Mentor: Positive DNS Reputation to Skim-Off Benign Domains in Botnet C&C Blacklists

Nizar Kheir, Frédéric Tran, Pierre Caron, Nicolas Deschamps

Mentor: Positive DNS Reputation to Skim-off Benign Domains in Botnet C&C Blacklists

Nizar Kheir, Frédéric Tran, Pierre Caron, and Nicolas Deschamps

Orange Labs, Issy-Les-Moulineaux, France
{name.surname}@orange.com

Abstract. The Domain Name System (DNS) is an essential infrastructure service on the internet. It provides a worldwide mapping between easily memorizable domain names and numerical IP addresses. Today, legitimate users and malicious applications use this service to locate content on the internet. Yet botnets increasingly rely on DNS to connect to their command and control servers. A widespread approach to detect bot infections inside corporate networks is to inspect DNS traffic using domain C&C blacklists. These are built using a wide range of techniques including passive DNS analysis, malware sandboxing and web content filtering. Using DNS to detect botnets is still an error-prone process; and current blacklist generation algorithms often add innocuous domains that lead to a large number of false positives during detection.
This paper presents a new system called Mentor. It implements a scalable, *positive DNS reputation* system that automatically removes benign entries within a blacklist of botnet C&C domains. Mentor embeds a crawler system that collects statistical features about a suspect domain name, including both web content and DNS properties. It applies supervised learning to a labeled set of known benign and malicious domain names, using its features set in order to build a DNS pruning model. It further processes domain blacklists using this model in order to skim-off benign domains and keep only true malicious domains for detection. We tested our system against a wide set of public botnet blacklists. Experimental results prove the ability of this system to efficiently detect and remove benign domain names with a very low false positives rate.

1 Introduction

The Domain Name System (DNS) constitutes a core infrastructure component of the internet. It provides a global hierarchical service that associates internet domains with their corresponding IP addresses [15]. Today, internet users access the DNS system to locate and retrieve content such as web servers, hosting and mailing services. Because of its global reach, DNS is nowadays being used to share knowledge about malware threats, including infected websites and domain callbacks [7]. As soon as malware infects a terminal, it establishes a Command and Control (C&C) channel with an attacker in order to download updates, retrieve commands and steal data. Yet malware implements multiple mechanisms

to locate its C&C server, and DNS still constitutes the most widespread technique being used today, including hard-coded domains, DGA [3] and domain flux [10]. A common approach to fight against malware is to use blacklists of botnet C&C domains [2, 6]. It observes traffic at system egress points and drops connections towards known malicious domain names. When malware can no longer connect to its C&C server, its effect would be neutralized because it would no longer be accessible to the remote attacker.

Domain C&C blacklists are currently being generated using a wide range of techniques such as passive DNS traffic analysis, malware sandboxing and web content filtering. Yet the wide use of domain blacklists for botnet detection is still confronted with the large amount of false positives included in these blacklists. Hence, security administrators are still reluctant in using domain blacklists as a way to automatically drop botnet C&C communications. In fact these are not reliable enough to be used as a proactive security solution [9], and so they are mostly being used for passive detection and alerting. Yet malware implements multiple obfuscation mechanisms that make difficult the correct identification of its main command and control channels, as follows.

Firstly, extraction of C&C domains during dynamic sandbox analysis is error-prone since malware triggers multiple network connections in addition to its main C&C activity. The conficker.C malware provides a typical example where it randomly selects a C&C domain out of 50 thousand possible domain names created daily for this purpose. The use of similar techniques by malware clearly makes difficult the correct extraction and maintenance of domain blacklists.

On the other hand, negative DNS reputation provides an alternative approach to sandbox analysis as it does not require collecting and executing malware samples. It aims at observing DNS traffic and collecting features that characterize a botnet signaling activity. For instance, the Notos system in [2] uses evidence-based features such as the number of malware that connected to a given domain name in order to measure the reliability of this domain. The Exposure system [6] also defines features that differentiate malicious and benign domains based on botnet C&C artifacts such as short domain TTL and abrupt changes in DNS requests towards a given domain. Unfortunately, modern botnets can easily avert negative DNS reputation systems using techniques such as random delays and noise injection within their main C&C communications. They also increasingly implement hybrid botnet topologies that distribute commands among a larger set of master C&C bots, thus reducing the coverage of DNS features during detection [12, 13]. In fact negative DNS reputation observes only artifacts associated with a known botnet signaling activity and so it is efficient only against known botnet C&C topologies. It cannot easily adapt to new botnet obfuscation techniques, thus limiting its coverage and increasing risks of false positives [20].

This paper addresses the limitations of current negative DNS reputation systems through the proposal of a new system called Mentor, that implements positive DNS reputation to separate malicious and benign domain names. Mentor searches for *artifacts that prove the innocuous nature of a domain name*. It acts

as a watchdog that processes domain blacklists generated by negative DNS reputation systems. It implements a crawler system that collects artifacts and builds a comprehensive set of features for every single domain in these blacklists. Mentor collects elements that characterize the legitimacy of a given domain name, including the popularity, cross-references, external links and the data hosted on this domain. It further applies machine learning techniques, using this set of features, in order to identify benign domain names and remove these from the initial domain blacklists, keeping only true C&C domains for botnet detection and prevention. Indeed Mentor uses a web crawler that actively connects to remote domains in order to build its features set, as opposed to negative DNS reputation that only uses passive traffic analysis.

To summarize, this paper makes the following two contributions:

- It proposes a comprehensive set of features that characterize the benign nature of a given domain. Our features complete current DNS reputation systems that use artifacts describing only malicious access to a given domain.
- It uses machine learning in order to build an automated, positive DNS reputation system that actively processes domain blacklists and eliminates false positives, with no need for human intervention.

Experimental results prove the ability of Mentor to efficiently identify and discard benign domain names from within domain blacklists, while also satisfying a very low false positives rate. This paper is structured as follows. Section 2 describes related work. Section 3 presents the architecture and workflow of Mentor. Section 4 provides experiments that we used in order to evaluate our system. Section 5 discusses the limitations of our system and finally section 6 concludes.

2 Related work

DNS is a core service that is widely used to locate content on the internet through association of domain names such as 'www.domainname.com' with routable IP addresses [15]. IP addresses are grouped within autonomous systems (AS) and so they are tightly coupled to a geographical location. On the other hand, domain names are grouped within administrative domains that can be associated with any IP address, regardless of the geographical position of the corresponding resource [17]. DNS is thus widely used by threat actors on the internet to associate IP addresses with domain names that are further used by infected nodes to locate their command servers [19].

Detection and extraction of domain callbacks first consisted of dynamically executing malware and observing its network activity [11, 16, 12]. After infecting a terminal, malware connects to a command and control server in order to get updates or retrieve commands. By observing network activity of malware in a dynamic analysis environment, we may pinpoint its main C&C channels and add these to domain blacklists. Hence, malware has been constantly developing new techniques to avoid being correctly analyzed, including the use of domain generation algorithms (DGA) [3] and detecting execution inside virtual

analysis environments [4]. Although several techniques have been proposed in the literature to thwart malware obfuscation mechanisms [22, 21], C&C domains discovered using these techniques are limited only to known malware samples that were correctly executed in sandbox environments. Yet they cannot identify C&C domains for unknown botnets and for malware that can efficiently avert detection during dynamic analysis.

Another trend of research consists of passively observing network activity and using machine learning to detect botnet C&C traffic [2, 6, 5, 13]. While certain techniques such as [5] and [13] observe only netflow data, others are mainly focused on DNS traffic and use negative reputation to detect malicious C&C domains [2, 6]. For example, authors in [2] build a dynamic DNS reputation system that uses both network and zone features of a domain. They make the assumption that the malicious use of DNS has unique characteristics and can be separated from benign, professionally provisioned DNS services. Therefore, they passively observe DNS queries and build models of known benign and malicious domains. These models are used to compute a reputation score for a newly observed domain, and which indicates whether this domain is indeed malicious or benign. The main drawback for this approach is that it needs a long enough history for a given domain name in order to assign a correct reputation score. It is inaccurate against frequently changing C&C domains, such as for hybrid botnet topologies that use multiple master C&C nodes to distribute commands.

Authors in [6] propose an alternative approach that applies machine learning to a set of 15 DNS features in order to identify malicious C&C domains. This approach builds a learning set of known benign and malicious domain names that it uses in order to train a DNS classifier. This classifier passively monitors real-time DNS traffic and identifies malware domains that do not appear in existing blacklists. Features in [6] are grouped into four categories, including time-based features, response features, TTL features and syntactical domain name features. Those features characterize anomalies in the way a given domain name is being requested, including abrupt changes in DNS queries towards this domain.

The system that we propose in this paper does not replace negative DNS reputation presented in [2] and [6]. Indeed, Mentor completes negative DNS reputation and it mainly searches for evidence about the benign nature of a domain, as opposed to [2] and [6] that search for evidence about the malicious nature of the same domain. In fact this paper proposes a positive DNS reputation system that processes blacklists of suspicious domains in order to remove false positives and keep only true malicious domains for botnet detection.

3 System description

The Mentor system includes a training phase where it builds a detection model using a training set of known malicious and benign domains. It further applies this model during detection to unqualified C&C blacklists in order to remove benign false positives and keep only confirmed malicious domains for detection. As in figure 1, the training phase implements a crawler system that builds a

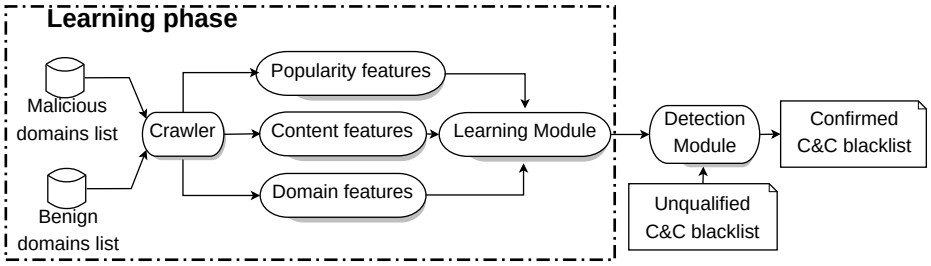


Fig. 1: Architecture and workflow of MENTOR

comprehensive set of features using the initial ground truth dataset. Features extracted by the crawler are further used as input to a supervised learning system. It implements machine learning techniques in order to build a classifier model that is further used during detection. This section provides the details of our features and describes the process that we use to build our detection model.

3.1 Features selection

As opposed to negative DNS reputation, our system finds evidence about the innocuous nature of a given domain. To determine the features that are indicative of a benign domain, we observed and studied the domain and content features of multiple known benign and malicious C&C domains, and that we obtained using our approach further presented in section 4.1. Following our analysis, we identified several distinctive features that we classify into three main categories, including popularity, content and domain-based features, as shown in figure 1. This section describes our set of features that we summarize in table 1, and explains our view about why they are indicative of benign domains.

Domain features describe empirical time-based observations about a domain name extracted from the public *whois* database. We consider that benign professional domain names are usually not changed and they are expected to remain accessible during the lifetime of their associated business. However, C&C domains are less static in nature. In order to enhance their resilience against detection and takedown, botnet herders frequently modify their callback domains using techniques such as DGA and domain flux. In fact, using the same C&C domain name during an extended period of time adds single nodes of failure in the botnet architecture. Therefore, botnet herders usually register short-lived domain names, along with short TTL values in order to frequently switch for other C&C domains, adding another level of complexity for botnet takedown. Hence, static registration information for professional domains show time-to-live values that are usually longer than other malicious C&C domains. Mentor accounts on this observation in order to introduce 4 domain features. They characterize such time-based differences between professional benign and malicious C&C domains.

Category	Id	Feature description
Domain features	1	Time since domain was first registered
	2	Time since domain was first created
	3	Time since domain was last changed
	4	Remaining time before domain expires
Content features	5	Ratio of text content w.r.t overall website content
	6	Number of entries in the site map of the website
	7	Number of entries in the robot file of a website
	8	Number of HTML descriptors metadata
	9	Number of HTML keywords metadata
	10	Number of HTML descriptors in the website title
	11	Number of displayable images hosted by the website
	12	Number of CSS style sheets for a website
Popularity features	13	Number of outbound links towards social networks
	14	Total number of outbound links
	15	Google page rank of the domain
	16	Number of inbound links from social networks

Table 1: Features set implemented by Mentor

They include the elapsed time since the domain was first *registered*, the elapsed time since the domain was first *created*, the elapsed time since the domain was last *changed*, and the remaining time before the domain *expires*.

Content features characterize differences in the structure and content of websites for both benign and malicious C&C domains. Benign professional domains usually seek higher rankings by the internet search engines such as *Google*, *Yahoo!* and *Bing*. They share rich HTML content, add metadata descriptors, and optimize the structure of their websites, which are all important elements that are used for domain indexing on the internet. On the other hand, C&C domains are used by malware to establish automated control paths between attackers and their remote infected bots. They are usually supplied by the attacker either statically using callback domains that are hard-encoded in malware payloads, or dynamically using domain flux techniques. Hence, malicious C&C domains do not seek good internet rankings. They usually share less human readable content and less metadata descriptors. Mentor capitalizes on these observations by introducing content features that separate benign and malicious C&C domains.

Mentor describes the content of a website using a set of 8 features, all being related to qualitative metrics used for domain indexing on the internet. We consider a domain to be aimed for benign usage when its content is likely to provide a higher indexing by internet search engines. Hence, the first feature provides the website text ratio. It characterizes the amount of human readable content hosted by a domain. It is evaluated as the ratio of textual content with respect to the overall content of a website. The second and third features respectively provide the size of the site map and robot files associated with a website. They

are evaluated as the number of entries in these files, and they characterize the structure of this website as seen by robots on the internet. These files determine how a website is displayed by search engines, which is an important property of professional benign domains. The fourth and fifth features provide the number of HTML descriptors and keywords metadata, and that determine the character strings that are most relevant for content indexing on this website. They describe the main area of interest which is being addressed by the website content. For example, a website that is aimed for health insurance would define keywords mostly related with the healthcare system. The sixth feature determines the number of HTML meta descriptors that are also represented in the website title. It provides an indication of whether the website title is randomly generated or if it was defined in compliance with the website content. Last of all, the seventh and eighth features respectively indicate the number of displayable images and css style sheets. They characterize the human friendly aspect and the way content is displayed by a website.

We admit that our features do not provide an exhaustive description of a website content. Nonetheless, they still provide enough evidence about the structure and content of a website. More importantly, they determine whether a website content is rather user friendly and eases human interaction, or if it is more likely to be addressed towards automated robots.

Popularity features describe a domain's popularity, including inbound and outbound links which characterize the user-friendly aspect of a website. As opposed to benign professional domains, malicious C&C domains are mostly aimed at sharing commands with remote infected terminals. They have characteristics that are different from other professional domains which are aimed at sharing content with benign users. For example, professional domains share human-readable content that can be appreciated or commented on social networks. Hence, they can be referenced by, yet they include outbound links towards social networks such as linkedin, twitter and facebook. Besides, the overall internet popularity also provides indicators about the professional or malicious nature of a domain. In fact, professional domains may share business partnerships, sponsors or media articles, which are described with inbound and outbound links towards external domains, and which also increases a domain's popularity. Therefore, we characterize the popularity of a domain using 4 elementary features. They include the number of outbound links towards social networks, the total number of outbound links, the google pagerank as an indicator of the domain's popularity as well as its total inbound links, and the number of inbound links from social networks that we obtain using the moonsy[1] application.

3.2 Detection model

The use of machine learning for botnet detection constitutes a real challenge as it implements statistical features that can be often bypassed by botnet herders.

[1] `http://www.moonsy.com`

In fact, modern botnets implement obfuscation mechanisms that make difficult to differentiate benign and malicious C&C activity using only network features. Hence, this paper offers a new set of features that leverages artifacts such as popularity, content descriptors, and timeline of a given domain. Indeed it is difficult for an attacker to setup malicious C&C domains that can bypass our features. Attackers need a longer time and effort to build a large enough set of popular domains including rich enough content descriptors in order to avoid detection, while also providing multiple redundant domains for botnet command and control.

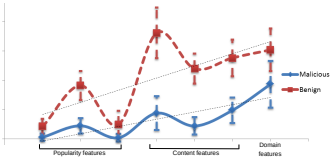
In order to build our domain classification system, we tested multiple supervised learning algorithms, including selective bayesian classifiers [14], SVM, J48 and C4.5 decision trees [8, 18]. First of all, decision trees offer a way to express structures in data. They provide a classic way to represent information from a machine learning algorithm. Besides, SVM provides an extension to nonlinear models that is based on the statistical learning theory. On the other hand, Bayes models provide a probabilistic classifier based on the Bayes Theorem. In the context of this paper, we evaluated the detection rates, including false positives and negatives, that we obtained by applying each of these learning algorithms, through application to our labeled ground truth dataset. We obtained a higher accuracy using the Bayesian classifier, and therefore we use this algorithm to build our domain classification model.

4 Experiments

This section provides the details of our experiments, including the dataset that we used in order to build and evaluate our system. First, we build a ground truth dataset of malicious and benign domain names that we process, using our crawler system, in order to train our classifier model. Then we evaluate the contribution of our features towards detection, and we adopt a cross-validation process in order to evaluate the accuracy of our system. Last of all, we tested Mentor against real public botnet C&C blacklists in order to evaluate the performance of our system for different malware families, as well as its ability to characterize unknown domain names by the time of building our system.

4.1 Ground truth dataset

Mentor applies machine learning to a training set of malicious and benign domain names in order to build a classifier model. The quality of our classifier strongly depends on the coverage of the initial training set and the accuracy of the ground truth labels associated with this training set. In fact Mentor actively connects to the domain names in the training set and builds detection features on the fly before these can be used as input to the training model. Hence, we need a valid blacklist of botnet C&C domains that are all active by the time of building our model. Yet we want to ensure that our blacklist only includes malicious domains and is clear of misclassified benign domain names.



(a) Mentor detection features

Dataset	Training	Evaluation	TP	FP
Benign	400	100	99.02%	0.98%
Malicious	745	185	95.32%	4.68%
Overall	1145	285	97.17%	2.83%

(b) Cross-validation results

Fig. 2: Evaluation of Mentor against the ground truth dataset

Malicious training set: In order to build our ground truth malicious dataset, we applied a voting system including Google safe browsing and two publically accessible blacklists: '`malwaredomains.com`' and '`malwaredomainlist.com`'. We consider that a given domain is more likely to be malicious when it belongs to more than a single blacklist. We obtained a resulting list of 1.849 malicious domains that were matching both blacklists and that were accessible by the time of building our model. Then we discarded domain names in this blacklist that were identified as safe by the Google safe browsing API. This API identified 919 domains out of the initial 1.849 domains as being safe. We thus keep the remaining 930 confirmed malicious domains as input to our classifier model. In fact we applied a conservative filter in order to ensure that our malicious training set does not include safe domains before we build our detection model. Sure we cannot rule out the possibility of few domains being misclassified in our initial training set. However, these would be limited when compared to the confirmed malicious C&C domains and so they would have little impact on our classifier. We further evaluate in this section the accuracy of our classifier, including false positives, by testing against a wide set of malicious and benign domain names.

Benign training set: We build our training set of benign domain names using the top domains list in Alexa [1]. The Alexa web site provides the list of most popular domains on the internet. Yet we cannot be sure that all top domains on the list provided by Alexa are indeed benign domains since malicious domains may shortly appear in this list. Therefore, we cross-correlated the Alexa top domains list during a period of one week in order to discard as many suspect domains as possible prior to building our classifier model. In fact we consider that malicious domains may indeed appear in the top domains list, but these would be rapidly detected and so they will be soon removed from this list. According to Alexa, traffic ranks are updated on the website daily, therefore we daily extracted the top domains list from the alexa website during one week of observation. Then we kept only the top 500 domains that were constantly present on this list during the 7 days of observation. Although we cannot formally validate our list of benign domains, we believe that they have a strong evidence of being benign professional domain names since they constantly remained popular on the internet during one week. Our ground truth dataset thus included a list of 930 confirmed malicious and 500 confirmed benign domains that we used to train our classifier.

4.2 Domain classification model

We process the training set of malicious and benign domains, using the crawler module, in order to build detection features for our classifier. The crawler took 56 minutes to process the entire list of 1.430 domain names, using a desktop machine with dual core processor and 2Gb of memory. Figure 2a illustrates the distribution of our training set, including benign and malicious domains, against different categories of features provided by our system.

As in figure 2a, content features provide the best detection accuracy with a clear separation of malicious and benign domains. Benign domains in our training set clearly contain rich content, including domain descriptors and metadata that do not exist in malicious domains. For example, most of our malicious C&C domains do not include human friendly `HTTP` tags such as keywords and rich content descriptors that indeed exist in most benign domains in our training set.

Domain features, including information collected from the *whois* database, provide lower detection accuracy compared to content features. As shown in figure 2a, most of our malicious C&C domains still have lower TTL values than benign domains. However, TTL values for malicious C&C domains have a large standard deviation, which leads to overlap between our set of malicious and benign domains, thus reducing the overall detection accuracy of those features.

Last of all, popularity features also provide an overall good detection accuracy, and so they clearly separate malicious and benign domains. According to our learning set, C&C domains have much less incoming and outgoing links compared to other benign domains. On the other hand, and as shown in figure 2a, features describing the popularity on social networks provide a lower separation between malicious and benign domains in our training set. Hence, they would have a lower detection accuracy compared to other detection features.

In the remaining of this section, we build our classifier system using the detection features described in table 1, and we evaluate the accuracy of our model, including the true hit rate and false positives rate.

Cross-validation: We first evaluate the Mentor detection system by cross-validating our classifier model against the ground truth dataset at our disposal. We performed multiple experiments, each time randomly splitting our dataset into 80% of data that we used for training, and 20% that we used for evaluation. Then we used our training set as input to the selective Bayesian learning model in order to build our classifier. We further tested this classifier against the remaining evaluation set in order to evaluate the accuracy of our system.

The table of figure 2b summarizes the results of our cross-validation process. Mentor correctly classifies 99.02% of benign domains, with almost 0.98% misclassification rate regarding benign domains. On the other hand, it correctly classifies 95.32% of malicious domains, with an overall classification accuracy of 97.17%. Mentor has a higher accuracy when classifying benign domain names, including a higher hit rate and a lower false positives rate. In fact the features set that we use to build our classifier module characterizes the benign nature of a given domain. It searches for evidence that a given domain is being established

Blacklist	Date	Nb of entries	Malicious	Benign
Zeus tracker	20-01-2014	90	94.5%	5.5%
Palevo tracker	21-01-2014	26	92.4%	7.6%
SpyEye tracker	21-01-2014	125	96%	4%
Feodo tracker	21-01-2014	25	100%	0%
malwaredomains	20-01-2014	400	97.75%	2.25%

Table 2: Testing Mentor against public blacklists

for a professional, human friendly usage. Therefore, most domains in Alexa top domains list are clearly established for professional usage and so they were correctly classified by our system. On the other hand, although it achieved a very good detection accuracy, our system had a *relatively* higher false positives rate against malicious domains. We manually checked malicious domains that were misclassified by our system. Most of these domains are vulnerable benign domains that were exploited by an attacker and used for command and control. These domains were considered by our system as benign as they were first established for professional usage, before they have been hijacked by an attacker. We discuss in section 5 the limitation of our system against such benign domains as they are both used for benign and malicious purposes.

Real-world evaluation: In order to evaluate the performance of our system, we made real world experiments using public C&C blacklists that we extracted from the *abuse.ch* website[2], including *Zeus*, *SpyEye*, *Palevo* and *Feodo* malware blacklists. We also evaluated the ability of Mentor to detect C&C domains that were unknown by the time of building our system through testing against a more recent blacklist from *malwaredomains.com*. In fact we aim at validating through these experiments the ability of Mentor to correctly skim-off benign domains, and evaluate the consistency of our results for different blacklist categories.

Table 2 summarizes the main results of our experiments, including malicious and benign domains as identified by Mentor. The results in table 2 only include domain names that were accessible by the time of building our experiments. The Feodo blacklist included only 25 accessible domain names, all being identified as malicious C&C domains by our system. Because of the small number of domains in this blacklist, we manually checked each domain, and then we verified these domains using the google safe browsing API. All domains in this blacklist are indeed malicious C&C domains, and therefore Mentor achieves a 0% false positives rate for this blacklist.

On the other hand, Mentor achieved 97.75% true positives rate for the malwaredomains blacklist extracted on the 20th of January 2014. We believe this high matching rate is mainly because we trained the Mentor domain classifier using a similar older version of this blacklist. Indeed Mentor identified 9 benign

[2] `http://www.abuse.ch`

domains in this blacklist. Yet 7 domains were previously infected websites that are currently for rent or back under construction, and so we wouldn't consider them to be misclassified by our system. The remaining two accessible domains include *ankursociety.org* and *keymasconsultancy.co.uk*. We manually checked these two domains, and we scanned them using the Google safe browsing API. The *ankursociety.org* website was identified by the Google robot as previously distributing malicious content. The diagnostic provided by the Google safe browsing API indicates that *no malicious content on this website was detected during the last check*. The other domain was identified as suspicious by Google safe browsing, but it also indicates that no malicious content was recently identified on this website. The manual check of this domain reveals no malicious content, and it looks as a benign professional website for a UK-based company. These are clearly previously infected domains that are still present in the blacklist, and so we would consider them as true positives triggered by Mentor. Hence, Mentor has successfully skimmed-off two benign domains from the malwaredomains list, with no false positives.

As shown in table 2, our system achieved similar detection rates for the 3 remaining C&C blacklists, including *Zeus*, *Palevo* and *SpyEye*. Mentor detected 12 benign domains out of 241 suspicious domains in these blacklists. Hence, 95.1% of domain names in these blacklists were correctly identified as malicious C&C domains by our system. We checked the remaining 12 domains detected as benign by our system using Google safe browsing, and we manually observed their content for evidence about the malicious or benign nature of these domains. Indeed 6 domains were clearly identified as benign domains by the Google API. The manual analysis of these domains shows two blogsites and 4 professional domains, and so we would consider these as true positives. Four other domains seem to be hosting benign professional content. The Google robot previously detected suspicious content on these domains, but they were all currently identified as benign and so these are also true positives. On the other hand, the two remaining domains (*biozov.ru* and *psgtech72.com*) are clearly malicious domains and these were misclassified by our system. Therefore, Mentor achieved 0.8% false positives rate and skimed-off 10 benign domains during this experiment.

5 Discussion

Our system identifies and eliminates benign domains using content-based features and the popularity of a given domain name. It efficiently detects C&C domains when they are specifically built and established for this purpose. Nonetheless, as shown in our real world experiments described in section 4, Mentor is less efficient when detecting benign domain names that were hijacked and diverted by an attacker. Such compromised domains would be considered by Mentor as benign as long as their popularity and content are not hampered by the attacker. Note that the use of these domains for command and control is risky as they constitute single nodes of failure in the botnet architecture. In fact, administrators of these domains would rapidly take actions, as soon as they detect suspicious

usage of their websites, in order to prevent them from being used for malicious purposes. Yet multiple domains were identified as such by our system during our experiments in section 4. Hence, modern botnets increasingly adopt hybrid topologies including master bots that act as C&C servers and distribute commands towards slave bots. Such botnets are more robust as they include a larger set of master C&C nodes. However, they are efficiently detected by Mentor because master servers only act as C&C domains for other slave nodes, and so they would have different characteristics than other professional benign domains.

Mentor applies machine learning techniques to a statistical set of features in order to identify malicious domains. It would be unable to correctly classify C&C domains that share similar features with other benign domains such as high popularity, rich web content and long-lived domain names. These maneuvers would modify the statistical consistency of a malicious domain and so it would be identified as benign by our system. The use of these techniques by an attacker would require to carefully build C&C domains. It would also take a longer time for these domains to increase their popularity so they can no longer be detected by our system. Although being technically possible, these techniques cannot be easily automated. It would be difficult for an attacker to maintain a large enough set of C&C domains to ensure a better botnet resilience while also keeping its C&C domains under the detection radar of our system. Therefore, Mentor adds a new level of complexity for botnet herders in their struggle to keep their C&C domains undetected.

6 Conclusion

This paper presented a new system called Mentor, which implements positive DNS reputation to identify benign domains within a list of malicious C&C domain names. Positive DNS reputation measures the likelihood of a given domain name being innocuous, as opposed to negative DNS reputation that rather observes malicious artifacts for the same domain. Mentor describes a given domain name using three sets of features, including *popularity*, *content* and *domain-based* features. It implements an active crawler system that collects artifacts and content features from the remote domain itself and the public whois database. It groups all features for a given domain within a single vector and further applies machine learning techniques in order to separate benign and malicious domains. Mentor completes current negative DNS reputation systems; it processes domain blacklists generated by these systems and reduces their high false positives rate. Experimental results prove the ability of Mentor to efficiently identify and eliminate benign domains within blacklists of malicious C&C domain names, with a very low false positives rate.

References

1. Alexa web information company. http://www.alexa.com/topsites/, 2013.

2. M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster. Building a dynamic reputation system for dns. In *Usenix Security Symposium*, 2010.
3. 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 *USENIX Security Symposium*, 2012.
4. D. Balzarotti, M. Cova, C. Karlberger, C. Kruegel, E. Kirda, and G. Vigna. Efficient detection of split personalities in malware. In *International Symposium on Network and Distributed System Security (NDSS)*, 2010.
5. L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel. Disclosure: detecting botnet command and control servers through large-scale netflow analysis. In *Int. Annual Computer Security Applications Conference (ACSAC)*, 2012.
6. L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. Exposure: Finding malicious domains using passive dns analysis. In *Symposium on Network and Distributed System Security*, 2011.
7. H. Choi, H. Lee, H. Lee, and H. Kim. Botnet detection by monitoring group activities in dns traffic. In *Seventh International Conference on Computer and Information Technology*, 2007.
8. N. Cristianini and J. Shawe-Taylor. An introduction to support vector machines and other kernel-based learning methods. In *Cambridge University Press*, 2000.
9. M. Felegyhazi, C. Kreibich, and V. Paxson. On the potential of proactive domain blacklisting. In *Third USENIX LEET Workshop*, 2010.
10. T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling. Measuring and detecting fast-flux service networks. In *Symp. on Network and Distributed System Security*, 2008.
11. G. Jacob, R. Hund, C. Kruegel, and T. Holz. Jackstraws: Picking command and control connections from bot traffic. In *USENIX Security Symposium*, 2011.
12. N. Kheir and X. Han. Peerviewer: Behavioral tracking and classification of p2p malware. In *5th Int. Symposium on Cyberspace Safety and Security (CSS)*, 2013.
13. N. Kheir and C. Wolley. Botsuer: Suing stealthy p2p bots in network traffic through netflow analysis. In *12th Int. Conf. Cryptology and Network Security (CANS)*, 2013.
14. P. Langley and S. Sage. Induction of selective bayesian classifiers. In *10th international conference on Uncertainty in artificial intelligence*, pages 399–406, 1994.
15. P. Mockapetris. Dns encoding of network names and other types. RFC 1101, April 1989.
16. A. Moser, C. Kruegel, and E. Kirda. Exploring multiple execution paths for malware analysis. In *International Symposium on Security and Privacy*, 2007.
17. J. Postel. Domain name system structure and delegation. In *RFC 1591*, 1994.
18. J. R. Quinlan. C4.5: Programs for machine learning. In *Morgan Kaufmann Publishers*, 1993.
19. M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. A multifaceted approach to understanding the botnet phenomenon. In *6th ACM SIGCOMM conference on Internet measurement*, 2006.
20. S. Sinha, M. Bailey, and F. Jahanian. Shades of grey: On the effectiveness of reputation-based "blacklists". In *International Conference on Malicious and Unwared Software (Malware)*, 2008.
21. P. Wurzinger, L. Bilge, T. Holz, J. Goebel, C. Kruegel, and E. Kirda. Automatically generating models for botnet detection. In *14th European Symposium on Research in Computer Security (ESORICS)*, 2009.
22. S. Yadav, A. K. Reddy, A. N. Reddy, and S. Ranjan. Detecting algorithmically generated malicious domain names. In *10th ACM SIGCOMM conference on Internet measurement*, 2010.