A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints

Victor Le Pochat*, Tim Van hamme*, Sourena Maroofi§, Tom Van Goethem*,
Davy Preuveneers*, Andrzej Duda§, Wouter Joosen*, Maciej Korczyński§

*imec-DistriNet, KU Leuven
{firstname.lastname}@kuleuven.be

§Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG
{firstname.lastname}@univ-grenoble-alpes.fr

***Abstract*—In 2016, law enforcement dismantled the infrastructure of the Avalanche bulletproof hosting service, the largest takedown of a cybercrime operation so far. The malware families supported by Avalanche use Domain Generation Algorithms (DGAs) to generate random domain names for controlling their botnets. The takedown proactively targets these presumably malicious domains; however, as coincidental collisions with legitimate domains are possible, investigators must first classify domains to prevent undesirable harm to website owners and botnet victims.**

The constraints of this real-world takedown (proactive decisions without access to malware activity, no bulk patterns and no active connections) mean that approaches from the state of the art cannot be applied. The problem of classifying thousands of registered DGA domain names therefore required an extensive, painstaking manual effort by law enforcement investigators. To significantly reduce this effort without compromising correctness, we develop a model that automates the classification. Through a synergetic approach, we achieve an accuracy of 97.6% with ground truth from the 2017 and 2018 Avalanche takedowns; for the 2019 takedown, this translates into a reduction of 76.9% in manual investigation effort. Furthermore, we interpret the model to provide investigators with insights into how benign and malicious domains differ in behavior, which features and data sources are most important, and how the model can be applied according to the practical requirements of a real-world takedown.

I. Introduction

On November 30, 2016, a global consortium of law enforcement agencies and Internet stakeholders completed a four-year investigation aimed at dismantling the Avalanche infrastructure [31], which has been called "the world's largest and most sophisticated cybercriminal syndicate law enforcement has encountered" [94]. For seven years, this 'bulletproof hosting service' [13] offered services to cybercriminal operations through a 'crime-as-a-service' model [94], fully managing all technical aspects of carrying out malware attacks, phishing, and spam campaigns. It supported a botnet of a massive scale: Avalanche was responsible for two thirds of all phishing attacks in the second half of 2009 [8], and ultimately affected victims in over 180 countries with estimations of its monetary impact reaching hundreds of millions of euros worldwide [6]. The takedown operation in 2016 was supported by authorities from 30 countries and culminated in five arrests, 260 servers being taken offline and the suspension of over 800,000 domains [31].

Network and Distributed Systems Security (NDSS) Symposium 2020
23-26 February 2020, San Diego, CA, USA
ISBN 1-891562-61-4
https://dx.doi.org/10.14722/ndss.2020.24161
www.ndss-symposium.org

As part of this dismantling, a large domain takedown effort sought to disable the botnet's communication infrastructure. This effort targets the large sets of domains that the malware families of Avalanche generate through *domain generation algorithms* (DGAs). Through this 'domain fluxing' [71], infected hosts attempt to contact all generated domains, whereas the botnet master only needs to register one to continue operating the malware, decreasing the likelihood of blacklisting and takedown. However, as security researchers have reverse-engineered several of these DGAs [71], law enforcement is able to identify upfront which domains the malware will try, after which these can be blocked or seized. Over four yearly iterations of the Avalanche takedown, more than 4.3 million domains were thus prevented from being abused, making it the largest domain takedown so far [7].

Previous work related to DGAs focused on detecting *malicious* domains in regular traffic, relying on strong indicators of *ongoing* malware activity, to discover new malware families or find infected hosts inside a network [16], [82], [100]. In this paper, we address the orthogonal issue that the Avalanche takedown faces: given – presumably malicious – DGA domains that will be generated in the future and should *proactively* be taken down, we seek to detect those that accidentally collide with *benign* domains. In particular, we assess how we can effectively support law enforcement investigators with an automated domain classification to inform the appropriate takedown action in a real-world use case. This reduces the extensive manual effort previously invested in this classification, while still maintaining the high accuracy required in such a sensitive operation. Taking down benign domains may cause prejudiced service interruption and harm their owners. At the same time, we have to guarantee that no malicious domain is left untouched, as this would allow malicious actors to target infected users once again.

We are the first to develop an approach that can be used to effectively identify the domains registered with malicious intent, within the constraints of a real-world takedown operation. First, *bulk patterns* no longer apply, both for domains that are benign (due to the accidental uncoordinated collisions) and malicious (due to the low number of required domains). Second, as the takedown is *proactive*, we cannot search for malicious activity (any ongoing activity would mean that infected machines are implicated in actual attacks and defeat the proactive purpose of the takedown). Third, we *cannot actively contact domains* so that the takedown can occur stealthily (otherwise attackers could evade detection and undermine the takedown). Instead, we rely on capturing more generic differences in how benign and DGA-generated malicious domains are registered and operated.

We design a machine learning-based model for classifying benign and malicious domains, and we evaluate it on ground truth from the 2017 and 2018 iterations. Using a human-in-the-loop approach that combines automated classification and manual investigation targeted at the most difficult domains, we achieve an accuracy of 97.6% for the real-world Avalanche use case, ensuring high correctness while still vastly reducing manual effort: in the 2019 iteration, our approach reduced this effort by 76.9%. However, we go beyond reporting this metric with an extensive analysis of the benefits and limitations brought by the machine learning approach as well as the real-world setting. We provide an interpretation for the factors that impact the decisions of the model, giving insight into how the owners of benign and malicious domains behave differently and how the model uses this information to make decisions. These insights can help law enforcement in their choices regarding the acceptable performance and reliability of the model.

Malware creators increasingly employ techniques that make the takedown of their command and control infrastructure more complex, and the scale of malicious operations continually increases. Further automation of the takedown process with our classifier of malicious and benign domains can support law enforcement in coping with the increased complexity. However, we need to carefully design, evaluate, and analyze such an approach to cope with the constraints of a real-world application as to avoid any adverse effect on the legitimacy of the operation. This enables law enforcement to continue disrupting malware infrastructure and protecting potential victims.

In summary, our contributions are the following:

- We assess to what extent an automated approach can assist law enforcement investigators in correctly detecting the collisions with benign domains among registered domains implicated in the Avalanche takedown, without the ability to rely on bulk malicious registrations, ongoing malware activity or actively collected traffic.
- We develop a technique where we complement a machine learning model with targeted manual labeling of the most informative and difficult domains, to maintain performance across multiple takedown iterations while still vastly reducing the required manual investigative effort.
- We evaluate how well this approach performs and transfers for the 2017 and 2018 takedowns: we obtain an accuracy of 97.6%. The predictions of our model were used in the 2019 takedown, and we find a subsequent reduction in manual investigative effort of 76.9%.
- We critically examine the factors that impact the performance and decision-making process of our model. We find that time-based features are the most important ones, which at the same time are the most costly to evade. In terms of data set availability, WHOIS data greatly improves accuracy, which shows its importance for conducting effective cybercrime investigations.

II. Background

A. Domain generation algorithms

Machines in a botnet such as Avalanche communicate with the malicious actor through command and control (C&C) servers. Early malware hard coded the domain names or IP addresses of their C&C servers, so it was easy to obtain this information and either blacklist the servers or even take over the corresponding infrastructure (by pointing for instance the domains to 'safe' IP addresses and/or having hosting providers take C&C servers down), effectively stopping the malware from further malicious operation [18]. Malware has therefore evolved from hard coding the C&C server information to dynamically creating or updating it.

TABLE I. Examples of domains generated by Avalanche DGAs.

	Domain	Malware	Validity
1	0a85rcbe2wb5n5fkni4i4y[.]com	CoreBot	Jan 21, 2018
2	researchmadness[.]com	Matsnu	Jan 28-31, 2018
3	arbres[.]com	Nymaim	Mar 9, 2018
4	sixt[.]com	Nymaim	always

One technique of this dynamic approach is 'domain fluxing', in which domain generation algorithms (DGAs) create up to thousands of algorithmically generated domains (AGDs) every day [71]. The malware will then attempt to contact these domains and ignore the unavailable ones: the botnet owner therefore only needs to set up one of the generated domains to host a C&C server [18]. Avalanche combined this technique with 'fast fluxing', in which compromised machines hosting a proxy to the C&C server as well as the corresponding DNS entries of the AGDs rapidly switch [41], thus further evading blacklisting and takedown [31].

DGAs take as seeds parameters known to both the malware owner and the infected host, so that they both generate the same set of domains [18], [71]. These parameters such as the length of domains, top-level domains (TLDs) to use, or seeds for pseudo random number generators can be hard coded. More complex algorithms may depend on time: one of the inputs to the DGA is then the current time, either from the system clock or retrieved from a common source (e.g., GET requests to legitimate sites [99]). In this way, the DGA creates domains having a certain *validity period*: the time frame during which the seed timestamps make the DGA generate that domain, which the infected machines then attempt to reach. For Avalanche malware families, these validity periods range from 1 day (e.g. Nymaim) to indefinitely (e.g. Tiny Banker).

We can further distinguish between deterministic DGAs that know all parameters upfront and non-deterministic DGAs that know some parameters only at the time of generating the domains: e.g., the DGA of the Bedep family uses exchange rates as seeds [79]. Avalanche did not use any non-deterministic DGAs so for successfully reverse-engineered DGAs [3], [71], we can generate all potential AGDs ahead of their validity, by varying the timestamp that serves as input to the DGA.

Table I lists example names generated by DGAs, from malware hosted by Avalanche. While Example 1 appears random (a long name with many digits and no discernible words), certain DGAs generate names that look much more like legitimate domains. Example 2 shows a name generated based on a word list yielding domains that may correspond to a regular domain name. Example 3 shows a short yet randomly generated name for which there is a high probability of generating either a valid word or a plausible abbreviation. These last two examples have a high probability of generating domains that collide with existing benign domains.

Finally, certain malware families alter domain resolution on the infected host, generating traffic to hard-coded and otherwise benign domains that actually resolve to malicious IP addresses to circumvent domain-based filters [40]. While these domains are not algorithmically generated, they are present in malware code and traffic and must therefore also be classified as part of the takedown operation, to distinguish them from other hard-coded and actually malicious domains. Example 4 is one such instance using the domain of the Sixt car rental site. We include these domains in our classification, but for brevity, we refer to all domains to be classified as the 'registered DGA domains'.

B. Taking down the Avalanche infrastructure

The perpetrators behind the Avalanche infrastructure offered two services for rent by cyber criminals: registering domain names as well as hosting a layered network of proxy servers through which malware actors could control infected hosts and exfiltrate stolen data [3]. Avalanche thereby supported the operation of 21 malware families [5], controlling a botnet of an estimated one million machines at the time of takedown [3].

Prosecutors completed the first iteration of the takedown in November 2016, where the whole infrastructure was dismantled through arrests, server seizures, and domain name takedowns [31]. For the latter, the first iteration targeted live C&C domains, but also those that would be generated by the DGAs in the coming year, preemptively blocking these to prevent Avalanche from respawning. This effort has been repeated every year since, as in January 2020 infected machines on over two million IPs still contacted the Avalanche network [1], highlighting the potential damage if Avalanche were to respawn.

Coupled with the large number of malware families and the extensive amount of domains that these DGAs generate, this results in a large number of DGA domains to be processed. For the three yearly iterations from 2016 to 2018, this amounts to around 850,000 domains per year [5], [7], while the 2019 iteration looks ahead five years and therefore treats almost 2 million domains: this means more than 4.3 million targeted domains have been processed in total. For the DGA domains in the Avalanche takedown, law enforcement took one of three actions on the takedown date [4]:

- *Block registration:* for a not yet registered domain, the TLD registry blocks registration. This is the case for the vast majority of domains.
- *Seize domain:* for a domain registered by a seemingly malicious actor, it is seized from the original owner and 'sinkholed', i.e. it is redirected to servers of the Shadowserver Foundation. Optionally, domains are also transferred to the "Registrar of Last Resort". Through sinkholing, law enforcement can then track how many and which infected hosts attempt to contact the domains [1] and aid in mitigation through notifications to network operators and infected users [22]. Domain *seizures* require a legal procedure such as a court order, while organizations could also *request* a takedown through a 'takedown notice' [42].
- *No action:* for a domain registered by a seemingly benign actor (including domains sinkholed by other security organizations), no action is taken by law enforcement and the domain remains with its original owner.

TABLE II. NUMBER OF BENIGN AND MALICIOUS DOMAINS PER ITERATION. *: ACCORDING TO OUR CLASSIFICATION.

	2017	2018	2019–2024*
Benign	1397	1014	4945
Malicious	1145	402	1053
Classified	2542	1416	5998
Sinkholed	1177	594	2293
Total	3719	2010	8291

III. PROBLEM STATEMENT

A. Making accurate takedown decisions

The aim of the Avalanche takedown is to prevent the botnet owners from interacting with infected machines by blocking access to the required domains that the DGAs will generate in the year following the takedown. However, as these DGAs may generate labels that collide with benign sites, performing a blanket takedown of all generated domains would harm legitimate websites. For Avalanche, public prosecutors therefore first had to manually classify domains into benign and malicious: as shown in Table II, they had to determine an appropriate action for a few thousand registered DGA domains each year.

For registered domains, an incorrect decision may have unintended adverse effects [23], [42]. In case of the seizure of a benign domain, its legitimate owner can no longer provide its service to end users. Owners may experience lengthy downtime, as challenging an illegitimate seizure and regaining the domain can be an opaque and difficult process [42], [49]; it appears that this also holds for Avalanche domains [21], [66].

Conversely, not preemptively seizing a malicious domain allows the botnet to respawn and continue its malicious operation: as the takedown does not remove the malware from infected machines, these will continue to establish contact with DGA domains. Once the botnet owners can obtain such a domain, the attackers can launch new attacks or spread malware to additional hosts. The takedown efforts, intended to permanently stop the malware, are then effectively spoiled.

Manually classifying all DGA domains is a resource- and time-consuming process, where due to 'decision fatigue' [28], [90], the mental effort in making repetitive decisions could lead to biases. Given the severe consequences of incorrect classifications, our goal is to develop an automated approach to the classification of DGA domains that performs with high accuracy, in order to relieve human investigators from manual effort as much as possible. At the same time, this does not preclude a manual review of those domains that are the hardest to classify or that could have the most significant effects. In the analysis of our approach in Section V, we quantify how such a union of automated and manual classification can still lead to a significant reduction in required effort. Through such a reduction in manual effort and time, we can ensure the correctness of takedown decisions, thereby minimizing negative effects on website owners as well as end users.

B. Constraints for distinguishing malicious and benign domains

While our base goal is to distinguish malicious and benign domains, we cannot use previously proposed solutions as they rely on certain indicators that would not work for the Avalanche

TABLE III. OVERVIEW OF GOALS AND STRATEGIES FOR THE DIFFERENTIATION OF BENIGN AND MALWARE/DGA DOMAINS.

Context/Detection goal	Individual patterns	Proactive analysis	No active connections	Related work
Active malware domains within regular traffic	✗	✗	✓	[15], [16], [19]
Likely DGA domains within regular traffic	✗	✗	✓	[26], [78], [96]
Future malicious domains at registration	✗	✓	✓	[33], [38], [86]
Benign domains within known malware domains	✓	✗	✗	[47]
Benign domains within future DGA domains	✓	✓	✓	*Our work*

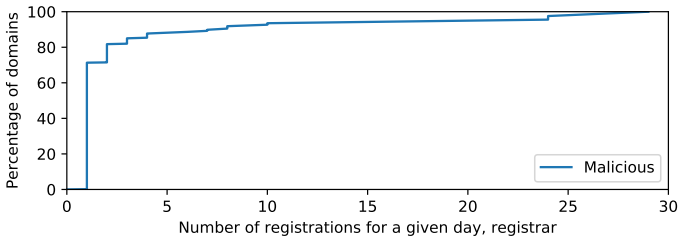


Fig. 1. Cumulative distribution of registration counts for a given day and registrar, for malicious domains from the 2017 and 2018 iterations.

use case. Concretely, these indicators no longer hold for malicious domains (e.g. bulk registration), cannot be observed by us (e.g. detecting malware activity), or are counterproductive (e.g. alerting the attacker). Table III summarizes how the different contexts, goals and strategies of previous works do not fully satisfy our requirements.

The reason is that the assumptions made in previous work no longer hold due to a different balance between malicious and benign domains: instead of detecting domains with clear malicious behavior among a (large) set of regular traffic, we assume that domains are malicious (they would be contacted by malware) and need to detect benign domains (i.e. accidental collisions). While in previous approaches, domains that do not exhibit strong indicators of maliciousness (offered by the former) are benign, the absence of such indicators in our use case means that we may not make such an assumption, and makes those previous approaches ineffective for Avalanche.

We translate these unique characteristics of the Avalanche takedown into three constraints. First, we need to take the characteristics of benign domains into account as well, by developing appropriate features that capture *individual differences in registration and configuration*. Second, as we cannot leverage ongoing malware activity itself, we need to develop features that allow for a *proactive analysis*. Third, attackers may not evade or detect data collection, so we may *not make any active connections* to domains in order to remain stealthy. In this section, we elaborate on these challenges and differences that make previous approaches ineffective for our use case.

a) Individual registration and configuration patterns: Previous work often assumes that specific (bulk) patterns in the setup of domains indicates maliciousness.

For example, PREDATOR [38] relies on the observation that in order to evade blacklisting, malicious spam domains are registered in bulk (over 50% in groups of ten or more at one registrar in five minute intervals), causing these temporal clusters to be similar in infrastructure, lexical composition and life-cycle stage. In a similar spirit, Premadoma [86] relies on similarities in registrant data and the prevalence of malicious domains at specific facilitators (such as registrars) to detect sustained large-scale malicious campaigns. However, these patterns are no longer usable for our set of domains. Attackers only need to register one of the domains that the DGA outputs at a given time, so they no longer need to register domains in bulk, as is necessary for spam domains, also reducing the likelihood that they share e.g. registrars. Figure 1 confirms this: 93.5% of malicious domains in the 2017 and 2018 iterations of the Avalanche takedown are registered in clusters of fewer than 10 domains at their given registrar in one day (as opposed to the five minute interval in PREDATOR [38]). Moreover, the accidentally colliding benign sites do not have any relationship and will therefore not share any properties either.

Systems such as DeepDGA [96] and FANCI [78] detect DGA domains from linguistic patterns in their label. However, we know that all domains are either generated by a DGA or hard coded in malware, so it would be incorrect to use such patterns to categorize them as malicious.

In summary, because of the characteristics of our domain set (singular malicious and unrelated benign domains, all output by a DGA), many of the assumptions that the above approaches make on patterns that determine maliciousness are no longer valid. We must therefore resort to capturing more generic, common registration and configuration patterns for individual domains. These patterns should not only capture 'obvious' maliciousness, but also properties that indicate benignness.

b) Proactive analysis: Previous work relies on observing ongoing malicious behavior: e.g. Exposure [19] leverages irregular DNS configurations and access patterns to detect 'domain flux' [41]; Pleiades [16] captures patterns in `NXDOMAIN` responses to DNS queries by active malware. These systems rely on ongoing malware activity that generates the analyzed traffic. Similarly, systems that use only the label to detect DGA candidates based on their appearance [26], [78], [96] need ongoing malware activity, otherwise infected hosts are not contacting malicious domains that are then visible in traffic.

Crucially, because malicious domains have to be taken down before they can cause any harm, we have to classify them proactively, i.e. before infected machines would actively query the malicious domain. This distinguishes our work from the above works, as we cannot analyze and rely on patterns within any (ongoing) malware activity. While we can and do use features similar to those from previous systems, we are restricted to detecting patterns in registration, configuration, and regular traffic. Moreover, we already know that a DGA generated the domains that we have to classify, meaning that we start with an assumption that the domains are malicious.

c) No active connections to domains: Internet measurements can be classified into two groups: passive collection, where already ongoing traffic is observed, and active collection, where new traffic is injected into the network. Notos [15] and Exposure [19] are examples of systems that analyze patterns in passively collected DNS queries. In contrast, Mentor [47] relies in part on website content features to measure positive domain reputation, requiring active and targeted data collection through crawling the domains.

While we have a similar goal to Mentor of detecting benign domains within presumably malicious domains, we avoid including features that require us to actively connect to domains. Malicious actors are namely known to detect active scanning and respond differently to appear more benign ('cloaking') [46], and could thus mislead our classification. More broadly, such probes could alert them of efforts to investigate and disrupt malicious infrastructures, allowing attackers to shift their approach or hide any traces to avoid repercussions [3]. A stealthier analysis without targeted active data collection therefore avoids endangering the effectiveness of ongoing investigations [19], [102].

C. Ground truth data

The advantage of our collaboration with law enforcement is that we can use their manual classification of benign and malicious domains from the takedown as a trustworthy source of ground truth. Previous studies mostly rely on publicly available blacklists and whitelists as the labeled ground truth [89], but malware blacklists have been found to contain benign parked or sinkholed domains and are ineffective at fully covering domains of several malware families [54], while lists of popular domains commonly used as whitelists can easily be manipulated by malware providers [56].

However, the real-world context of the Avalanche takedown affects the composition of our ground truth data. Concretely, our data set is relatively small, as seen in Table II. Plohmann et al. [71] have seen a similarly small proportion of registered domains among DGA domains. We can expect this number to be small: malicious actors only need to register few domains, as the malware will try all DGA-generated domains; conversely, benign actors are less likely to be interested in using the often random-looking domains generated by the DGAs. Previous studies are able to evaluate their approach on much larger data sets, albeit self-constructed and arbitrarily selected. Nonetheless, training on a small data set is a challenge that prosecutors would also face, and our analysis is therefore valuable for informing them on the feasibility, constraints and benefits of an automated approach for such a practical use case.

D. Ethical considerations

We use the data set of the Avalanche takedown shared with us by our law enforcement partner. We augment this data with third-party data, avoiding unnecessary active probes of both benign and malicious domains. However, given the sensitivity of the former and commercial agreements for the latter, we cannot share this data with external parties. We release the data processing scripts and resulting models at https://github.com/DistriNet/avalanche-ndss2020 to support reproducibility.

We assisted law enforcement agencies by applying our approach to the 2019 Avalanche iteration. While the use of machine learning for law enforcement purposes may be contested [69], human investigators may similarly make involuntary errors, e.g. due to 'decision fatigue' [28], [90].

IV. Data set analysis and feature extraction

To determine a suitable takedown action for algorithmically generated domains (AGDs), we search for relevant features providing a full view of their properties over time. We then create a classifier that detects whether patterns in these properties are more likely to correspond to a benign or malicious domain without having to rely on ongoing malware activity.

In this section, we first analyze how different data sources can track different stages of the domain *life cycle* and we discuss the *insights* on how features capture contrasting properties of benign and malicious domains. Then, we select the final set of *features* and discuss the reasons for omitting certain features.

A. Life cycle of a domain

To correctly identify the intent of a domain registration, we need to observe patterns in the domain life cycle, as they indicate who obtained the domain, how they use it, and how they value it. For each identified step, we determine which relevant features capture the actions of the domain owner and list sources that track this information. Through our analysis, we can then ensure that our selection of features and data sets appropriately covers each step.

L1. Choice of the domain name: The prospective owners of a domain (the registrants) must first choose the domain name that they want to purchase. Usually, the name is chosen to be easily memorized, sufficiently short, and representative of the service provided by the domain, but as malicious actors will need to produce domains in bulk, they will generate them automatically. The resulting names have a random or patterned appearance that we can capture in lexical features on the label itself in order to automatically detect DGAs [77], [78], [96].

L2. Registration of the domain: A registrant registers a domain through a registrar, typically paying a registration fee for at least 1 year [44] (although free and shorter offers exist [35] that tend to attract abuse [50]). The registrant identity, the registrar used, and the timestamps of the registration start and end are then made publicly available in the WHOIS database. We can then extract the registration patterns to distinguish benign and malicious sites [60]. Due to privacy concerns and regulations (e.g., the European General Data Protection Regulation), the publicly available identity of the registrant may be obfuscated: the real identity is then only available to the registrar and the top-level domain (TLD) registry. This data may be leveraged in collaborations with registries, e.g. for detecting malicious domains at registration time [86], [93].

L3. DNS configuration: Once a domain has been registered, its entry in the Domain Name System (DNS) must be configured to allow discovery of its services using the domain name. The nameserver is passed onto the TLD registry and will appear in its zone files. The domain resource records configured in the nameserver zone file then become available for querying. Active DNS data sets (collected by e.g., OpenINTEL [91]) rely on scanning zone files or popular domains to obtain these records, while passive DNS data sets (collected by e.g., Farsight Security [32]) extract them from monitored DNS responses. Both types of data sets have been used to detect malicious domain registrations and activity [19], [52], [84].

L4. Setup of the service infrastructure: The main purpose of a domain name is usually to provide a service for which an infrastructure needs to be set up. The records stored in DNS may reveal the hosting infrastructure or third-party service providers (e.g., cloud providers) from which

actors that enable malicious activity can be derived [72], [101]. A scan of open ports accompanied by "banner grabs" may reveal provided services and the content available through the service may reveal its purpose. Such an operation requires active probing of the domain, which either can be executed ad hoc or is already performed regularly by e.g. Censys [30] and Project Sonar [73], whose scale enables analyses of botnet devices [14]. Furthermore, certificates obtained by the domain owner for their service may also be tracked in Certificate Transparency logs [55].

L5. Service activity: Once the service is set up, end users can start interacting with it. Traffic to the service may be logged either at the server, the client, or in any network in-between. These logs can then be analyzed for multiple purposes. Malicious behavior can be detected and publicly shared in blacklists [54], [81], [101]. Commercial providers publish lists of the most popular websites that become base sets of seemingly benign domains [56]. The service may be crawled to populate search engine results or archive web content [37]: the latter enables longitudinal analyses of malicious activity [12], [83], [101]. These methods can be combined to calculate risk scores for the domain [43].

L6. Service unavailability and domain expiration: The unavailability of the services offered by the domain, either intentionally or unintentionally due to misconfigurations, may be detected by any of the previously discussed data sets depending on the type of disruption. Once a domain is no longer needed, it may expire: domains that are set to expire are often monitored for drop-catching [39], i.e., registering domains as rapidly after expiry as possible. Malicious actors also reuse previously expired domains to capitalize on the reputation of those domains [57], [97]. Alternatively, a service may be interrupted or a domain may be made unavailable for legal reasons, e.g., in takedown operations. As we study domains before they would be taken down, we do not consider this last step in our final feature set.

B. General insights

We want to design features that exhibit contrasting properties of benign and malicious domains and therefore provide a more accurate classification, while still acting within the constraints imposed by the Avalanche takedown use case (as outlined in Section III-B). This requires insights into the generic differences in behavior of legitimate and malicious actors with respect to their domains. We choose our features to capture the following three characteristics:

i1. Likelihood of collisions: Given that all domains are algorithmically generated, our target is to find "regular" (least random) looking domains as they are more likely to be a collision with a benign domain, which is opposite to other work that focuses on detecting DGAs solely based on how random their domain names appear [77], [78], [85], [96].

i2. Investment in the domain: Obtaining and (validly) maintaining a domain requires an investment from its owner, both monetary for paying the registration fee and in effort for setting up DNS and WHOIS records correctly and installing services attached to the domain. While benign owners value their domains and are willing to make such an investment, the opposite is true for malicious actors: they want to set up a campaign with minimal cost and effort to maximize their revenue. Certain indicators imply high investment, such as long-term registration (benign domains tend to be older, while malicious domains tend to be registered shortly before the start of the validity period [19], [20], [36], [71]) or valid DNS and WHOIS records (invalid, obfuscated or repeated values hint at malicious practices [93]).

i3. Website popularity: Establishing a website that attracts sufficient traffic and is therefore perceived as popular, requires significant effort in creating content and building an audience. Website popularity is therefore an indication of benignness: malicious actors will not make the effort of setting up real websites on dormant domains, especially as it is not required for the correct operation of botnets. Regular users as well as web crawlers are also unlikely to end up on these domains. Moreover, if the domain has not yet been generated by a DGA, its traffic is low or non-existent, so we can assume that any traffic that the domain draws is legitimate.

C. Summary of feature sets

We aim to capture the broadest view possible of the life cycle of the domains to classify, and select the features and the data sources that provide their values accordingly, further inspired by our general insights. While potentially useful, certain features are not applicable to our use case or would have unwanted consequences for required data collection or wider applicability of our approach. We elaborate on the reasons for not retaining these features in Section IV-D.

Table IV gives a summary of the 36 features that we compute. We distinguish between six feature sets: for each feature set, we describe what it represents, which features it includes, how it is obtained, and how complete its coverage is. We indicate for each feature 1) whether it is binary or continuous, 2) whether our intuition is that higher or true values indicate a benign or malicious domain,[1] 3) which life cycle step from Section IV-A it covers, and 4) which insight from Section IV-B is illustrated.

For each domain, we know the start and end dates of their validity period, i.e. when their respective DGA would generate the domain. We also retrieve the date when a malware family started being active from DGArchive [71], where available.

a) Two *lexical* features capture the linguistic structure of the domain name. We compute the domain name length, as shorter domains tend to be more popular and expensive, and the ratio of digits in the domain name, as domains with more digits tend to be less readable. Both features discard the TLD.

b) Seven *popularity-based* features capture whether a domain hosts a website that appears to attract regular visitors. Three features use data obtained through the Wayback Machine API[2]: the number of unique pages captured on the domain, the time between the first capture of any page and the takedown, and the time between this first capture and the start of the AGD validity period.

[1]Note that this is only an intuition—our classifier can detect edge cases that provide contrary evidence.

[2]https://archive.org/help/wayback_api.php

TABLE IV. OVERVIEW OF THE FEATURES USED IN OUR CLASSIFIER. WE INDICATE WHICH OUTCOME (BENIGN OR MALICIOUS) A HIGHER OR TRUE VALUE DENOTES AND HOW THE FEATURE COVERS THE DOMAIN LIFE CYCLE AND INSIGHTS.

Set	#	Description	Type	Outcome	Life cycle step (Section IV-A)	Insight (Section IV-B)	Source
Lexical	1	Domain name length	Continuous	Malicious	L1. Domain choice	i1. Likelihood	[16]
	2	Digit ratio	Continuous	Malicious	L1. Domain choice	i1. Likelihood	[19]
Popularity	3	Number of pages found in Wayback Machine	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	4	Time between first entry in Wayback Machine and takedown	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	5	Time between first entry in Wayback Machine and start of malware validity period	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	6-9	Presence in Alexa, Majestic, Quantcast, and Umbrella top websites rankings	Binary	Benign	L5. Activity	i3. Popularity	[58]
CT	10	TLS certificate found in Certificate Transparency logs	Binary	Benign	L4. Infrastructure	i2. Investment	*New*
WHOIS	11	Time between WHOIS creation date and start of AGD validity period	Continuous	Benign	L2. Registration	i2. Investment	*New*
	12	Time between WHOIS creation date and start of malware family activity	Continuous	Benign	L2. Registration	i2. Investment	*New*
	13	Time between WHOIS creation data and takedown date	Continuous	Benign	L2. Registration	i2. Investment	[36]
	14	Time between WHOIS creation date and WHOIS expiration date	Continuous	Benign	L2. Registration	i2. Investment	[47]
	15	Renewal of domain seen in WHOIS data	Binary	Benign	L2. Registration	i2. Investment	[38]
	16	Domain uses privacy/proxy service	Binary	Malicious	L2. Registration	i2. Investment	*New*
	17	WHOIS registrant email is a temporary/throwaway email service	Binary	Malicious	L2. Registration	i2. Investment	*New*
	18	WHOIS registrant phone number is valid	Binary	Benign	L2. Registration	i2. Investment	*New*
Passive DNS	19	Number of passive DNS queries	Continuous	Benign	L5. Activity	i3. Popularity	[58]
	20	Time between first and last seen passive DNS query	Continuous	Benign	L5. Activity	i3. Popularity	[58]
	21	Time between first seen passive DNS query and takedown	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	22	Time between first seen passive DNS query and start of AGD validity period	Continuous	Benign	L5. Activity	i3. Popularity	*New*
	23-29	Presence of passive DNS query for resource record: `A`, `AAAA`, `CNAME`, `MX`, `NS`, `SOA`, `TXT`	Binary	Benign	L5. Activity	i3. Popularity	*New*
Active DNS	30	Time between first seen DNS record and takedown	Continuous	Benign	L3. DNS config.	i2. Investment	*New*
	31	Time between first seen DNS record and start of AGD validity period	Continuous	Benign	L3. DNS config.	i2. Investment	*New*
	32-36	Number of days DNS record was seen for resource records `A`, `AAAA`, `MX`, `NS`, `SOA`	Continuous	Benign	L3. DNS config.	i2. Investment	*New*

Four features capture whether the domain is present at any point in time in the Alexa[3], Majestic[4], Quantcast[5], and Umbrella[6] top websites rankings. These rankings serve as an approximation of popularity from different vantage points: web browser visits, incoming links, tracking script/ISP data, and DNS traffic, respectively. Although they can contain malicious domains and are susceptible to malicious manipulation [56], we assume that presence in these lists still serves as a reasonable indication of benign intent. We retrieve historical data from an archive of historical top websites rankings [76].

c) One *Certificate Transparency* feature captures whether Certificate Transparency logs contain a certificate that was valid on the date of the takedown, i.e. whether the owner had obtained a TLS certificate for the domain. The feature in this set uses data obtained through an API from Entrust[7], which tracks Google Certificate Transparency logs [63]. Certificate Transparency logs have the most complete coverage of issued TLS certificates [92]. Recent browser policies that enforce logging further increase uptake [75].

d) Eight *WHOIS* features capture the registration cycle of a domain as well as registrant details. We base four features on the time between the WHOIS creation date and the start of the AGD validity period, the start of malware family activity, the takedown date, and the WHOIS expiration date respectively. For an additional feature, we compute whether the domain has been renewed at least once by the latest registrant, i.e. we find at least two records with different expiration dates.

We capture the validity of registrant data in three features. We determine if the domain uses a privacy/proxy service (replacing real registrant data with generic data) by checking for keywords (e.g. "privacy", "proxy") in the WHOIS registrant records. While legitimate users may prefer to use such a service to hide personal information [51], malicious domains also tend to use these services [24]. We also determine whether the WHOIS registrant email is a disposable address: as the email account can no longer be accessed after some time, this indicates that the owner does not consider the domain to be important. We test non-default/non-proxy email addresses against a manually curated list of disposable domains[8]. Finally, we check whether the WHOIS registrant phone number is valid: malicious actors would not want any trace leading to their real identity and therefore resort to fake (e.g., automatically generated) contact information. We test the validity of phone numbers using an API from numverify[9].

WHOIS-based features are based on historical data generously provided to us by DomainTools[10]. To observe long-term and renewed registrations, we obtain historical records spanning their full data collection period. The data reflects a state before the introduction of the European General Data Protection Regulation, so it contains more domains with publicly available contact details. We elaborate on the continued availability of such details in Section VI-B.

e) Eleven *passive DNS* features capture both the period and frequency of DNS resolutions for a particular domain, providing a viewpoint on both domain age and popularity. We retrieve the number of passive DNS queries: when more queries (for any resource record) have been made for the domain, the domain appears to be more popular. We base three features on the time between the first seen passive DNS query and the last seen query, the takedown date, and the start of the AGD validity period respectively. Finally, we record the presence of at least one passive DNS query for resource records `A`, `AAAA`, `CNAME`, `MX`, `NS`, `SOA`, and `TXT`: more (requested) record types with a value indicate proper domain setup and usage.

[3]https://www.alexa.com/topsites
[4]https://majestic.com/reports/majestic-million
[5]https://www.quantcast.com/top-sites/
[6]https://umbrella-static.s3-us-west-1.amazonaws.com/index.html
[7]https://www.entrust.com/ct-search/
[8]https://github.com/ivolo/disposable-email-domains
[9]https://numverify.com/
[10]https://whois.domaintools.com/

The features in this set use passive DNS data generously provided to us by Farsight Security[11]. We retrieve aggregated data spanning the full data collection period (i.e., since 2010 [32]). For each resource record value seen, the aggregated data contains the number of queries and the timestamps when it was first and last seen.

f) Seven active DNS features capture the availability of DNS records for a particular domain. We base two features on the time between the first seen DNS record and the takedown date, and the start of the AGD validity period respectively. We also record the number of days any DNS record value was seen for resource records `A`, `AAAA`, `MX`, `NS`, and `SOA`.

The features in this set use active DNS data generously provided to us by the OpenINTEL[12] project [91]. We cap the data period at 333 days (i.e. starting from January 1 of the relevant year). While OpenINTEL collects data actively, it complies with our requirement that we do not contact domains ourselves. Moreover, data collection is not targeted at specific domains, yet sufficiently comprehensive to also capture most of the registered Avalanche domains as it covers full zone files.

D. Omitted features

Given our use case of proactive takedowns, we cannot consider features that try to detect ongoing malicious operations directly, as the maliciously registered domain does not yet necessarily exhibit such behavior at the time of the takedown: malicious actors can leave these domains dormant right until a DGA generates the domain and infected hosts start contacting the domain. This means for example that we do not verify whether a C&C server is running on the domain and do not check malware blacklists.

Approaches for detecting AGDs, especially per single domain, are often based on lexical features that seek to discover patterns unlikely to occur in "human-generated" domain names [77], [78]. However, all of our candidate domains have been generated by a DGA, which leads us to use only a limited set of lexical features to find the domains that are more likely to be potential collisions (short and few digits).

Detecting patterns from DNS logs [20] that indicate fast flux services [41], often used by command and control servers, is not applicable as the malicious domains would only start operating in fast flux during the validity period of the AGD.

Following our observation from Section III-B that bulk patterns do not apply for malware domains, we do not use approaches and features that rely on clustering domains [16] and batches of similar registrations [38], such as timing patterns or shared registrars.

The type of network could be an appropriate feature to take into account while the domain is active [20], with more trust in government or business networks hosting benign sites and domains in residential networks potentially being hosted by an infected machine. However, as a maliciously registered domain does not yet have to be actively malicious before the DGA generates the domain, its IP address can easily be set to a benign network (without the need for that network to actually host the domain) [62], thereby misleading our classifier.

Data collected through a crawl of candidate domains such as properties of the site content could indicate legitimately used domains [47]. However, following our stealth constraint from Section III-B and due to the need for historical data, we cannot do an active crawl of domains ourselves. We also cannot rely on existing third-party repositories of website crawls (e.g. the Internet Archive [2], Common Crawl [25] or Censys [30]): they do not provide historical data, do not crawl sufficiently regularly to capture recent data, do not have a consistent set of crawled domains and/or do not have sufficient domain coverage. Their data would therefore not be comprehensively representative of domain web content at the time of the takedown.

We do not include the malware family as a feature: as Avalanche provided domain registration as a service [3], we do not expect differences in behavior between the 21 supported malware families. Moreover, such a feature would go against our goal of capturing general differences in behavior between benign and malicious domains. We design the other features to represent distributions, for which the model can interpret the differences, whereas the malware family feature can only serve to refine the model for specific families. Finally, benign domains accidentally 'belong' to a certain malware family, so the feature is irrelevant in terms of registration behavior. We already capture relevant characteristics of the DGA in derived features such as the domain length that capture randomness in generated domains and therefore the likelihood of collisions.

We want to evaluate our approach as if it were deployed at the time of the takedown, so we do not use features for which we lack available historical data, as we would only be able to obtain the current state, which for malicious domains is post-takedown. They include the features that require active probing or data collection such as the website properties discussed earlier or the existence of search engine results for the domain, which could serve as an additional indicator of popularity. However, if they meet the applicable requirements and constraints, we can add such features in an actual takedown as we can then collect accurate data.

V. Analysis of machine learning-based classification

To evaluate to what extent machine-learning based approaches can reduce the effort of law enforcement to execute a takedown, we develop and evaluate a classifier that decides whether future DGA domains are likely to be benign or malicious. The goals of our analysis are threefold: we want to evaluate the raw performance of the classifier, but also gain insights into its decision-making process to finally thoroughly assess the benefits and limitations of automated approaches for domain classification. Moreover, given that not all data sources are equally easy to collect, we assess their impact on the correctness of our classification.

A. Experimental protocol

We first design an experimental protocol to determine the most appropriate machine learning-based solution and evaluate it in a way that is accurate and representative of real-world takedowns. Given the investigative setting and our intention to

[11] https://www.farsightsecurity.com/solutions/dnsdb/
[12] https://www.openintel.nl/

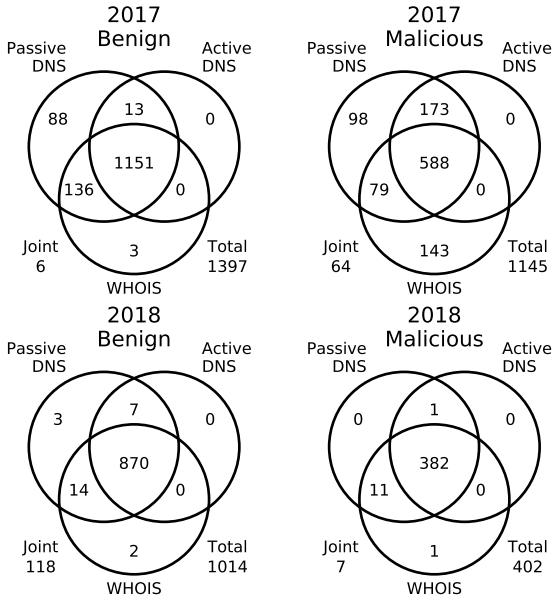


Fig. 2. Number of domains where certain data sets are available, after removing sinkholed domains, for the 2017 and 2018 iterations. We separately mark the remainder of domains where only the joint data set (comprising lexical, popularity-based, and Certificate Transparency features) is available.

thoroughly analyze the resulting model, we restrict our selection of machine learning algorithms to those that are sufficiently interpretable. Moreover, as we systematically develop high-level features that capture the full domain life cycle, we do not require automated feature engineering. Therefore, we would not benefit from a deep learning approach and only face drawbacks from its increased complexity, so we do not consider it further.

Before classifying benign and malicious domains, we discard domains that were already sinkholed by security organizations to study botnet behavior. These organizations can sinkhole the domains either because they detect that botnet hosts are already contacting the domain (whose validity period therefore starts before and extends beyond the takedown date), or because they generate the domains output by the DGA upfront. The sinkholed domains can be considered neither a benign collision, as they do not host real content and may even mimic the malware C&C server, nor a registration made with malicious intent, as they will not communicate with actual malware. This means that they would confuse our model, and should be removed upfront by preprocessing the data. We detect sinkholed domains by matching DNS and WHOIS records with those of the sinkhole providers collected in SinkDB [10], by Alowaisheq et al. [12], and by Stampar et al. [87], [88]. Table II summarizes the distribution of domains across classes.

We execute our protocol with four machine learning algorithms: decision tree, gradient boosted tree, random forest, and support vector machine. We split data sets in a training and test set according to the considered iterations. When training and testing on the same iteration, we split the ground truth according to a 10-fold cross validation procedure. Otherwise, we construct the training and test sets from the separate iteration ground truths as applicable. We perform all model training and analysis using `scikit-learn` [67]. We elaborate on the different steps of this protocol in Appendix A.

We run our experimental protocol for all domains of the 2017, 2018 and 2019 takedown iterations. We only evaluate performance with the manually labeled ground truth that we obtained from law enforcement for the 2017 and 2018 iterations (Section III-C). In 2019, our model was used in the real-world classification effort, so a performance evaluation would be biased since we contributed to the ground truth.

As we want to measure the performance of our approach as if it were deployed at the time of the takedown operation, we use historical data that reflects the state of the domains as of each takedown, i.e. November 30 of each year. Data for the malicious domains collected after the takedown would refer to sinkholing and domain transfer infrastructure, making it a signal for maliciousness that would heavily bias our classifier.

As shown in Figure 2, we cannot obtain all data sets for all domains: this is because the third-party source could not collect relevant data (e.g. no WHOIS record is available or the domain was never seen at passive DNS sensors). In order to still generate a prediction for all domains, we develop an *ensemble model*. We train a model for each combination of available feature sets, where a domain is included in the training set if at least those data sets are available. To classify a domain, we use the output of the model of the domain's available data sets.

B. Results

Given that we are the first to analyze the specific issue of preemptively deciding whether DGA domains are actually malicious or accidentally benign for a real-world takedown (which brings about certain constraints), we are not able to compare our performance results with previous work. Instead, we go beyond reporting basic metrics and critically examine how its performance translates into a real-world reduction in effort, whether our solution correctly captures differences between benign and malicious domains, and how much it depends on the availability of different data sets.

a) Model performance: Appendix B lists the relative performance of the four machine learning algorithms that we evaluate: we conclude that a gradient boosted tree classifier yields the best performance while still being sufficiently interpretable. We therefore analyze only its results.

We first train a *base* ensemble model, varying the training and test sets over the 2017 and 2018 iterations. From the performance metrics in Table V, we can see that concept drift [95] occurs: performance drops when deploying our model across iterations instead of within. This suggests that over time, patterns that distinguish benign and malicious actors emerge or change, and these are therefore not captured by a model trained on only a single iteration.

We therefore develop an *extended* ensemble model, where we combine ground truth from a previous iteration with manual, *a priori* classifications of a subset of domains in the target iteration. This enables us to improve model performance by capturing the novel patterns in the new iteration, while still reducing manual effort overall.

We evaluate this extended model trained on all of the 2017 and part of the 2018 ground truth and tested on the remaining 2018 domains. Based on Figure 3, we empirically set the proportion of the 2018 ground truth that is (randomly) selected

TABLE V. PERFORMANCE METRICS FOR THE BASE ENSEMBLE MODEL, VARYING THE TRAINING AND TEST SET OVER THE 2017 AND 2018 ITERATIONS.

Training \ Test	Accuracy		F_1 score		Precision		Recall	
	2017	2018	2017	2018	2017	2018	2017	2018
2017	93.4%	84.3%	92.6%	73.4%	92.6%	70.8%	92.7%	76.1%
2018	76.1%	96.3%	70.9%	93.5%	78.6%	92.7%	64.6%	94.3%

TABLE VI. PERFORMANCE METRICS FOR MODELS TRAINED ON THE 2017 AND (FOR THE EXTENDED MODEL) 15% OF THE 2018 ITERATION.

Ensemble model		Accuracy	F_1 score	Precision	Recall	FNR	FPR	Effort reduction
Base		84.3%	73.4%	70.8%	76.1%	23.9%	12.4%	100.0%
Extended	a priori	86.4%	78.6%	70.5%	88.6%	2.3%	2.0%	100.0%
Base	a posteriori	97.3%	95.3%	94.2%	96.5%	3.5%	2.4%	70.3%
Extended	a priori + a posteriori	97.6%	95.8%	94.3%	97.4%	2.6%	2.3%	66.2%

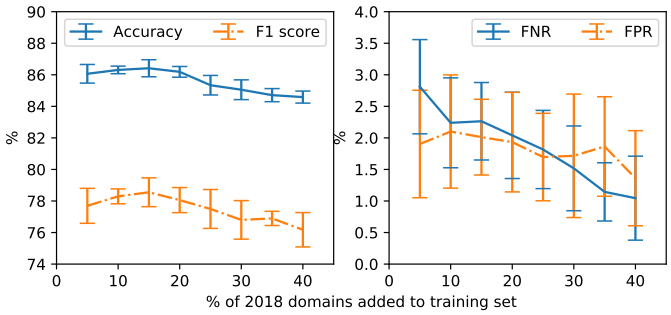


Fig. 3. Performance metrics (mean and standard deviation) for the extended a priori ensemble model, trained on the 2017 and a varying part of the 2018 ground truth.

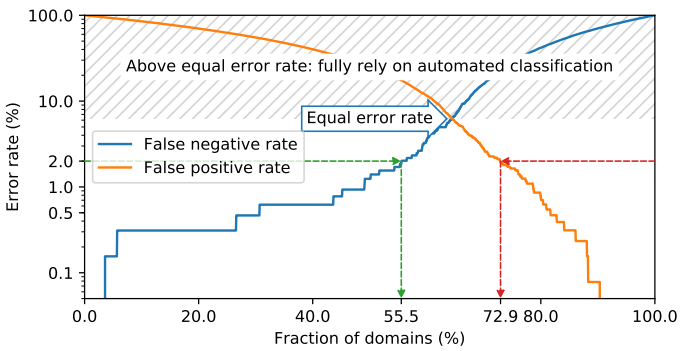


Fig. 4. FNR and FPR as a function of the fraction of domains with a score below a certain value. By choosing the maximum error rate, we determine the fraction of domains that can be automatically classified.

to be manually classified and added to the training set at 15%, as it represents the best trade-off between improved performance and limited additional effort. We repeat this random selection ten times and report average results. Table VI shows that this extended a priori ensemble model improves on the base model.

However, some misclassifications still occur in this extended a priori model. The gradient boosted tree model outputs a score that reflects its confidence in its prediction. We can leverage these scores to develop a directed semi-automated approach: uncertain domains are manually investigated in more detail *a posteriori*. We examine how effective this approach is in further improving performance while still reducing investigative effort.

We explain this approach using the extended model for domains where all data sets are available, which allows us to simplify and visually support our explanation, but then apply it to the extended ensemble model. Figure 4 shows the false negative and positive rates as a function of the fraction of domains with a score below a certain value. By choosing a target maximum FNR and FPR, we can determine the lower and upper bounds on the maliciousness score; these bounds are determined based on the training set, so they do not necessarily reflect the exact actual error rates on the test set. Domains with scores within these bounds have to be verified manually, while domains with a lower and higher score are automatically classified as benign and malicious, respectively.

For the extended model on domains with all data sets available as represented in Figure 4, when setting a 2% error tolerance, 55.5% of domains have a maliciousness score below the lower bound set by 2% FPR (i.e. are benign), while $(100\% - 72.9\%) = 27.1\%$ of domains exceed the upper bound set by 2% FNR (i.e. are malicious). $55.5\% + 27.1\% = 82.6\%$ of domains therefore no longer need to be manually inspected. Only $72.9\% - 55.5\% = 17.4\%$ of domains still require further manual investigation.

When we apply this a posteriori approach to the extended ensemble model evaluated on all domains from the 2017 and part of the 2018 iteration (by choosing appropriate bounds for each component model), we obtain an accuracy of 97.6%; overall, the performance metrics in Table VI indicate a very high performance. The effective FNR and FPR are 2.6% and 2.3%, comparable to the target error rate of 2%.

Overall, this approach reduces manual effort by 66.2%, accounting for the 15% of domains manually classified a priori. When the error tolerance is 1% and 0.5%, the fraction of automatically classified domains is 52.5% and 35.7% respectively. The score thresholds become very strict when very low error tolerances must be maintained, reducing the fraction of domains that can be automatically classified. The comparable effort reduction for an ensemble model trained on the 2017 and 2018 and tested on the 2019 iteration and a 2% error tolerance amounts to 76.9%, again achieving a significant reduction in manual effort.

b) Feature analysis: By using gradient boosted trees, we can measure how important individual features are to the overall performance. As we want to make an accurate assessment for

TABLE VII. IMPORTANCE SCORES OF THE TOP 10 FEATURES IN THE FULL FEATURE SET FOR THE EXTENDED A PRIORI ENSEMBLE MODEL.

#	Set	Feature	Score
14	WHOIS	Time between WHOIS creation and expiration date	0.230
13	WHOIS	Time between WHOIS creation and takedown date	0.219
21	Passive DNS	Time between first passive DNS query and takedown	0.057
20	Passive DNS	Time between first and last seen passive DNS query	0.049
11	WHOIS	Time between WHOIS creation date and AGD validity	0.041
15	WHOIS	Renewal of domain seen in WHOIS data (Unknown)	0.040
34	Active DNS	Days DNS record was seen for resource record MX	0.040
15	WHOIS	Renewal of domain seen in WHOIS data (False)	0.037
31	Active DNS	Time between first seen DNS record and AGD validity	0.029
3	Popularity	Number of pages found in Wayback Machine	0.028

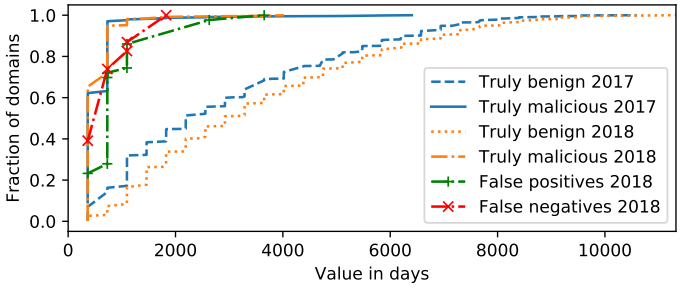


Fig. 5. Cumulative distribution function of the values of benign, malicious, false positive, and false negative domains for the time between WHOIS creation and expiration date.

the full feature set, we calculate importance scores for the extended model on domains where all data sets are available.

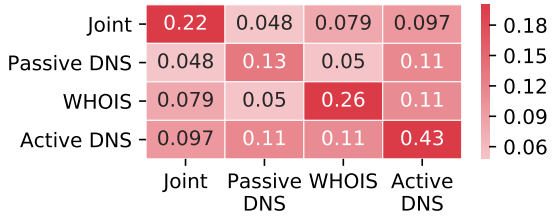
We show the ten most important features in Table VII and find that they primarily capture the age and activity period of a domain. When malware creators want to evade our classifier, they would primarily want to influence these features. Figure 5 shows how the distributions of values for the most impactful feature (time between WHOIS creation and expiration date) are clearly distinct for benign and malicious domains. Misclassified benign domains (false positives) actually show a 'malicious' character, i.e. they are young; the malicious domains in our test set (from 2018) are never old, so other (but less expressive) features impact whether they are classified correctly.

Consistent with our second insight from Section IV-B, time-based features are costly and difficult to evade: attackers have to register a domain name for a longer period of time, which translates into a higher monetary cost, and register it earlier, which is hard to achieve retroactively. In an extreme case, the domain name would have to be registered before the malware family becomes active.

c) Data set comparison: We assess the impact of the availability of each data source on our performance starting from the extended a priori ensemble model, after which we retrain models with one feature set omitted each time. We join lexical, popularity-based, and Certificate Transparency features into a joint feature set, as they are the easiest to acquire and are always available, which leaves us with four feature sets: joint, WHOIS, passive DNS, and active DNS.

Figure 6 illustrates the performance of the models where one data set is discarded. We observe that missing WHOIS data has the most severe impact, significantly harming performance. Discarding the joint data set may actually improve performance, as its non-time-based features may lack sufficiently distinctive

TABLE VIII. AVERAGE COVARIANCE BETWEEN FEATURES OF ONE SET, FOR THE DOMAINS FROM THE 2017 AND 2018 ITERATIONS.



patterns, but it remains necessary for domains that lack any other data set (but these are likely candidates for manual verification).

Missing passive or active DNS data has a less pronounced effect. We find some degree of redundancy between passive and active DNS data, as their time-based features in particular represent similar concepts and are therefore intuitively dependent. We confirm this effect with the covariance between feature sets shown in Table VIII: passive and active DNS data are relatively highly correlated with each other.

This effect means that passive and active DNS (as well as WHOIS) data all capture important and hard-to-evade time-based patterns, but that one missing data set can be substituted by the others without a significant loss in performance. This becomes important when considering that data sets such as WHOIS that lead to better performance may come with a significant cost to acquire. In Section VI-B, we elaborate on the implications of our findings on future takedown operations.

d) Conclusion: We find that an approach combining primarily automated classification and targeted manual investigation across multiple iterations achieves the best compromise of high accuracy and low manual effort, with less than 3% mistakes. This reduces investigative effort by up to 76.9%, depending on the tolerated error rate, freeing up time to focus on those domains that are the hardest to classify.

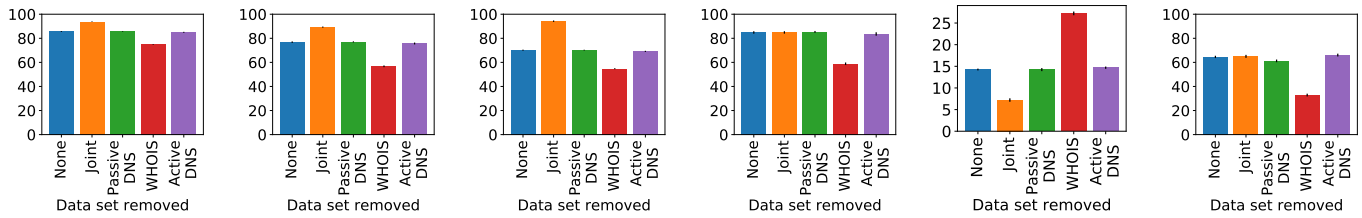
Our analysis of features and data sets shows that time-based features are the most important ones, which at the same time increases the cost and difficulty of evading our classifier. However, our performance depends on data sources with a high cost of acquisition, in particular WHOIS data. We continue our discussion of these aspects in the next section.

VI. DISCUSSION

In this section, we elaborate on the factors that may influence the applicability of our approach to future takedowns. We first explain how a high cost and effort for attackers complicates the evasion of our classifier and may therefore discourage malicious actors. We then highlight how recent developments in the availability of data sets may have a negative impact on the performance of our approach.

A. Evasion

Previous work [38], [60] pointed out that attackers may develop bypasses to mislead a classifier like ours and therefore evade detection and subsequent takedown of their malicious domains, especially as we cannot rely on detecting the malicious activity that would be required for the correct functioning of the botnet. We discuss potential evasion strategies and how difficult



(a) Accuracy. (b) F_1 score. (c) Precision. (d) Recall. (e) Equal error rate. (f) Reduction in effort (for 2% error rate).

Fig. 6. Performance metrics (mean and standard deviation, in percent) of extended a priori ensemble models where one data set is omitted.

they are for malicious actors to deploy. This proactive analysis allows for anticipating changes in attacker behavior, developing additional features that are even harder to circumvent and implementing infrastructural measures that complicate evasion.

Features that leverage the properties of the DGA itself, such as lexical features, can be evaded by redesigning DGAs. While it is feasible to carefully engineer DGAs to be more resilient against detection [85], such a DGA should generate domains that appear very similar to benign domains (e.g., only short domains). This yields a higher risk of collisions and fewer domains available for registration, endangering uninterrupted control of the botnet.

Popularity-based features require setting up a website for discovery by web crawlers, and generating traffic, or at least the appearance thereof. Website popularity rankings can easily be manipulated at scale [56], allowing attackers to insert their domains and appear as benign. If malicious actors can have a presence within the networks where passive DNS data is collected, they could also insert DNS traffic that makes the domain appear regularly visited. Given that the attackers control their infected machines, the botnet itself could be leveraged for this purpose. However, as the traffic of infected machines can be monitored, these queries can be detected, revealing those domains that the malicious actors have registered upfront. Finally, the presence of certain DNS resource records can be forged by inserting fake records, but as some records require values of a specific format, their validity could be verified, as maintaining valid records requires more effort.

Given recent efforts to increase the ubiquity of TLS encryption by making free and automated TLS certificates available [11], malicious actors can relatively easily obtain them for malicious domains and therefore appear in Certificate Transparency logs. However, such a process still requires additional effort that is not strictly necessary for the correct operation of the C&C server. While the choice to obtain a paid certificate indicates a willingness to invest in the domain (and therefore suggests benignness), the use of a free certificate does not necessarily imply maliciousness.

Features that consider the age of a domain can be thwarted by registering malicious domains (long) before they become valid. However, it requires prolonged registrations and the corresponding payment of registration fees, which runs counter to minimizing the cost of the malicious campaign. Moreover, the longer a domain with malicious intent has been registered, whether active or dormant, the more susceptible it is to being blacklisted/taken down or to the attackers being identified.

Acquiring and managing domains may incur a significant (manual) effort. If the process is automated, certain registration patterns can emerge that make it easier to identify the maliciously registered domains [86], [93]. Malicious actors might attempt to compromise existing or reuse expired domains to exploit the (residual) trust in these domains [57] (for example their age). However, it would require even more effort, as they would need to find eligible domains, attempt to compromise them or monitor their expiration status to take them over at the right time, and finally deploy the malicious operation. As domains are randomly generated by a DGA and often have a short validity, the likelihood of success is low.

To circumvent features that use WHOIS registrant records, malicious actors could insert forged yet realistically-looking data. However, if these records are automatically generated, detection becomes feasible and accurate [86], [93]. Manual effort in creating fake records quickly becomes infeasible given the need to keep registering domains as they become (in)valid.

In summary, while the publication of features allows for an attacker to develop techniques to evade them, many of these would go against the goal of malware operators to set up these domains with low effort and at low cost. Moreover, if the attacker behavior would significantly shift, other evasion countermeasures and detection strategies remain available, although they might require increased effort and involvement by relevant stakeholders. Finally, we find time-based features to be the most important ones: they are particularly costly and hard to evade.

B. *Availability of data sets*

Our features come from different data sources that each present their own issues in terms of acquisition, affecting not only law enforcement but also adversaries seeking to evade the model. Moreover, our evaluation of the importance of different data sources for correctly classifying domains shows that the data sets that contribute the most to our model's performance have a significant cost in terms of money and effort.

WHOIS data in particular provides the highest accuracy, but obtaining it may be challenging. From a technical standpoint, WHOIS data is not machine-readable nor has a standard format [27], so it requires (sometimes manual) parsing. Moreover, access is rate limited [59].

Public availability of WHOIS data is also affected by privacy concerns [74] as well as strict limitations on the collection and dissemination of personal data due to privacy regulations. This triggered ICANN to adopt the "Temporary Specification for

gTLD Registration Data", which allows generic TLD registries to redact personal data in WHOIS records, while having the intent to provide vetted partners such as law enforcement agencies with privileged access [45]. As a result of the European General Data Protection Regulation, European country-code TLD registries have also started to withhold personal data [29]. Security researchers have voiced concerns that the unavailability of such data to them could significantly hamper efforts to identify and track malicious actors [34], [70].

Passive DNS data collection may also have privacy implications [52], and requires sufficient storage and processing resources. Active DNS data collection has similar storage and resource needs, especially to ensure that records are updated sufficiently frequently. The coverage of both data sets also depends on cooperation of third parties: passive DNS requires access to recursive resolvers ideally deployed all over the world, and active DNS collection often relies on zone files that must then be shared by registries. Although law enforcement may gain more extensive access, they may be more limited in terms of resources, and delays in procedures to obtain data may hamper swift action. Conversely, commercial providers that can deploy more extensive resources may not be able to access more sensitive information. Finally, from a cost perspective, these commercial providers may charge significant amounts, especially for historical data.

We see that our approach becomes less effective if certain data sets would be unavailable, and our discussion shows that comprehensive coverage of data sets comes at great cost. However, we can still achieve reasonable performance even with missing data, and we see that data sets are partially correlated. The continued availability of these data sets is therefore important to counter future malicious operations, but not to such an extent that their absence would be disrupting the effectiveness of takedowns.

VII. Related work

a) Classifiers for detecting malicious domains: Numerous works have addressed the problem of designing classifiers to distinguish benign from malicious web pages and domains. Ma et al. [60] classified malicious URLs based on lexical and host-based features, comparing multiple feature sets and classifiers. Felegyhazi et al. [33] designed a classifier seeded with known malicious domains that uses DNS and WHOIS data. Antonakakis et al. [15] proposed Notos, which outputs a reputation score based on the determination of the reputation of domain clusters obtained from network properties, DNS data, and the ground truth on benign and malicious domains. Bilge et al. [19], [20] proposed Exposure, which uses DNS-based and domain name features to detect domains contacted by infected machines within passive DNS traffic. Frosch et al. [36] proposed Predentifier, which combines passive DNS, WHOIS, and geolocation data to detect botnet command and control servers. Hao et al. [38] proposed PREDATOR, a classifier for malicious domains based on features available at the time of registration and the identification of batch registrations. Spooren et al. [86] developed Premadoma, a model to detect malicious domains at the time of registration, leveraging features based on infrastructural reputation and registrant similarity, and discussed the challenges and tactics for deploying the model in an operational setting. Machlica et al. [61] created a model that uses two levels of classifiers to improve detecting malicious domains using lexical and traffic-based features. Kidmose et al. [48] and Zhauniarovich et al. [102] surveyed approaches to detecting malicious domains from (enriched) DNS data.

b) Classifiers for detecting algorithmically generated domains: Earlier work in detecting algorithmically generated domains (AGDs) identified clusters of likely candidates. Yadav et al. [99], [100] evaluated several statistical measures for classifying groups of domains as algorithmically generated or not based on character distributions within the domain names and the IP addresses to which they resolve. Yadav and Reddy [98] applied similar statistical measures on successful and failed domain resolutions. Antonakakis et al. [16] proposed Pleiades, which clusters non-existent domains based on character distributions within the domain names and on the querying hosts, using the strategy on DNS traffic from large ISPs to discover six DGAs that were unknown at that time. Krishnan et al. [53] detected hosts in a botnet by analyzing patterns in DNS queries for non-existent AGDs through sequential hypothesis testing. Mowbray et al. [64] detected hosts that query domains with an unusual length distribution, deriving 19 DGAs of which nine were previously unknown.

Later work moved towards detecting AGDs per single domain name. Schiavone et al. [77] proposed Phoenix, which uses linguistic features to detect potential AGDs, afterwards using linguistic, IP-based and DNS-based features to cluster domains and extract properties of the DGAs that generated them. Abbink and Doerr [9] and Pereira et al. [68] highlighted how most classifiers focus on detecting the randomness in AGDs and are therefore not able to correctly classify dictionary-based DGAs, and proposed new methods for detecting such DGAs. Multiple deep learning-based approaches have since been proposed [82]. Spooren et al. [85] found one such deep learning model by Woodbridge et al. [96] to outperform the human-engineered features of the model by Schüppen et al. [78].

c) Takedowns of botnet infrastructures: Previous coordinated takedowns of botnet infrastructures have been studied to evaluate their effectiveness over time in preventing further abuse. Nadji et al. [65] presented rza, a tool that uses a passive DNS database to analyze and improve the effectiveness of botnet takedowns. They evaluated the tool for three malware families and found mixed long-term impact of takedown operations. Asghari et al. [17] analyzed the institutional factors that influenced the cleanup effort of the Conficker worm, finding that cleanup was slow and that large-scale national initiatives did not have a visible impact. Shirazi [80] surveyed and taxonomized 19 botnet takedown initiatives from 2008 to 2014. Plohmann et al. [71] analyzed the structure of DGAs for 43 malware families and variants, and analyzed registrations of their AGDs, finding domains missed in takedowns, families for which few domains were sinkholed, and slowness in seizing AGDs registered by malicious actors. Alowaisheq et al. [12] studied the life cycle of takedown operations across sinkholes and registrars based on passive DNS and WHOIS data, finding several flaws that would allow malicious actors to regain control of some sinkholed domains. Hutchings et al. [42] provided insights into the effectiveness of takedown efforts by interviewing key actors, finding that law enforcement faces more challenges than commercial enterprises in effectively carrying out takedown operations.

VIII. Conclusion

Taking down the domains that compromised machines use to communicate with command and control servers is an effective measure to disrupt botnets such as Avalanche. However, law enforcement must take care not to affect any legitimate domains that happen to collide with algorithmically generated domains. For Avalanche, prosecutors manually conducted this classification process, requiring large amounts of time and effort as well as allowing for human error.

We therefore develop an automated approach for classifying benign and malicious registered DGA domains, within the constraints of the real-world takedown context that make previous approaches inapplicable: we cannot rely on bulk patterns, detecting ongoing malware activity or actively connecting to domains. We propose a hybrid model that balances automation with manual classification to achieve a higher performance as well as vastly reduce investigator effort. We develop and evaluate our approach to represent the Avalanche takedown most truthfully, such that our results and findings reflect the utility of automated domain classifiers in a real-world takedown scenario, such as for our contribution to the 2019 iteration.

Given the increasing number and size of cybercrime operations, automated tools can assist law enforcement investigators in avoiding any harmful impact of their operation, especially on uninvolved legitimate parties. These tools will allow them to stay one step ahead of malicious actors and impair their activities with the goal of shielding end users from any harm.

Acknowledgment

We thank the reviewers for their valuable and constructive feedback, as well as the Security Analytics SIG at DistriNet, the Drakkar group at LIG, and Paul Vixie. We thank our partners for providing access to the Avalanche ground truth data: Benedict Addis of RoLR, Sascha Alexander Jopen and his team at Fraunhofer FKIE, and the law enforcement agencies involved. We thank Farsight Security for providing access to the DNSDB data as well as the DNSDB data contributors; DomainTools for providing historical WHOIS data; the OpenINTEL team, in particular Roland van Rijswijk-Deij, for their help in obtaining the OpenINTEL data; Roman Huessy at abuse.ch for the SinkDB data; and Daniel Plohmann for access to DGArchive.

This research is partially funded by the Research Fund KU Leuven. Victor Le Pochat holds a PhD Fellowship of the Research Foundation - Flanders (FWO). This work was partially supported by SIDN, the .NL Registry and AFNIC, the .FR Registry under the COMAR project. The research leading to these results was made possible by OpenINTEL (https://www.openintel.nl/), a joint project of SURFnet, the University of Twente, SIDN and NLnet Labs.

References

[1] Avalanche stats by subregion. The Shadowserver Foundation. [Online]. Available: https://avalanche.shadowserver.org/stats/

[2] (2013, Sep.) Wayback Machine APIs. The Internet Archive. [Online]. Available: https://archive.org/help/wayback_api.php

[3] "Declaration of special agent Aaron O. Francis in support of application for an emergency temporary restraining order and order to show cause re preliminary injunction," in *United States of America v. "flux" a/k/a "ffhost", and "flux2" a/k/a "ffhost2"*. District Court, Western District of Pennsylvania, Nov. 2016. [Online]. Available: https://www.justice.gov/opa/page/file/915231/download

[4] "Preliminary injunction," in *United States of America v. "flux" a/k/a "ffhost", and "flux2" a/k/a "ffhost2"*. District Court, Western District of Pennsylvania, Dec. 2016. [Online]. Available: https://www.justice.gov/opa/page/file/917581/download

[5] (2017, Dec.) Avalanche year two, this time with Andromeda. The Shadowserver Foundation. [Online]. Available: http://blog.shadowserver.org/news/avalanche-year-two-this-time-with-andromeda/

[6] "Operation Avalanche: A closer look," Eurojust, EU publication QP-01-17-801-EN-N, Apr. 2017. [Online]. Available: http://www.eurojust.europa.eu/doclibrary/Eurojust-framework/Casework/Operation%20Avalanche%20-%20A%20closer%20look%20(April%202017)/2017-04_Avalanche-Case_EN.pdf

[7] (2018, Dec.) Avalanche 1,2,3.... The Shadowserver Foundation. [Online]. Available: http://blog.shadowserver.org/news/avalanche-123/

[8] G. Aaron and R. Rasmussen, "Global phishing survey: Trends and domain name use in 2H2009," Anti-Phishing Working Group, APWG Industry Advisory, May 2010. [Online]. Available: https://docs.apwg.org/reports/APWG_GlobalPhishingSurvey_2H2009.pdf

[9] J. Abbink and C. Doerr, "Popularity-based detection of domain generation algorithms," in *12th International Conference on Availability, Reliability and Security*, ser. ARES '17, 2017, pp. 79:1–79:8.

[10] abuse.ch. (2019) SinkDB. [Online]. Available: https://sinkdb.abuse.ch/

[11] M. Aertsen, M. Korczyński, G. C. M. Moura, S. Tajalizadehkhoob, and J. van den Berg, "No domain left behind: Is Let's Encrypt democratizing encryption?" in *2017 Applied Networking Research Workshop*, ser. ANRW '17, 2017, pp. 48–54.

[12] E. Alowaisheq, P. Wang, S. Alrwais, X. Liao, X. Wang, T. Alowaisheq, X. Mi, S. Tang, and B. Liu, "Cracking the wall of confinement: Understanding and analyzing malicious domain take-downs," in *26th Annual Network and Distributed System Security Symposium*, ser. NDSS '19, 2019.

[13] S. Alrwais, X. Liao, X. Mi, P. Wang, X. Wang, F. Qian, R. Beyah, and D. McCoy, "Under the shadow of sunshine: Understanding and detecting bulletproof hosting on legitimate service provider networks," in *2017 IEEE Symposium on Security and Privacy*, ser. SP '17, 2017, pp. 805–823.

[14] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai botnet," in *26th USENIX Security Symposium*, ser. USENIX Security '17, 2017, pp. 1093–1110.

[15] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, "Building a dynamic reputation system for DNS," in *19th USENIX Conference on Security*, ser. USENIX Security '10, 2010, pp. 273–289.

[16] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From throw-away traffic to bots: Detecting the rise of DGA-based malware," in *21st USENIX Security Symposium*, ser. USENIX Security '12, 2012, pp. 491–506.

[17] H. Asghari, M. Ciere, and M. J. van Eeten, "Post-mortem of a zombie: Conficker cleanup after six years," in *24th USENIX Security Symposium*, ser. USENIX Security '15, 2015, pp. 1–16.

[18] T. Barabosch, A. Wichmann, F. Leder, and E. Gerhards-Padilla, "Automatic extraction of domain name generation algorithms from current malware," in *IST-111/RSY-026 Symposium on Information Assurance and Cyber Defence*. NATO Science & Technology Organization, 2012.

[19] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "EXPOSURE: Finding malicious domains using passive DNS analysis," in *18th Annual Network and Distributed System Security Symposium*, ser. NDSS '11, 2011.

[20] L. Bilge, S. Sen, D. Balzarotti, E. Kirda, and C. Kruegel, "Exposure: A passive DNS analysis service to detect and report malicious domains,"

ACM Transactions on Information and System Security, vol. 16, no. 4, pp. 14:1–14:28, Apr. 2014.

[21] boker *et al.* (2018, Dec.) Domain seized. [Online]. Available: https://www.namepros.com/threads/domain-seized.1116091/

[22] O. Cetin, C. Gañán, L. Altena, T. Kasama, D. Inoue, K. Tamiya, Y. Tie, K. Yoshioka, and M. van Eeten, "Cleaning up the internet of evil things: Real-world evidence on ISP and consumer efforts to remove Mirai," in *26th Annual Network and Distributed System Security Symposium*, ser. NDSS '19, 2019.

[23] Y. T. Chua, S. Parkin, M. Edwards, D. Oliveira, S. Schiffner, G. Tyson, and A. Hutchings, "Identifying unintended harms of cybersecurity countermeasures," in *2019 APWG Symposium on Electronic Crime Research*, ser. eCrime '19, 2019.

[24] R. Clayton and T. Mansfield, "A study of Whois privacy and proxy service abuse," in *13th Annual Workshop on the Economics of Information Security*, ser. WEIS '14, 2014.

[25] Common Crawl Foundation. Common Crawl. [Online]. Available: https://commoncrawl.org/

[26] R. R. Curtin, A. B. Gardner, S. Grzonkowski, A. Kleymenov, and A. Mosquera, "Detecting DGA domains with recurrent neural networks and side information," in *14th International Conference on Availability, Reliability and Security*, ser. ARES '19, 2019, pp. 20:1–20:10.

[27] L. Daigle, "WHOIS protocol specification," Internet Requests for Comments, RFC Editor, RFC 3912, Sep. 2004.

[28] S. Danziger, J. Levav, and L. Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 17, pp. 6889–6892, 2011.

[29] DENIC. (2018, May) DENIC putting extensive changes into force for .DE Whois lookup service by 25 May 2018. [Online]. Available: https://www.denic.de/en/whats-new/press-releases/article/denic-putting-extensive-changes-into-force-for-de-whois-lookup-service-as-of-25-may-2018/

[30] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman, "A search engine backed by Internet-wide scanning," in *22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15, 2015, pp. 542–553.

[31] "'Avalanche' network dismantled in international cyber operation," Europol, Dec. 2016. [Online]. Available: https://www.europol.europa.eu/newsroom/news/%E2%80%98avalanche%E2%80%99-network-dismantled-in-international-cyber-operation

[32] Farsight Security. Passive DNS historical internet database: Farsight DNSDB. Farsight Security. [Online]. Available: https://www.farsightsecurity.com/solutions/dnsdb/

[33] M. Felegyhazi, C. Kreibich, and V. Paxson, "On the potential of proactive domain blacklisting," in *3rd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More*, ser. LEET '10, 2010.

[34] A. J. Ferrante, "The impact of GDPR on WHOIS: Implications for businesses facing cybercrime," *Cyber Security: A Peer-Reviewed Journal*, vol. 2, no. 2, pp. 143–148, 2018.

[35] Freenom. (2017) Free and paid domains. [Online]. Available: https://www.freenom.com/en/freeandpaiddomains.html

[36] T. Frosch, M. Kührer, and T. Holz, "Predentifier: Detecting botnet C&C domains from passive DNS data," in *Advances in IT Early Warning*, M. Zeilinger, P. Schoo, and E. Hermann, Eds. Fraunhofer Verlag, Feb. 2013, pp. 78–90. [Online]. Available: http://publica.fraunhofer.de/documents/N-227985.html

[37] D. Gomes, J. Miranda, and M. Costa, "A survey on web archiving initiatives," in *International Conference on Theory and Practice of Digital Libraries*, ser. TPDL '11, 2011, pp. 408–420.

[38] S. Hao, A. Kantchelian, B. Miller, V. Paxson, and N. Feamster, "PREDATOR: Proactive recognition and elimination of domain abuse at time-of-registration," in *2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, 2016, pp. 1568–1579.

[39] S. Hao, M. Thomas, V. Paxson, N. Feamster, C. Kreibich, C. Grier, and S. Hollenbeck, "Understanding the domain registration behavior of spammers," in *2013 Internet Measurement Conference*, ser. IMC '13, 2013, pp. 63–76.

[40] M. Heinemeyer. (2018, Mar.) How malware abused sixt.com and breitling.com for covert command & control communication. Darktrace. [Online]. Available: https://www.darktrace.com/en/blog/how-malware-abused-sixt-com-and-breitling-com-for-covert-command-control-communication/

[41] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling, "Measuring and detecting fast-flux service networks," in *15th Annual Network and Distributed System Security Symposium*, ser. NDSS '08, 2008.

[42] A. Hutchings, R. Clayton, and R. Anderson, "Taking down websites to prevent crime," in *2016 APWG Symposium on Electronic Crime Research*, ser. eCrime '16, 2016.

[43] IBM Security. IBM X-Force Exchange. frequently asked questions. [Online]. Available: https://exchange.xforce.ibmcloud.com/faq

[44] Internet Corporation for Assigned Names and Numbers. (2012, Feb.) How long does a registration last? Can it be renewed? [Online]. Available: https://www.icann.org/resources/pages/faqs-84-2012-02-25-en#7

[45] ——. (2018, May) Temporary specification for gTLD registration data. Internet Corporation for Assigned Names and Numbers. [Online]. Available: https://www.icann.org/resources/pages/gtld-registration-data-specs-en

[46] L. Invernizzi, K. Thomas, A. Kapravelos, O. Comanescu, J.-M. Picod, and E. Bursztein, "Cloak of visibility: Detecting when machines browse a different web," in *2016 IEEE Symposium on Security and Privacy*, ser. SP '16, 2016, pp. 743–758.

[47] N. Kheir, F. Tran, P. Caron, and N. Deschamps, "Mentor: Positive DNS reputation to skim-off benign domains in botnet C&C blacklists," in *29th IFIP International Information Security and Privacy Conference*, ser. SEC '14, 2014, pp. 1–14.

[48] E. Kidmose, E. Lansing, S. Brandbyge, and J. M. Pedersen, "Detection of malicious and abusive domain names," in *2018 1st International Conference on Data Intelligence and Security*, ser. ICDIS '18, Apr. 2018, pp. 49–56.

[49] K. Kopel, "Operation seizing our sites: How the federal government is taking domain names without prior notice," *Berkeley Technology Law Journal*, vol. 28, no. 4, pp. 859–900, 2013.

[50] M. Korczyński, S. Tajalizadehkhoob, A. Noroozian, M. Wullink, C. Hesselman, and M. van Eeten, "Reputation metrics design to improve intermediary incentives for security of TLDs," in *2017 IEEE European Symposium on Security and Privacy*, ser. EuroS&P '17, 2017, pp. 579–594.

[51] M. Korczyński, M. Wullink, S. Tajalizadehkhoob, G. C. M. Moura, A. Noroozian, D. Bagley, and C. Hesselman, "Cybercrime after the sunrise: A statistical analysis of DNS abuse in new gTLDs," in *13th ACM Asia Conference on Computer and Communications Security*, ser. ASIACCS '18, 2018, pp. 609–623.

[52] A. Kountouras, P. Kintis, C. Lever, Y. Chen, Y. Nadji, D. Dagon, M. Antonakakis, and R. Joffe, "Enabling network security through active DNS datasets," in *Research in Attacks, Intrusions, and Defenses*, ser. RAID '16, 2016, pp. 188–208.

[53] S. Krishnan, T. Taylor, F. Monrose, and J. McHugh, "Crossing the threshold: Detecting network malfeasance via sequential hypothesis testing," in *43rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, ser. DSN '13, 2013.

[54] M. Kührer, C. Rossow, and T. Holz, "Paint it black: Evaluating the effectiveness of malware blacklists," in *17th International Symposium on Research in Attacks, Intrusions and Defenses*, ser. RAID '14, 2014, pp. 1–21.

[55] B. Laurie, A. Langley, and E. Kasper, "Certificate Transparency," Internet Requests for Comments, RFC Editor, RFC 6962, June 2013.

[56] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," in *26th Annual Network and Distributed System Security Symposium*, ser. NDSS '19, 2019.

[57] C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis, "Domain-Z: 28 registrations later measuring the exploitation of residual trust in domains," in *2016 IEEE Symposium on Security and Privacy*, ser. SP '16, 2016, pp. 691–706.

[58] P. Lison and V. Mavroeidis, "Neural reputation models learned from passive DNS data," in *2017 IEEE International Conference on Big Data*, ser. Big Data '17, 2017, pp. 3662–3671.

[59] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul, "Who is .com?: Learning to parse WHOIS records," in *2015 Internet Measurement Conference*, ser. IMC '15, 2015, pp. 369–380.

[60] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: Learning to detect malicious web sites from suspicious URLs," in *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09, 2009, pp. 1245–1254.

[61] L. Machlica, K. Bartos, and M. Sofka, "Learning detectors of malicious web requests for intrusion detection in network traffic," Feb. 2017, arXiv:1702.02530.

[62] L. B. Metcalf, D. Ruef, and J. M. Spring, "Open-source measurement of fast-flux networks while considering domain-name parking," in *2017 Learning from Authoritative Security Experiment Results Workshop*, ser. LASER '17, 2017, pp. 13–24.

[63] B. Morton. (2016, Oct.) Protect your domain with CT search. [Online]. Available: https://www.entrustdatacard.com/blog/2016/october/protect-your-domain-with-ct-search

[64] M. Mowbray and J. Hagen, "Finding domain-generation algorithms by looking at length distribution," in *2014 IEEE International Symposium on Software Reliability Engineering Workshops*, 2014, pp. 395–400.

[65] Y. Nadji, M. Antonakakis, R. Perdisci, D. Dagon, and W. Lee, "Beheading hydras: Performing effective botnet takedowns," in *2013 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '13, 2013, pp. 121–132.

[66] S. Pal. (2019, Dec.) Sinkholed. [Online]. Available: https://susam.in/blog/sinkholed/

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[68] M. Pereira, S. Coleman, B. Yu, M. De Cock, and A. C. A. Nascimento, "Dictionary extraction and detection of algorithmically generated domain names in passive DNS traffic," in *21st International Symposium on Research in Attacks, Intrusions, and Defenses*, ser. RAID '18, 2018, pp. 295–314.

[69] N. Petit, "Artificial intelligence and automated law enforcement: A review paper," *SSRN Electronic Journal*, 2018. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3145133

[70] D. Piscitello. (2018, Oct.) ICANN GDPR and WHOIS users survey. a joint survey by the anti-phishing working group (APWG) and the messaging, malware and mobile anti-abuse working group (M^3AAWG). [Online]. Available: https://www.m3aawg.org/sites/default/files/m3aawg-apwg-whois-user-survey-report-2018-10.pdf

[71] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla, "A comprehensive measurement study of domain generating malware," in *25th USENIX Security Symposium*, ser. USENIX Security '16, 2016, pp. 263–278.

[72] M. Z. Rafique, T. Van Goethem, W. Joosen, C. Huygens, and N. Nikiforakis, "It's free for a reason: Exploring the ecosystem of free live streaming services," in *23rd Annual Network and Distributed System Security Symposium*, ser. NDSS '16, 2016.

[73] Rapid7. Project Sonar. [Online]. Available: https://www.rapid7.com/research/project-sonar/

[74] S. Rodota, "Opinion 2/2003 on the application of the data protection principles to the Whois directories," Article 29 Data Protection Working Party, Jun. 2003. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2003/wp76_en.pdf

[75] Q. Scheitle, O. Gasser, T. Nolte, J. Amann, L. Brent, G. Carle, R. Holz, T. C. Schmidt, and M. Wählisch, "The rise of certificate transparency and its implications on the Internet ecosystem," in *2018 Internet Measurement Conference*, ser. IMC '18, 2018, pp. 343–349.

[76] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of Internet top lists," in *2018 Internet Measurement Conference*, ser. IMC '18, 2018, pp. 478–493.

[77] S. Schiavoni, F. Maggi, L. Cavallaro, and S. Zanero, "Phoenix: DGA-based botnet tracking and intelligence," in *11th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, ser. DIMVA '14, 2014, pp. 192–211.

[78] S. Schüppen, D. Teubert, P. Herrmann, and U. Meyer, "FANCI : Feature-based automated NXDomain classification and intelligence," in *27th USENIX Security Symposium*, ser. USENIX Security '18, 2018, pp. 1165–1181.

[79] D. Schwarz. (2015, Apr.) Bedep's DGA: Trading foreign exchange for malware domains. Arbor Networks. [Online]. Available: https://web.archive.org/web/20160114122355/https://asert.arbornetworks.com/bedeps-dga-trading-foreign-exchange-for-malware-domains/

[80] R. Shirazi, "Botnet takedown initiatives: A taxonomy and performance model," *Technology Innovation Management Review*, vol. 5, no. 1, pp. 15–20, Jan. 2015.

[81] S. Sinha, M. Bailey, and F. Jahanian, "Shades of grey: On the effectiveness of reputation-based "blacklists"," in *3rd International Conference on Malicious and Unwanted Software*, ser. MALWARE '08, 2008, pp. 57–64.

[82] R. Sivaguru, C. Choudhary, B. Yu, V. Tymchenko, A. Nascimento, and M. De Cock, "An evaluation of DGA classifiers," in *2018 IEEE International Conference on Big Data*, ser. Big Data '18, 2018, pp. 5058–5067.

[83] K. Soska and N. Christin, "Automatically detecting vulnerable websites before they turn malicious," in *23rd USENIX Security Symposium*, ser. USENIX Security '14, 2014, pp. 625–640.

[84] A. Sperotto, O. van der Toorn, and R. van Rijswijk-Deij, "TIDE: Threat identification using active DNS measurements," in *Proceedings of the SIGCOMM Posters and Demos*, ser. SIGCOMM Posters and Demos '17, 2017, pp. 65–67.

[85] J. Spooren, D. Preuveneers, L. Desmet, P. Janssen, and W. Joosen, "Detection of algorithmically generated domain names used by botnets: A dual arms race," in *34th ACM/SIGAPP Symposium on Applied Computing*, ser. SAC '19, 2019, pp. 1916–1923.

[86] J. Spooren, T. Vissers, P. Janssen, W. Joosen, and L. Desmet, "Premadoma: An operational solution for DNS registries to prevent malicious domain registrations," in *35th Annual Computer Security Applications Conference*, ser. ACSAC '19, 2019, pp. 557–567.

[87] M. Stampar. (2018, Oct.) Email addresses used in WHOIS registrations of sinkholed malicious/malware domains. [Online]. Available: https://gist.github.com/stamparm/9726d93fd0048aee6c54ec88a8e85bfc

[88] M. Stampar *et al.* (2019) maltrail: Malicious traffic detection system. [Online]. Available: https://github.com/stamparm/maltrail

[89] M. Stevanovic, J. M. Pedersen, A. D'Alconzo, S. Ruehrup, and A. Berger, "On the ground truth problem of malicious DNS traffic analysis," *Computers & Security*, vol. 55, pp. 142–158, 2015.

[90] J. Tierney, "Do you suffer from decision fatigue?" Aug. 2011. [Online]. Available: https://www.nytimes.com/2011/08/21/magazine/do-you-suffer-from-decision-fatigue.html

[91] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras, "A high-performance, scalable infrastructure for large-scale active DNS measurements," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 6, pp. 1877–1888, June 2016.

[92] B. VanderSloot, J. Amann, M. Bernhard, Z. Durumeric, M. Bailey, and J. A. Halderman, "Towards a complete view of the certificate ecosystem," in *2016 Internet Measurement Conference*, ser. IMC '16, 2016, pp. 543–549.

[93] T. Vissers, J. Spooren, P. Agten, D. Jumpertz, P. Janssen, M. V. Wesemael, F. Piessens, W. Joosen, and L. Desmet, "Exploring the ecosystem of malicious domain registrations in the .eu TLD," in *Proceedings of the 20th International Symposium on Research in Attacks, Intrusions, and Defenses*, ser. RAID '17, 2017, pp. 472–493.

[94] R. Wainwright and F. J. Cilluffo, "Responding to cybercrime at scale: Operation Avalanche - a case study," Europol; Center for Cyber and Homeland Security, The George Washington University, Issue Brief 2017-03, Mar. 2017. [Online]. Available: https://cchs.gwu.edu/sites/g/files/zaxdzs2371/f/Responding%20to%20Cybercrime%20at%20Scale%20FINAL.pdf

[95] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Machine Learning*, vol. 23, no. 1, pp. 69–101, Apr. 1996.

[96] J. Woodbridge, H. S. Anderson, A. Ahuja, and D. Grant, "Predict-

ing Domain Generation Algorithms with Long Short-Term Memory Networks," Nov. 2016, arXiv:1611.00791.

[97] W. Xu, K. Sanders, and Y. Zhang, "We know it before you do: Predicting malicious domains," in *Virus Bulletin Conference*, Sep. 2014, pp. 73–77.

[98] S. Yadav and A. L. N. Reddy, "Winning with DNS failures: Strategies for faster botnet detection," in *7th International ICST Conference on Security and Privacy in Communication Networks*, ser. SecureComm '11, 2011, pp. 446–459.

[99] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan, "Detecting algorithmically generated malicious domain names," in *10th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '10, 2010, pp. 48–61.

[100] ——, "Detecting algorithmically generated domain-flux attacks with DNS traffic analysis," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1663–1677, Oct. 2012.

[101] B. Z. H. Zhao, M. Ikram, H. J. Asghar, M. A. Kaafar, A. Chaabane, and K. Thilakarathna, "A decade of mal-activity reporting: A retrospective analysis of Internet malicious activity blacklists," in *14th ACM Asia Conference on Computer and Communications Security*, ser. ASIACCS '19, 2019, pp. 193–205.

[102] Y. Zhauniarovich, I. Khalil, T. Yu, and M. Dacier, "A survey on malicious domains detection through DNS data analysis," *ACM Computing Surveys*, vol. 51, no. 4, pp. 67:1–67:36, Jul. 2018.

APPENDIX A
MACHINE LEARNING PROTOCOL

Machine learning algorithms are trained on a training set Tr and evaluated on a test set Te. As explained in Section V, if we need to train and test on the same iteration, we split using a k-fold cross validation procedure: the data is split in k folds, with every fold being used once as the test set, while we use the $k-1$ others for training, and finally, we average results over k experiments. We set k to 10. The advantage of using cross validation is that we can reduce bias in the composition of the selected training and test set, even with a relatively small data set.

Most ML algorithms have different hyperparameters to tune. Tuning on the test set would lead to highly biased results. Therefore, we have to split the training set Tr into a set for training Tr' and another one for validation V. We again use a 10-fold cross validation procedure. We treat and calculate the upper and lower bounds for the extended a posteriori model as hyperparameters.

We evaluate the following performance metrics over the test set:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

$$precision = \frac{tp}{tp + fp} \tag{2}$$

$$recall = \frac{tp}{tp + fn} \tag{3}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{4}$$

where tp, tn, fp, fn stand for the number of true positives, true negatives, false positives and false negatives, respectively. Malicious domains are considered positive, benign domains are negative. Precision represents the fraction of samples identified as malicious that are actually malicious, while recall represents the fraction of malicious samples that were correctly identified. The F_1 score summarizes these two metrics, and is a superior metric compared to accuracy when dealing with unbalanced datasets, therefore we optimize for it.

Due to incompleteness of our data sets (e.g., WHOIS records not containing a parseable phone number), certain domains have missing feature values. We impute them (i.e., substituted them with plausible values to avoid bias) as follows (the feature numbers correspond to those defined in Section IV-C):

- No Wayback Machine data: feature values (3-5) are set to zero as no data means that the Wayback Machine has not found any page on the domain, suggesting unpopularity.
- No WHOIS timestamps: feature values (11-14) are set to the mean, as no data implies that data could not be parsed or retrieved, not that the data does not exist (e.g., all domains have a registration date). By using the mean, we do not attach any statistical meaning to the absence of data and do not skew the distribution.
- Less than two WHOIS records: the renewal feature (15) gets a third value that indicates that only one historical WHOIS record was available (preventing a comparison of expiration dates).
- No WHOIS registrant records: features that rely on an address, an email address, or a phone number (16-18) get a third value that indicates that we do not have a value for the corresponding field.
- No passive or active DNS data: continuous feature values (19-22, 30-36) are set to zero and binary feature values (23-29) to false as no data means that DNS records for the domain were never queried, suggesting unpopularity.

APPENDIX B
EVALUATION OF MACHINE LEARNING ALGORITHMS

Table IX presents the performance metrics of the machine learning algorithms that we evaluate in Section V-B, for a base ensemble model trained and tested on the initial 2017 iteration. The results show that gradient boosted trees consistently outperform the other ML algorithms.

TABLE IX. PERFORMANCE METRICS OF THE EVALUATED MACHINE LEARNING ALGORITHMS.

Metric	Decision Tree	Gradient Boosted Tree	Random Forest	Support Vector Machine
Accuracy	88.6%	93.4%	92.8%	86.4%
Recall	86.6%	92.7%	92.6%	77.9%
Precision	87.8%	92.6%	91.5%	90.6%
F_1 score	87.2%	92.6%	92.0%	83.8%