chen, and Tim Helming. 2017. Domain Valuation: How To Value a Domain Name. http://blog.domaintools.com/2011/01/domain-valuation-how-to-value-a-domain-name/. (2017).

[24] Tobias Lauinger, Abdelberi Chaabane, Ahmet Buyukkayhan, Kaan Onarlioglu, and William Robertson. 2017. Game of Registrars: An Empirical Analysis of Post-Expiration Domain Name Takeovers. In *Proceedings of the USENIX Security Symposium*.

[25] Tobias Lauinger, Kaan Onarlioglu, Abdelberi Chaabane, William Robertson, and Engin Kirda. 2016. WHOIS Lost in Translation:(Mis) Understanding Domain Name Expiration and Re-Registration. In *Proceedings of the 2016 ACM on Internet Measurement Conference*. ACM, 247–253.

[26] Chaz Lever, Robert Walls, Yacin Nadji, David Dagon, Patrick McDaniel, and Manos Antonakakis. 2016. Domain-Z: 28 registrations later measuring the exploitation of residual trust in domains. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 691–706.

[27] Justin Ma, Lawrence K Saul, Stefan Savage, and Geoffrey M Voelker. 2009. Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1245–1254.

[28] Chris Matyszczyk. 2010. Cowboys fire coach, forget to renew Web domain. https://www.cnet.com/news/cowboys-fire-coach-forget-to-renew-web-domain/. (2010).

[29] Tyler Moore and Richard Clayton. 2014. The ghosts of banking past: Empirical analysis of closed bank websites. In *International Conference on Financial Cryptography and Data Security*. Springer, 33–48.

[30] Tyler Moore and Benjamin Edelman. 2010. Measuring the perpetrators and funders of typosquatting. In *International Conference on Financial Cryptography and Data Security*. Springer, 175–191.

[31] Nick Nikiforakis, Luca Invernizzi, Alexandros Kapravelos, Steven Van Acker, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. 2012. You are what you include: large-scale evaluation of remote javascript inclusions. In *Proceedings of the 2012 ACM conference on Computer and communications security*. ACM, 736–747.

[32] Johann Schlamp, Georg Carle, and Ernst W Biersack. 2013. A forensic case study on as hijacking: The attacker's perspective. *ACM SIGCOMM Computer Communication Review* 43, 2 (2013), 5–12.

[33] Johann Schlamp, Josef Gustafsson, Matthias Wählisch, Thomas C Schmidt, and Georg Carle. 2015. The abandoned side of the Internet: Hijacking Internet resources when domain names expire. In *International Workshop on Traffic Monitoring and Analysis*. Springer, 188–201.

[34] MG Siegler. 2010. Foursquare Goes Dark Too. Unintentionally. https://techcrunch.com/2010/03/27/foursquare-offline/. (2010).

[35] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. 2014. The Long" Taile" of Typosquatting Domain Names.. In *USENIX Security Symposium*. 191–206.

[36] Thomas Vissers, Timothy Barron, Tom Van Goethem, Wouter Joosen, and Nick Nikiforakis. 2017. The Wolf of Name Street: Hijacking Domains Through Their Nameservers. In *Proceedings of 24th ACM Conference on Computer and Communications Security (CCS)*.

[37] Thomas Vissers, Wouter Joosen, and Nick Nikiforakis. 2015. Parking Sensors: Analyzing and Detecting Parked Domains. In *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS)*.