Exploring the ecosystem of malicious domain registrations in the .eu TLD

Thomas Vissers[1], Jan Spooren[1], Pieter Agten[1], Dirk Jumpertz[2], Peter Janssen[2], Marc Van Wesemael[2], Frank Piessens[1], Wouter Joosen[1], and Lieven Desmet[1]

[1] imec-DistriNet, KU Leuven, Belgium
{firstname.lastname}@cs.kuleuven.be,
[2] EURid VZW, Belgium
{firstname.lastname}@eurid.eu

Abstract. This study extensively scrutinizes 14 months of registration data to identify large-scale malicious campaigns present in the .eu TLD. We explore the ecosystem and modus operandi of elaborate cybercriminal entities that recurrently register large amounts of domains for one-shot, malicious use. Although these malicious domains are short-lived, by incorporating registrant information, we establish that at least 80.04% of them can be framed in to 20 larger campaigns with varying duration and intensity. We further report on insights in the operational aspects of this business and observe, amongst other findings, that their processes are only partially automated. Finally, we apply a post-factum clustering process to validate the campaign identification process and to automate the ecosystem analysis of malicious registrations in a TLD zone.

Keywords: malicious domain names, campaigns, DNS security

1 Introduction

The Domain Name System (DNS) is one of the key technologies that has allowed the web to expand to its current dimensions. Virtually all communication on the web requires the resolution of domain names to IP addresses. Malicious activities are no exception, and attackers constantly depend upon functioning domain names to execute their abusive operations. For instance, phishing attacks, distributing spam emails, botnet command and control (C&C) connections and malware distribution: these activities all require domain names to operate.

Widely-used domain blacklists are curated and used to stop malicious domain names[3] shortly after abusive activities have been observed and reported. As a consequence, attackers changed to a hit-and-run strategy, in which malicious domain names are operational for only a very small time window after the initial registration, just for a single day in 60% of the cases [11]. Once domain names

[3] We use the term *malicious domain name* whenever we refer to a domain name that is registered to be bound to a malicious service or activity.

have fulfilled their purpose, attackers can simply abandon them and register a new set of domain names to ensure continuity of their criminal activities [24].

This strategy is economically viable to the attackers when the cost of registering a domain name is minimal. However, this approach requires repetitive and often automated domain name registrations. We refer to these series of malicious domain names registered by a single entity as *campaigns*. To obscure their actions, attackers often use fake registration details and need to switch between identities, registrars and resellers to avoid detection.

Moreover, we have observed that certain underground services pop up to facilitate the bulk domain registration process for abusive activities. For instance, on the darknet forum "AlphaBay", we found several instances of "Domain and Email Registration as a Service". In one example[4], cyber criminals register new domain names and create fresh, private email accounts that are sold to be used for illegal activities, such as carding.

The sheer volume of malicious domain names, as well as the fact that the registration process is being automated and monetized, illustrates the need for strong insights into the modus operandi of cybercriminals to produce effective countermeasures.

In this paper, we focus on the malicious campaign ecosystem by extensively leveraging the registrant and registration details, with the goal to better understand how miscreants operate to acquire a constant stream of domain names. We rigorously investigate 14 months of .eu domain registrations, a top 10 ccTLD [15] for the European Economic Area. Overall, the dataset of this study contains 824,121 new domain registrations; 2.53% of which have been flagged as malicious by blacklisting services.

Among others, the following conclusions can be drawn from this in-depth assessment:

1. While most malicious domains are short-lived, a large fraction of them can be attributed to a small set of malicious actors: 80.04% of the malicious registrations are part of just 20 long-running campaigns. We identified campaigns that were active for over a year, and campaigns that registered more than 2,000 blacklisted domains. (Section 3)
2. The campaign identification process suggests that 18.23% of malicious domains does not end up on a blacklist. (Section 3.3)
3. The malicious domain registration process is only partially automated: underground syndicates work along office hours, take holiday breaks and make human errors while registering domains. (Section 4)
4. Ecosystem analysis can be automated and reproduced by leveraging clustering algorithms. In our experiment, the 30 largest clusters formed by agglomerative clustering encompass 91.48% of blacklisted campaign registrations. These clusters exhibit a clear mapping with manually identified campaigns. (Section 5)

[4] `http://pwoah7foa6au2pul.onion/forum/index.php?threads/%E2%96%84-%E2%96%88-%E2%98%85-paperghost-%E2%98%85-%E2%96%88-%E2%96%84-fresh-non-hacked-private-email-logins-lower-your-fraud-detection-score-2.71566`

The remainder of this paper is structured as follows. First, in Section 2, we introduce the data set used in this research, along with initial insights. Next, we perform a large scale experiment to manually identify malicious campaigns (Section 3), followed by several analyses to gather more insights (Section 4). In Section 5, we follow up with a method to automate campaign identification. We discuss applications and limitations in Section 6, followed by a summary of related work in Section 7. Lastly, we conclude this study in Section 8.

2 Datasets and initial findings

In this section, we present the data used in this research and give initial insights based on a first, high-level analysis.

2.1 Registration data

We analyzed 824,121 .eu domain registrations between April 1, 2015 and May 31, 2016. We inspected the following fields:

Basic registration information contains the domain name, the date and time of registration, and the registrar via which the registration happened.

Contact information of the registrant contains the company name, name, the language, email address, phone, fax, as well as postal address information. We decomposed two additional attributes from the email address: the email account and the email provider.

Nameservers or glue records that are responsible for resolving entries within the domain. We enriched the nameserver data with their geographical location by resolving the NS records and adding IP geolocation data.

2.2 Blacklists

To capture whether or not a domain was used in malicious activity, a set of public blacklists was queried on a daily basis. Each new domain is monitored daily during 1 month after registration. Afterwards, all domains were checked once more 4 months after the last registration in our dataset. The following blacklist services have been used:

dbl.spamhaus.org blacklist [21]. This Spamhaus blacklist is queried using their DNS API, and provides indicators for botnet C&C domains, malware domains, phishing domains, and spam domains.

multi.surbl.org blacklist [20]. SURBL features a combination of different lists, such as abuse, phishing, malware, etc. The combined SURBL list is queried over DNS.

Google's Safe Browsing list [7]. Google's Safe Browsing list is queried via a Web API, and provides indicators for malware domains, phishing domains, and domains hosting unwanted software, as described in [8].

2.3 Preliminary insights

Given the data described above, we present a preliminary analysis to provide insights in the general trends and patterns of malicious registrations.

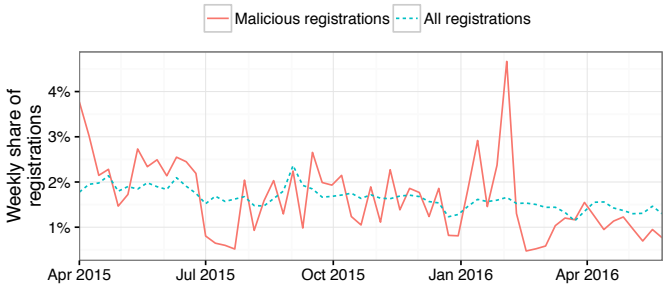

Fig. 1: Weekly share of malicious and all registrations over time.

Observing the 824,121 registrations that occurred between April 1, 2015 and May 31, 2016, we find that 2.53% end up on a blacklist. This corresponds to a total of 20,870 registrations used by cyber criminals in the given 14 month time span. Figure 1 shows the weekly share of both malicious and all registrations over this period. The differences in intensity of malicious registrations are moderately correlated with those of all registrations ($\rho = 0.54$). However, the variance of malicious registrations is clearly much larger. Most of the increased malicious activity, for instance at the start of February 2016, can be attributed to a single malicious campaign. These cases are discussed in depth in Section 3.

The selected blacklists return metadata that encode the reason(s) why a particular domain name was flagged. In our records, 93.68% of the blacklisted domains in the dataset is labelled for spam, 2.09% for malware infrastructure, 0.57% for unwanted software, and 3.22% for phishing activities.

Most domains appear on blacklists very shortly after their registration. More specifically, 72.93% of malicious domains were flagged within 5 days of delegation. 98.57% of malicious registrations are listed on a blacklist in their first month.

3 Campaign identification experiment

Typically, illegal online activities do not occur in an isolated or dispersed fashion [5,11]. Instead, malicious actors commonly set up campaigns that involve multiple, tightly related abusive strategies, techniques and targets. Through an in-depth, a posteriori analysis of the .eu dataset, we assessed whether such patterns can be identified between domain registrations and to what extent these registrations happen in bulk.

Ultimately, we manually identified 20 distinct campaigns responsible for the vast majority of malicious registrations. A campaign represents a series of registrations over time, with strong similarities in terms of registration data (e.g. the registrar, the registrant's address information, phone number or email address, and the set of nameservers). Moreover, a campaign can most probably be attributed to a single individual or organization. In this section, we first give a more thorough description on how these campaigns were identified, followed by some general insights into their characteristics.

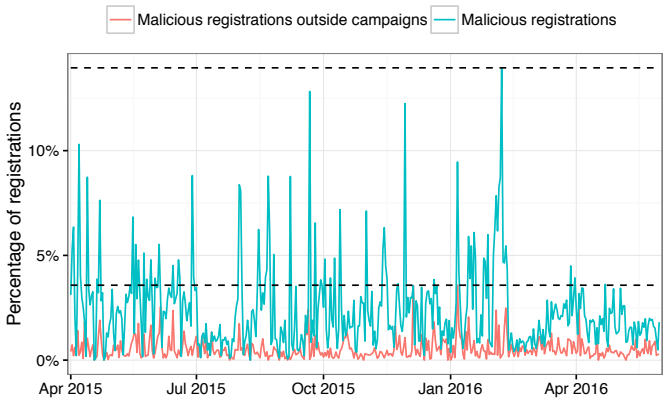

Fig. 2: Daily percentage of malicious registrations, including and excluding campaign registrations. The dotted lines represent the highest daily concentration of both sets.

3.1 Campaign identification process

As malicious registrations often occur in batches [11,10], high temporal concentrations can serve as a preliminary indicator of campaign activity. Figure 2 plots the relative amount of malicious registrations on each day. That graph can be used to identify the time periods in which the amount of malicious registrations was surging. If a campaign was responsible for a high concentration of malicious registrations, a substantial subset of registrations within that timeframe should be related to each other. Hence, all malicious registrations that occurred in that time span are examined to find common characteristics in the registration data. These can be recurring values or distinct patterns in the email address, the address info, the registrar, the registrant name, etc. To detect useful outliers, we visualized correlations between registration fields. For example, by plotting the email providers of the registrants versus the country listed in their street address (as shown in Figure 3), multiple hotspots of malicious registrations can be found that contribute to one or more campaigns. These unique combinations

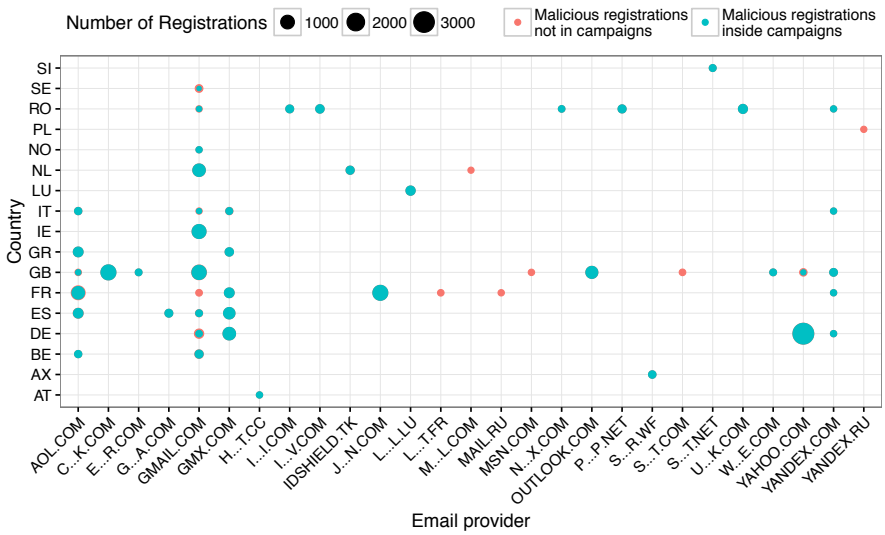

Fig. 3: Malicious registrations, grouped by email provider and country of the registrant. For visibility, combinations with less that 50 registrations are left out of the figure. Moreover, email providers with less than 50 distinct email addresses in the dataset have been obfuscated for privacy reasons.

and patterns form the basis of the manually assigned campaign selection criteria. To evaluate these, we apply them to the full dataset, i.e. on both benign and blacklisted registrations, over all 14 months. If the criteria match multiple active days and contain a substantial number of blacklisted domains, they are withheld as a new campaign. This process was repeated iteratively, reducing the number of malicious concentrations each time.

Over the complete dataset, we identify 20 distinct campaigns. A variety of attributes of the registration details have been used to characterize a campaign, the specifics for each campaign are listed in Table 1.

3.2 General campaign observations

The activity of the 20 identified campaigns is depicted in Figure 4. A first observation is that most of the campaigns are long-living: only one campaign runs for less than a month, while some campaigns run up to a year and more[5].

Secondly, campaigns strongly vary in their activity patterns. Some campaigns are active on almost a daily basis (e.g. campaign c_19), whereas others only have a few distinct active days throughout their lifetime (e.g. campaign c_07). Simi-

[5] Note that some campaigns might be running even longer than 372 days, as they might have been active before the starting date of our dataset (campaigns c_01 - c_05) or they may still be active past the time span that is covered in our dataset.

Table 1: Attributes used to express the selection criteria of a campaign. ● represents a string match, and ☆ a regular expression pattern.

	Campaign																			
Criteria	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
domain name	–	–	–	–	☆	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
registrar	–	–	–	●	–	–	–	–	●	–	–	●	–	–	●	–	–	–	–	●
nameservers	–	–	–	☆	–	–	–	●	–	–	–	–	–	–	☆	–	–	–	–	●
name	☆	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
address	–	●	●	☆	–	●	–	–	–	–	–	–	●	●	☆	●	–	–	–	–
organization	☆	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
email account	–	–	☆	☆	–	–	●	–	–	–	–	☆	–	–	–	–	–	–	●	–
email provider	●	–	●	●	●	–	●	–	●	●	●	–	–	–	☆	●	–	●	●	●

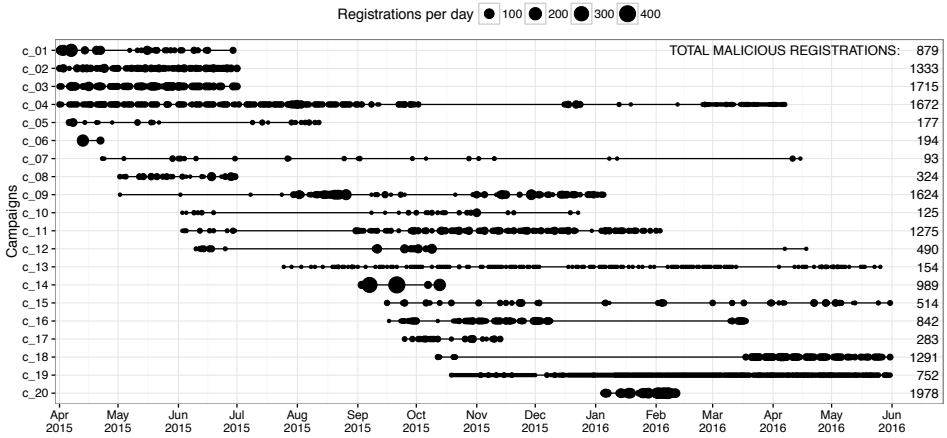

Fig. 4: Campaign duration and activity over time. The black lines represent the overall duration of the campaign, while the black dots indicate the number of malicious registrations on that day.

larly, campaigns vary in concentration. An intense, six week campaign was for instance responsible for almost 2,000 new registrations (c_20), whereas one steady campaign ran over 10 months and produced only 154 malicious registrations (c_13).

A third observation is that campaigns contribute to a large fraction of malicious registrations found in the .eu registration data. Together, the 20 campaigns cover 16,704 domain registrations, that appeared on blacklists. This represents 80.04% of the 20,780 blacklisted registrations in our dataset.

Lastly, not all registrations identified as part of a campaign are flagged as malicious. In total, 19.30% of the campaign registrations we identified are not known as abusive domains by blacklisting services. A more in-depth analysis of these potential false positives is discussed in Section 3.3. Note that to avoid any bias, Figure 4 only include registrations that appeared on blacklists, and thus represent a lower bound of campaign activity.

3.3 Validation of campaign selection criteria

As briefly mentioned in the previous section, 19.30% of the registrations associated with malicious campaigns do not appear on blacklists. We expect that various reasons contribute to this mismatch:

1. **Incomplete coverage by blacklists.** As blacklists are not exhaustive oracles, we expect that certain domains in a campaign may simply not have been picked up by the specific set of blacklists used.
2. **Not abused.** It is possible that a number of campaign registrations simply has not been used for malicious purposes (yet).
3. **False positives.** Some of our campaign criteria might not be strict enough, introducing false positive matches.

Figure 5 depicts in red the percentage of registrations for each individual campaign that appears on a blacklist. There are three campaigns with less than 60% of their registrations blacklisted: c_05, c_11 and c_15. In the remainder of this section we validate the quality of the campaign selection criteria. We attempt to gauge the real false positive rate by inspecting domains belonging to campaigns, but do not appear on blacklists. A high false positive rate would imply that the selection criteria are imprecise and include a significant set of domains that were registered without any malicious intent. In contrast, a very high true positive rate implies that the selection criteria are substantially more exhaustive in defining domains with malicious intent compared to blacklisting services.

Transitive attribution. To assess the prevalence of incomplete blacklists and not-active malicious domains, we examine the registrant data of false positives in order to find undeniable traces that connect them to malicious domains. We base this transitive attribution on phone numbers as these are uniquely assigned identifiers that were never used in our campaign selection criteria. Thus, if the registrant's phone number is identical to that of a blacklisted registration, we consider the domain name to be part of the malicious campaign and assume that it has either not been abused yet, or was not picked up by a blacklist. In total, 3,252 campaign domains are transitively considered as malicious. As shown in yellow in Figure 5, 14 of the 20 identified campaigns are thereby completely validated.

A threat to using phone numbers to identify malicious registrants arises when an attacker retrieves the WHOIS information of a legitimate .eu domain and falsely uses it for his own registration. With three small experiments, we try to invalidate the presence of this scenario in the transitive attributed set. Firstly, we measure the time interval between the registration time of a transitively attributed domain and of the blacklisted domain that it was associated with. We find that for 2,058 domains, the malicious registration (with the same phone number) occurred within 60 seconds of the transitively attributed registration. We argue that it is virtually impossible for an attacker to observe a new registration (which is non-public information in the .eu zone), query its WHOIS data

and subsequently make a similar registration in that time interval. In a second experiment, we argue that an attacker would not exploit a benign registrant's information if those contact details are already tainted. In that regard, we filter out 965 of the remaining domains that were registered after a prior registration with the same phone number was already blacklisted. Lastly, we consult a phone number verification tool [22] and identify invalid phone numbers for 189 of the 229 remaining domains. We presume that a malicious actor would not steal benign registrant details with an invalid phone number while attempting to mimic a legitimate registrant. In the end, we observe one of these three indicators for 3,212 (98.77%) of the transitively attributed domains and conclude that this attribution is justified.

In-depth analysis of campaign c_15. After the transitive attribution step, still 30.6% of the registrations in campaign c_15 remain potential false positives. This set of domain names is further investigated.

Within campaign c_15, all domain names are composed of concatenated Dutch words (mostly 2 words, but sometimes up to 4). The same words are frequently reused, indicating that a limited dictionary was used to generate the domain names. The remaining 583 potential false positives domain names were split up by a native Dutch speaker in segments of existing words. 396 of these unflagged domain names turned out to be exclusively constructed out of Dutch words used in blacklisted domains of the campaign. As this is a very specific pattern, these domains have been labeled as *validated true positive*. The remaining domain names had either one word segment in common (172 domains) or no common word at all (15 domains). Thereafter, a new iteration of transitive attribution strategy was applied on that remaining set. Hereby, 147 registrations shared a phone number with the previously validated registrations, reducing the potential false positives to just 40 registrations.

Interstingly, we find that 95 out of the 98 registrant names that are used in c_15 can be generated with on of the the Laravel Faker generator tool forks [16] using its nl-NL language option.

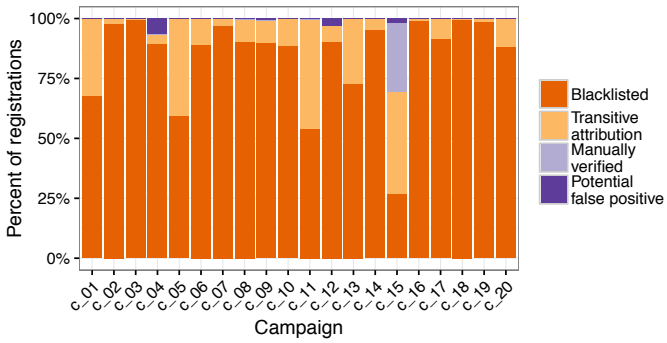

Fig. 5: Extended false positive analysis of each campaign.

Manual analysis of the remaining false positives. After the transitive attribution and the analysis of c_15, the residual potential false positives in all campaigns were further investigated manually, by querying DNS records, visiting websites, and searching on blacklists (e.g. *URLVoid* [23]) and search engines. Only two additional domains could be validated as true positives: one registration in campaign c_04 was identified as a phishing website by FortiGuard [6], and one registration in campaign c_15 sent out unsolicited to a temporary email account on *email-fake.com*.

Summary of validation. Of the 20,698 campaign registrations, 16,704 domains (80.73%) were flagged by blacklisting services, 3,252 registrations (15.71%) were linked to malicious domains via transitive attribution, and 552 (2.67%) have been manually validated as registered with malicious intent.

To conclude, the campaign selection criteria resulted in only 190 potential false positives (i.e. 0.92%). This is a strong indicator that the selection criteria are sufficiently accurate to perform a representative analysis and to give us the necessary insights into the malicious domain ecosystem.

4 Insights into malicious campaigns

In this section, we discuss several interesting observations regarding malicious campaigns, found during our assessment.

Abuse indicators and categories Overall, the vast majority of blacklisted domains (93.68%) were associated with spam domains. As listed in Table 2, all campaigns follow this general distribution, except for c_19 where nearly 28% is linked to botnet C&C servers.

Spamhaus DBL and SURBL are the two abuse sources that cover the largest number of domains. While there is a considerable overlap, both are required to get an exhaustive coverage of all campaigns. In particular, c_1 and c_19 are exclusively flagged by just one of the two sources. Interestingly, Google Safe Browsing was not involved in flagging domains in any of the campaigns. Presumably, Safe Browsing focuses more on malware delivery, as opposed to malicious infrastructure.

Cross-campaign characteristics. Some interesting characteristics exist across multiple campaigns. For instance, c_03, c_04 and c_20 generate the registrant's email address from its name followed with a numerical suffix. Similarly, the registered domain names in c_05 and c_11 follow clear character patterns with numerical suffixes. Another returning peculiarity is the discrepancy between the registrant's street address and his country. c_07, c_9, c_13 and c_14 use valid street addresses located outside of Europe (US and Panama) in combination with a European country (Norway, Ireland and others). Presumably, this is to partly confuse the residential requirements for registering a .eu domain. In the

Table 2: The different types of abuse, the blacklists and registration timing patterns per campaign. A small fraction of blacklisted domains has a missing abuse type. The max. burst represents the highest number of registrations that occurred within a 60-second time span.

	Abuse types					Blacklist sources			Registration timing patterns		
Campaign	Spam	Botnet	Malware	Phishing	Unwanted	Spamhaus	SURBL	Google SB	Day of week (Mon-Sun)	Hour of day (00-23h)	Max. burst
c_01	100.00%						100.00%				99
c_02	100.00%					100.00%	27.53%				59
c_03	100.00%					99.48%	86.82%				51
c_04	99.88%		0.12%	1.38%		99.64%	76.26%				28
c_05	83.05%					12.99%	77.97%				9
c_06	100.00%					87.63%	12.37%				3
c_07	91.40%					91.40%	1.08%				10
c_08	100.00%					100.00%	3.70%				19
c_09	99.63%		0.12%	1.97%		99.26%	28.45%				46
c_10	99.20%			1.60%		78.40%	90.40%				48
c_11	85.18%		0.08%			16.00%	77.02%				59
c_12	99.59%			0.20%		99.39%	74.29%				23
c_13	96.75%					81.82%	19.48%				1
c_14	100.00%					84.43%	86.05%				132
c_15	97.28%					73.35%	33.46%				13
c_16	100.00%			0.12%		100.00%	43.71%				8
c_17	100.00%					100.00%	8.83%				18
c_18	99.85%			0.15%		99.77%	28.04%				10
c_19	72.07%	27.93%				100.00%					5
c_20	99.29%		0.96%			99.14%	7.58%				19
All malicious	93.68%	1.27%	0.85%	3.22%	0.57%	81.07%	50.04%	1.81%			

case of c_10, a fixed street address is listed throughout the campaign while 10 different countries are combined with it.

Registration process is not fully automated. While performing the in-depth analysis of the malicious domain registrations, we found multiple indications that the malicious registration process in (at least some of) the campaigns is not fully automated: syndicates work along office hours and make human errors while registering domains.

Office hours and holiday breaks. As expected, the overall registrations in the .eu zone follow a weekly pattern. Figure 6 demonstrates this by zooming into 1 month of registrations. During weekends, a significantly smaller amount of registrations occurs than during the week. On average, week days have 2.34 times more registrations than weekend days. For blacklisted registrations, the difference is even more prominent. During weekend days, 3.85 times less malicious registrations occur as compared to weekdays. Moreover, several weekend days have no blacklisted registrations at all. Table 2 displays this behavior separately for each campaign.

As already mentioned in Section 2, the distribution over time of malicious registrations is much more fluctuating than those of all registrations (Figure 1). Interestingly, the longer drops in malicious registration activity coincide with holiday periods. The most significant one starts at the first week of July and continues for several weeks, concurring with the summer holidays. The other major periods of recess correspond to Labour day weekend (May 1), the Christmas holidays (last week of December) and the beginning of Lent or Carnival (mid-February).

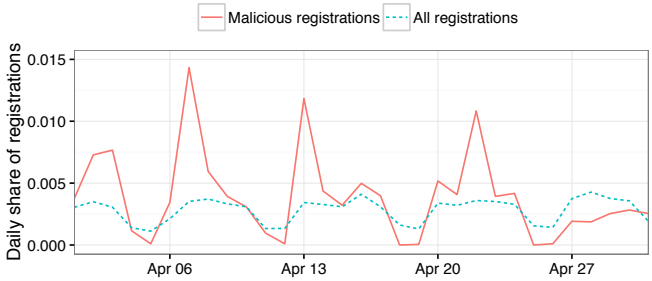

Fig. 6: Daily share of all and malicious registrations between April 1, 2015 and April 20, 2015. A clear weekly pattern is measured for both.

There are multiple hypotheses to explain these registration patterns:

1. Malicious actors might deliberately mimic normal registration patterns to avoid detection.
2. There might be a lower demand for new malicious domains during holidays, when potential victims are less active online.
3. Cybercriminal activities could be managed as any other business and are therefore equally susceptible to vacation periods.

To substantiate the latter hypothesis, we also zoomed in into the variation in registration time per campaign. Interestingly, as shown in Table 2 displays this separately for each campaign, we identified that some of the campaigns clearly align with a typical day at the office. For instance, in campaign c_11 and c_18 syndicates are working 8 to 10 hours a day, and the daily pattern of c_11 even suggests that there is sufficient time to take a lunch break. In contrast, the daily registration pattern of campaign c_19, further illustrated in Figure 7, hints at a more automated process. The vast majority of registrations are made daily at midnight and 1 PM. Furthermore, campaigns such as c_14 are registering at a rate up to 132 new domains per minute, suggesting underlying automation.

Minor inconsistencies in the data. We observe a number of inconsistencies in several registration details of certain campaigns. These inconsistencies could be

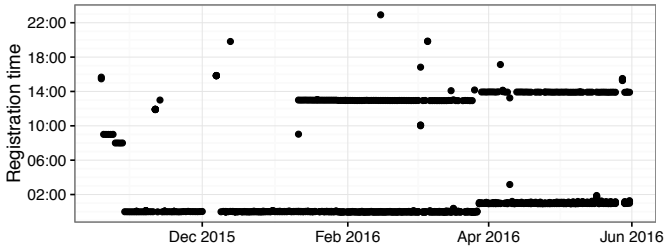

Fig. 7: Times of registrations for campaign c_19. Note the impact of daylight saving time starting from the last Sunday of March.

the consequence of small errors or typos, suggesting that some of the data has been manually entered into scripts or registration forms, or that different input validation rules have been applied by registrars or resellers.

As listed in Table 3, we encounter a few cases where registration fields belonging to the same registrant vary typographically inside a single campaign.

In campaign c_15, we also observed registrant names for which the name field has been filled inconsistently, leading to name patterns such as *Lastname Lastname* or *Firstname Firstname Lastname*.

Table 3: Minor inconsistencies found in the registration details campaign domains. Some registration details have been obfuscated for privacy reasons.

Attribute	Inconsistencies	
c_04 street	P.O BOX 3...4	P.O BOX 3...4,
c_11 city	AIX EN PROVENCE	AIX-EN-PROVENCE
c_11 street	1... ROUTE D AVIGNON	1... ROUTE D'AVIGNON
c_16 street	947 C...R	9457 C...R

Adaptive registration strategies. Several campaigns alter their strategies throughout their lifetime. For instance, five campaigns have registered domains via multiple registrars: c_01, c_03, c_11, c_12 and c_16. Figure 8 illustrates how campaign c_11 sequentially changes between 4 registrars over the entire duration of the campaign. Malicious actors might change registrars for economic reasons (cheaper domain registrations) or to evade detection. Alternatively, the change in registrar can be triggered by an intermediate reseller that changes registrar.

Table 4 lists for each campaign the amount of adaptive registration details that were used throughout its lifespan. While five campaigns use just a single phone number and email address, the large majority leverage multiple registration details. The email providers that are categorized as "Campaign" indicate that a domain name that was registered as part of the campaign, was later used as the email provider for a new registration.

As primary indicators of evasion sophistication, we list two metrics. Firstly, we give the maximum number of domains for which a campaign has reused a single phone number or email address. Secondly, we measure the longest period during which a registrants phone or email address has been reused. c_15,

c_12 and c_8 demonstrate the highest sophistication in terms of minimizing the reuse of registrant details. However, c_15 uses many different self-registered email providers and only reuses details sparsely over a long period. In other words, they leverage a more elaborate strategy than c_12 and c_8, where registrant details seem to be automatically generated in a hit-and-run fashion The success of c_15's strategy is supported by its low blacklist presence. In contrast, c_2, c_11 and c_18 deploy exhibit more simple and high-volume strategies.

Table 4: The amount of registrars, phone numbers, email addresses and types of email providers used per campaign.

		Campaign																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	Nb of registrars	3	1	2	1	1	1	1	1	1	1	4	2	1	1	1	3	1	1	1	1
	Nb of phones	4	3	19	54	1	2	1	29	14	1	2	29	1	1	97	8	1	4	1	13
	Max domains per phone	338	1026	385	169	177	158	93	20	590	125	1220	24	154	989	16	372	283	1265	752	237
	Max phone usage (days)	90	71	69	276	129	1	359	2	155	204	246	15	307	41	232	147	50	75	226	35
	Nb of email addresses	6	18	71	54	177	2	1	29	13	1	2	29	29	1	98	8	1	4	1	14
	Max domains per email	263	103	68	169	1	158	93	20	590	125	1240	24	126	989	16	373	283	1265	752	237
	Max email usage (days)	50	8	14	267	–	1	359	2	155	204	157	15	255	41	232	147	50	75	226	35
Email Providers	Public	–	1	1	2	–	–	–	6	1	–	–	1	–	1	–	3	1	1	1	1
	Private	5	–	–	–	–	2	1	–	–	1	1	–	1	–	–	–	–	–	–	–
	Campaign	–	–	–	–	–	–	–	–	–	–	–	–	28	–	98	–	–	–	–	–
	WHOIS privacy	–	–	–	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

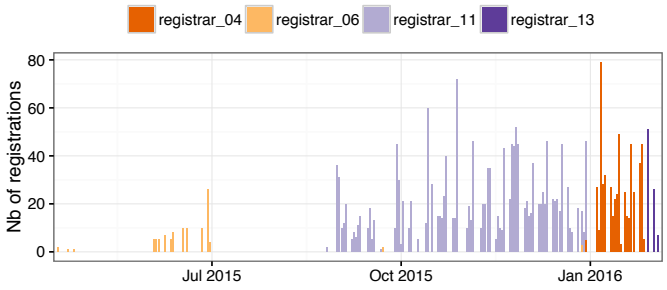

Fig. 8: Registrations per day and per registrar of campaign c_11.

Related campaigns. By searching for overlaps between campaigns in their registrants' details, as well as temporal characteristics (simultaneous or chained activity), we have identified that several campaigns are likely related to each other:

- c_02 and c_03 have registrants with the same phone number
- c_08 and c_12 have registrants with the same phone number, email and address
- c_16 and c_18 have registrants with the same address

Similarly, the abrupt ending of campaigns c_01, c_02 and c_03 suggest that these campaigns might be of the same actor, or depend on the same reseller or registrar that ended their service.

Most active malicious actors. Table 5 gives the highest represented malicious registrars, registrants and email providers in our dataset. Most surprisingly, 49.6% of all the malicious domain names are registered with one single registrar. Furthermore, it used by half of all the campaigns we identified. We argue that this registrar is either very *flexible* in accepting registrations, or has the most interesting price setting for bulk registrants. Note that this registrar only accounts for 2.27% of all benign registrations. This observation confirms earlier findings in [13,11] that a handful of registrars accounts for the majority of spammer domains.

The most used malicious email providers are all popular public webmail providers. The situation is different compared to the registrars as gmail.com has the largest share of malicious registrations but also well-represented in benign registrations. In contrast, aol.com and yahoo.com do have a large fraction of malicious registrations.

Over 3,000 malicious registrations can be attributed to just 3 registrants who are predominantly malicious. Related to the reasoning in Section 3.3, we suspect that non-blacklisted registrations of these registrants are likely malicious as well.

Table 5: Top 3 most malicious registrars, email providers and registrants. For each entry, we list their contribution to all malicious and benign registrations, their toxicity and the campaigns that are associated with them. The toxicity expresses the percentage of malicious registrations within that entity.

	Nb of malicious	Contribution Malicious	Benign	Toxicity	Associated campaigns
1. registrar_5	10,353	49.61%	2.27%	36.25%	1 2 3 4 9 10 12 13 14 17
2. registrar_3	3,004	14.39%	2.64%	12.41%	3 7 8 12 16 18
3. registrar_7	2,327	11.15%	0.46%	38.67%	20
1. gmail.com	4,221	20.23%	24.79%	2.08%	4 14 19
2. yahoo.com	3,348	16.04%	1.49%	21.85%	2 3 4 8 20
3. aol.com	2,134	10.23%	0.31%	46.28%	8 9
1. m...s@c...k.com	1,265	6.06%	0.00%	99.37%	18
2. abuse@j...n.com	1,240	5.94%	0.12%	54.89%	11
3. n...t@gmail.com	989	4.74%	0.01%	95.37%	14

5 Automating campaign identification

In the previous section, we discussed a large-scale experiment in which we manually identified large campaigns from a corpus of malicious registrations. The criteria that defined these campaigns were mainly recurring registrant and nameserver details. In this section, we use that knowledge to automate the campaign identification process by using a clustering algorithm. The results serve to both validate the manual experiment, as well as to demonstrate the capabilities of automatic campaign identification to aid ecosystem analyses in TLDs.

5.1 Clustering process

Algorithm. *Agglomerative clustering* is chosen as the basis to perform automatic campaign identification. It is a hierarchical clustering algorithm that works by iteratively merging the two clusters that are closest to each other [12]. We adopt the complete linkage criterion to determine the distance between clusters. Using this criterion, the distance is equal to that of the most dissimilar instances of both clusters, promoting a high density. There are two main reasons for opting for agglomerative clustering.

1. The algorithm does not require a predetermined number of clusters, allowing us to statistically evaluate the optimal number of clusters afterwards.
2. Given the results from Section 3, we presume that about 80% of malicious domains can be grouped into clusters. Agglomerative clustering allows the remaining independent domains to have their own singleton cluster, without necessarily polluting the large clusters.

Feature set. For each of the 20,870 blacklisted registrations, we extract 13 features. There are two general **registration features**, *domain length* and *registrar*. Next, we have ten **registrant features**: *name, street, city, region, country, zip code, phone number, email account* and *email provider*. Lastly, two **nameserver features** were included, the *nameserver domain names* and their *geographical location*.

Agglomerative clustering uses the Euclidean distance measure to calculate the distance between two instances. However, except for *domain length* and *address score*, all features in our set are categorical, not numeric. In order to accommodate these features, we apply one-hot encoding [18]: for each possible category in our set, a new binary feature is created. Each instance that carried that value will receive a value of 1 in the new binary encoded feature, all others are set to 0. Naturally, one-hot encoding dramatically increases the number of features, more specifically from 13 to 30,843.

Cutoff selection. Agglomerative clustering has no predefined stopping criteria and merges clusters until only one remains. Using the campaign labels from the manual analysis in Section 3, we calculate the *V-measure* after each merging step to statistically express the mapping between clusters and campaigns. The V-measure is the harmonic mean of the *homogeneity* and *completeness score* [19]. The former is a metric that represents how homogeneous each cluster is in terms of campaign labels, the latter measures whether the instances of a certain label are all assigned to the same cluster[6]. The highest V-measure is observed at a cutoff of 432 clusters, where the homogeneity is 0.90 and the completeness score 0.86.

[6] For instances without campaign labels, the registrant's phone numbers are set as their label.

5.2 Results

In the selected model, very large clusters have formed. Namely, 80% of domains reside in the 39 largest clusters, while a long tail of 227 clusters consisting out of only 5 registrations or less. In other words, the clustering algorithm forms a Pareto distribution similar to the manual campaign identification in Section 4. Furthermore, the top 30 clusters represent 91.48% of blacklisted registrations that reside within the 20 manually identified campaigns.

Using Figure 9, we analyze the top 30 largest clusters and their correlation to the campaign labels from the manual analysis. The clustering algorithm largely aligns with the manual campaign identification, with most clusters mapping to a single campaign. The notable exceptions being the two largest clusters. The first cluster encompasses 2,052 domains of both c_02 and c_03. This is in line with our previous speculation (Section 4) that c_02 and c_03 are related given their synchronized ending and the fact that they share registrants with the same phone number. The same is true for the second cluster, as both c_16 and c_18 clearly share registrants with the same address.

Cluster 16 is the only automatically identified cluster that solely exists out of domains without campaign labels. When inspecting those domains, we find that this cluster is likely related to or part of c_20. More specifically, their active days align and the same registrar is used for all registrations in both sets, as shown in the bottom part of Figure 10.

Several clusters also contain a small amount of instances without a campaign label. We distinguish two cases: instances that closely align to a campaign, but were not selected because of too narrow selection criteria; and instances that have no campaign affinity, but are most probably merged because the clustering algorithm has executed too many merges. The former are labeled as (Related) in Figure 9, the latter as (Unrelated).

18 of the 20 manually identified campaigns are represented in the top 30 clusters. The smallest identified campaign, c_07, is not found in this subset of clusters because it is simply too small. Cluster 30 contains 110 domains, more than c_07 encompasses as a whole. However, we find that c_07 is completely and homogeneously represented by the 35th cluster. The second campaign that is missing is c_15. As mentioned in Section 3.3, this campaign was selected by a unique and complex address formatting pattern. Since the clustering algorithm only performs binary matches on these fields, it is less effective at detecting these more advanced similarities. As shown in the top part of Figure 10, c_15 is spread out over 18 clusters, that essentially represent 18 different registrants that are reused throughout the campaign. The affinity between those clusters is clear when considering their active days.

In conclusion, the manual and automatic campaign identification results align to a large extent. We find that, when performing automatic detection using clustering, we achieve a more exhaustive identification of clear similarities as opposed to manual identification (e.g. cluster 16). However, the automatic approach has difficulties to detect more advanced similarity patterns (e.g. c_15). In future

work, more sophisticated techniques, such as n-grams, can be integrated into the clustering algorithm to detect more advanced similarity patterns.

In general, the outcome of the clustering algorithm both validates the approach of the manual analysis, as well as demonstrates the capabilities of automatic and reproducible campaign identification using registrant and nameserver details.

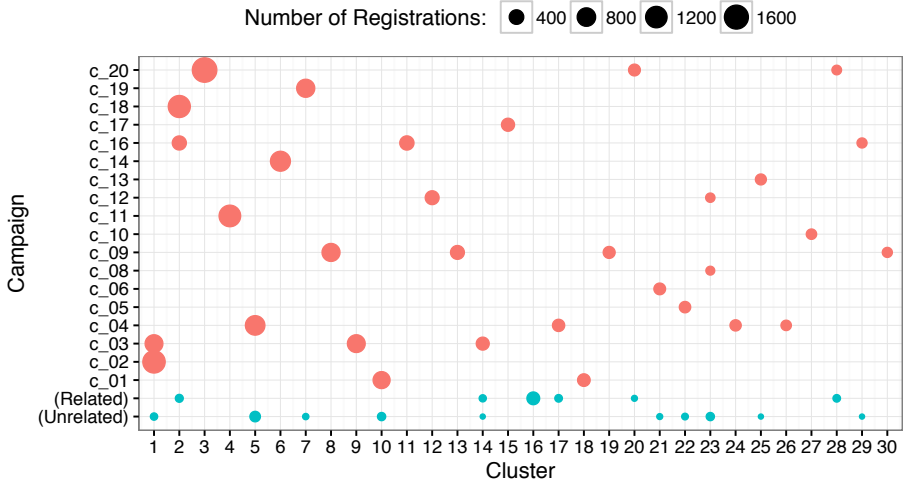

Fig. 9: Mapping of the top 30 clusters to campaign instances. The bottom two rows represent domains without a campaign label: the (Related) row groups the registrations that closely align with campaigns, the (Unrelated) groups registrations without campaign affinity. The clusters are ranked from large to small.

6 Discussion and limitations

In this section, we want to discuss the relevance and applications, as well as the limitations of our study.

Applications. Given the exploratory nature of this research, we anticipate several applications and next steps.

The relevance of this work is not limited to .eu domains. Presumably, malicious actors do not restrict their potential to a single TLD. Furthermore, bulk registrations can be made across multiple TLDs using the same registrar. Therefore, the findings and methods described in this paper can most likely be applied to other or across TLDs. To reproduce this study, access to registrant and nameserver details of registrations is required. This data can generally be obtained by downloading zone files and WHOIS data.

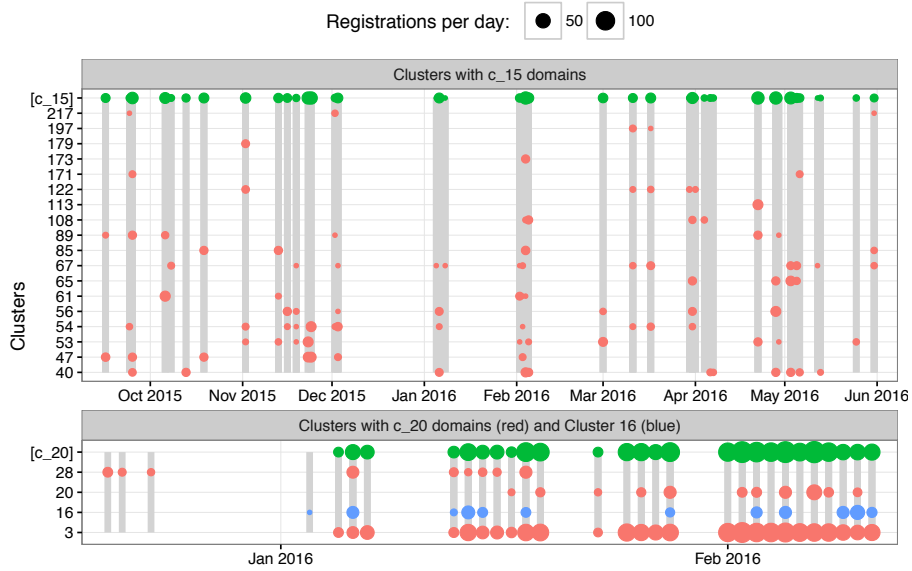

Fig. 10: Related temporal activity, highlighted in gray. **Top**: Domains of campaign c_15 are spread over many clusters. **Bottom**: Cluster 16 maps to clusters of c_20 domains.

Additionally, we demonstrate that automatic campaign identification using clustering is a feasible strategy. Moreover, 18.38% of registrations in the identified campaigns are not present on blacklists. This entices interesting opportunities to extend the coverage of blacklists. Although the proposed system relies on a post-factum analysis, it could create opportunities to stop ongoing campaigns.

Limitations. We note four limitations and potential validity threats.

Firstly, the main subjects of this research are domain names that are registered with malicious intent. However, backlists also contain legitimate registrations that have been compromised later on. We argue that the prevalence of these cases is minimal, since 98.57% of blacklisted registrations were already flagged within the first 30 days of registration. Furthermore, compromised benign domains would appear as outliers in our data and could thus hardly pollute campaign analyses.

Secondly, both the manual and automatic identification rely on patterns in the registration data. Malicious actors can leverage this dependency by constantly using different registration data and patterns. However, the cost for attackers would increase to achieve this higher level of circumvention. Furthermore, it is hard not to exhibit any pattern when performing bulk registrations (same registrars, time patterns, fake identity generating tools,...).

Additionally, several registrars offer anonymization services to their customers, obscuring the registrant contact information to the registry. Evidently,

this diminishes the ability to differentiate between registrations and conceals information that can be used to identify domains registered by the same entity. In the case of .eu, the use of such obfuscation services is not allowed by the registry's terms and conditions. During our analysis, we find that such services were only deployed by c_05 which could have impacted this campaign.

Finally, our research is based on a set of publicly available blacklists that are, at least to some extent, incomplete. A more complete ground truth would likely improve the performance of our approach.

7 Related work

Prior to our research, Hao et al. [11] studied the domain registration behavior of spammers. They reported that most spam domains are very short-lived. More specifically, 60% of these domains were active for only a single day. Spammers are registering many "single-shot" domains to minimize interference by blacklists. To counter this strategy, the authors explore various features on which spam domains exhibit distinctive behavior. For instance, in contrast with benign registrations, they find that malicious domains are more often registered in batches.Recently, Hao et al. implemented many features discussed in that prior work to create a machine learning-based predictor capable of detecting malicious domains at time-of-registration [10]. The three most dominant features of their classifier are authoritative nameservers, trigrams in domain names and the IP addresses of nameservers.

While both papers approach malicious domains as a two-class problem (benign vs. malicious registrations), many of their features essentially depend on returning characteristics of different underlying malicious campaigns. In this work, we are the first to shift the focus to the campaigns itself, exploring their modus operandi and different identifying characteristics.

A method related to ours was proposed by Felegyhzi et al. [5], who investigated the feasibility of proactive domain blacklisting, by inferring other malicious registrations from known-bad domains through shared nameservers and identical registration times. The proposed system shortens the time required to blacklist malicious domains, while providing important insights regarding the similarities of registrations within campaigns. Additionally, Cova et al. [4] indentified different rogue antivirus campaigns by looking at the hosting infrastructure and registration details (including the registrant's email) of different domains.

Related studies concentrate on DNS traffic of newly registered domains to characterize malicious behaviour [3,9,1,2,14]. These systems mainly focus on the initial operational DNS patterns of domain names.

Other important efforts regarding malicious domains come from the study of domain generation algorithms (DGAs). Recent work by Plohmann et al. [17] demonstrates the increasing importance of understanding DGAs to thwart C&C communication. Using reimplementations of these algorithms, the authors execute forward generation of domain lists, which enables proactive identification of C&C domains.

8 Conclusion

In this study, we analyzed the maliciously-flagged .eu registrations over a 14-month period. This paper is the first to extensively dissect the underbelly of malicious registrations using registrant details to identify its operational components, namely campaigns. We explored the ecosystem and modus operandi of elaborate malicious actors that register massive amounts of domains for short-lived, malicious use.

By searching for shared characteristics, we established that at least 80.04% of all malicious registrations can be attributed to 20 campaigns with varying duration and intensity. Moreover, the coverage of blacklists can be extended by 19.30%, using the information from the campaign identification. After a rigorous evaluation, we are able to confirm that the vast majority of these previously undetected registrations are genuinely related to malicious activity; at most 0.92% are false positive registrations.

Our study demonstrates the potential to leverage the registrant details and other registration characteristics to identify large campaigns. Aided by an automatic identification process, this insight can be used to easily track and interfere with massive, long-running campaigns and to preemptively extend blacklists with malicious domains that have yet to be actively used by a cybercriminal.

Acknowledgements

We thank the reviewers for their valuable feedback. We would also like to express our gratitude to the PC chairs and in particular our shepherd, for supporting us in improving the paper.

References

1. Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., Feamster, N.: Building a dynamic reputation system for dns. In: Proceedings of the 19th USENIX Conference on Security. pp. 18–18 (2010)
2. Antonakakis, M., Perdisci, R., Lee, W., Vasiloglou, II, N., Dagon, D.: Detecting malware domains at the upper dns hierarchy. In: Proceedings of the 20th USENIX Conference on Security. pp. 27–27
3. Bilge, L., Sen, S., Balzarotti, D., Kirda, E., Kruegel, C.: Exposure: a passive dns analysis service to detect and report malicious domains. ACM Transactions on Information and System Security (TISSEC) 16(4), 14 (2014)
4. Cova, M., Leita, C., Thonnard, O., Keromytis, A.D., Dacier, M.: An analysis of rogue av campaigns. In: International Workshop on Recent Advances in Intrusion Detection. pp. 442–463. Springer (2010)
5. Felegyhazi, M., Kreibich, C., Paxson, V.: On the potential of proactive domain blacklisting. In: Proceedings of the 3rd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More. pp. 6–6 (2010)
6. FortiGuard Center: Antispam - IP & Signature Lookup (2017), `https://www.fortiguard.com/more/antispam`

7. Google: Google Safe Browsing (2016), https://developers.google.com/safe-browsing/
8. Google: Unwanted Software Policy (2016), https://www.google.com/about/company/unwanted-software-policy.html
9. Hao, S., Feamster, N., Pandrangi, R.: Monitoring the initial dns behavior of malicious domains. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 269–278. ACM (2011)
10. Hao, S., Kantchelian, A., Miller, B., Paxson, V., Feamster, N.: Predator: Proactive recognition and elimination of domain abuse at time-of-registration
11. Hao, S., Thomas, M., Paxson, V., Feamster, N., Kreibich, C., Grier, C., Hollenbeck, S.: Understanding the domain registration behavior of spammers. In: Proceedings of the 2013 Conference on Internet Measurement Conference. pp. 63–76 (2013)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics (2001)
13. Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Félegyházi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Weaver, N., Paxson, V., Voelker, G.M., Savage, S.: Click trajectories: End-to-end analysis of the spam value chain. In: Proceedings of the 2011 IEEE Symposium on Security and Privacy. pp. 431–446. SP '11, IEEE Computer Society, Washington, DC, USA (2011), http://dx.doi.org/10.1109/SP.2011.24
14. Moura, G.C., Müller, M., Wullink, M., Hesselman, C.: ndews: A new domains early warning system for tlds. In: NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium. pp. 1061–1066. IEEE (2016)
15. Myles, P.: DomainWire Global TLD Report 2016/2, https://www.centr.org/library/statistics-report/domainwire-global-tld-report-2016-2.html
16. Pattanai: Faker is a PHP library that generates fake data for you, https://github.com/teepluss/laravel-faker
17. Plohmann, D., Yakdan, K., Klatt, M., Bader, J., Gerhards-Padilla, E.: A comprehensive measurement study of domain generating malware. In: 25th USENIX Security Symposium. pp. 263–278 (2016)
18. Scikit-learn developers: Encoding categorical features (2017), http://scikit-learn.org/stable/modules/preprocessing.html
19. Scikit-learn developers: Homogeneity, completeness and V-measure (2017), http://scikit-learn.org/stable/modules/clustering.html
20. SURBL: SURBL - URI Reputation Data (2016), http://www.surbl.org
21. The Spamhaus Project Ltd.: The Domain Block List (2016), https://www.spamhaus.org/dbl/
22. Twilio: Lookup (2017), https://www.twilio.com/lookup
23. URL Void: Website Reputation Checker Tool (2016), http://www.urlvoid.com/
24. Vixie, P.: Domain name abuse: How cheap new domain names fuel the ecrime economy. Presentation at RSA Conference 2015 (2015)