Unsupervised Clustering for Identification of Malicious Domain Campaigns

Michael Weber
Palo Alto Networks
3000 Tannery Way
Santa Clara, CA 95054, USA
mweber@paloaltonetworks.com

Jun Wang
Palo Alto Networks
3000 Tannery Way
Santa Clara, CA 95054, USA
junwang@paloaltonetworks.com

Yuchen Zhou
Palo Alto Networks
3000 Tannery Way
Santa Clara, CA 95054, USA
yzhou@paloaltonetworks.com

ABSTRACT

New malicious domain campaigns often include large sets of domains registered in bulk and deployed simultaneously. Early identification of these campaigns can often be accomplished with distance functions or regular expressions of registered domains, but these methods may also miss some campaign domains. Other studies have used time-of-registration features to help identify malicious domains. This paper explores the use of unsupervised clustering based on passive DNS records and other inherent network information to identify domains that may be part of campaigns but resistant to detection by domain name or time-of-registration analysis alone. We have found that using this method, we can achieve up to 2.1x expansion from a seed of known campaign domains with less than 4% false positives. This could be a useful tool to augment other methods of identifying malicious domains.

CCS CONCEPTS

•Security and privacy → Malware and its mitigation;

KEYWORDS

Unsupervised machine learning, clustering, malware detection, DBSCAN, Agglomerative clustering

ACM Reference format:
Michael Weber, Jun Wang, and Yuchen Zhou. 2016. Unsupervised Clustering for Identification of Malicious Domain Campaigns. In *Proceedings of Radical and Experiential Security Workshop, Songdo, Incheon, Korea, Jun 4-8, 2018 (RESEC'18),* 7 pages.
DOI:

1 INTRODUCTION

One class of malicious on-line activity involves registration of domains that take advantage of a topical event. The domain names often utilize typo-squatting of legitimate domains names or names that indicate some relevance to legitimate services. Recent examples of this include malicious campaigns released after the Equifax data breach or critical software bug updates.

RESEC'18, Songdo, Incheon, Korea

978-x-xxxx-xxxx-x/YY/MM…$15.00
DOI:

In the case of the Equifax breach, the credit reporting agency set up a legitimate website, www.equifaxsecurity2017.com, to help people determine whether they had been affected. This triggered one or more malicious campaigns that registered hundreds of domains that appeared similar to the real URL. For example, one such domain was www.equifaxsecurity3017.com

Detection of such domains can sometimes be achieved through analysis of the registered domains alone by using regular expressions or distance function to identify similar domains. This has also been done with time-of-registration features [2][6][7]. This is particularly useful when the campaign domains are registered in bulk at the same time. However, identification by domain name alone can often be evaded. For example, the attacker can choose to register domains at different time and/or with sufficiently distinct domain names, such as www.ewuifactssecutity3017.com. However, even with these varieties, we observe that malicious domains belong to the same campaign still share many common characteristics such as IP subnet, ASN, DNS TTL, Whois information, and many other attributes. Based on such an observation, this paper discusses the use of unsupervised clustering of domains by using passive DNS records and other factors to complement existing methods and identify campaign domains that might not be identified otherwise.

The features used in this analysis have been determined by passive DNS information, along with Whois and BGP data to collect network properties and behaviors related to domains. Clustering is performed to group domains that have similar characteristics. By using a few seed domains from known campaigns, additional domains can be identified by being clustered with the seed domains. Many of the features used in the clustering have been used in other studies related to malicious classification [1][4][5]. The contribution of this paper is to demonstrate the use of such features with unsupervised clustering in order to expand identification of malicious campaigns based on a small set of known seed domains. We have seen that a small set of seed domains can be expanded 2.1x with less than 4% false positives.

2 CLUSTERING MOTIVATION AND METHODOLOGY

There have been several studies that use supervised methods to classify domains as malicious or benign [1][4][5]. Rather than directly classifying domains, this study seeks to group similar domains together and identify malicious campaign domains that might go undetected otherwise. These unsupervised methods appear to be effective in identifying previously undetected domains used in specific campaigns. This can be a useful tool in blocking these topical campaigns early when they pose the greatest threat.

2.1 Methodology

The methodology of this study focused on malicious campaigns related to the Equifax breach disclosure and fake software updates that occurred in September 2017. Passive DNS traffic was analyzed for the week following the disclosure on September 7th, and features were created for all observed domains based on this data. The passive DNS traffic was provided by Farsight Security [11]. Additional features were created based on Whois and BGP data. The features are discussed in the following section.

Some passive DNS records were filtered out as part of preprocessing of the data. Since the goal is to find domains that are part of a new topical campaign, any domains that are older than one year were removed. Additionally, any domains found the Alexa top 250,000 in August 2017 are considered benign or unrelated to the campaigns and were filtered out. This still yielded a large number of unique domains seen during the week. A random sampling of approximately 1% of the domains was chosen to create a list of 100,000 domains for analysis. In addition, only passive DNS "A" (host) records were used. Although additional useful features could be generated for other record types, only "A" records were used for this initial analysis.

The 100,000 domains were analyzed with various clustering algorithms and parameters to determine effectiveness of the process and which algorithms delivered the best results. The clustering algorithms were chosen based on their suitability for this application and their ability to scale to the number of domains required. Scikit-learn [3] was used as the clustering framework for each of the algorithms.

2.2 Validation Set

A set of ground truth domains related to the campaigns was generated with a combination of distance functions, regular expressions, and manual review. Domains registered within a couple days of the start of campaign were analyzed and domain names with a small edit distance from the legitimate Equifax domain were flagged. Regular expressions, such as eq.*fax were applied to domains seen the following week in passive DNS. These were then reviewed to manually remove domains that did not appear related.

This yielded 1541 unique domains that could be part of the targeted campaigns. Examples of malicious campaign domains include:

- ecuifaxsecurity2017.com
- edquifaxsecurity2017.com
- eequifaxsecurity2017.com
- apptraffic2update.club
- apptraffic4update.bid
- apptrafficforupdates.stream

Since this list relied on domain name similarity rather than a more definitive list of campaign domains, there is noise in this list. For example, even though two domains may appear similar, they may not be part of the same campaign or even be malicious. For our purposes, however, this does not inhibit the detection process because the analysis only looks at clusters of domains with similar features to known seed domains. Any benign domains in the seed list are unlikely to have large clusters of similar domains.

The ground truth domains were randomly split into two groups, 20% as the campaign seed domains and 80% as the campaign verification domains. Clusters generated for all domains were then analyzed with respect to the seed domains. Clusters with more than 10% seed domains were considered candidate clusters of campaign domains. On average, a cluster that is completely malicious would be expected to be 20% seed domains and 80% validation domains. The threshold of 10% was chosen to capture clusters with at least half of the expected seed domains. In Section 5, different threshold values are tested to determine the best setting.

The rest of the domains in the cluster were analyzed against the campaign verification list. A minimum cluster size of five was also established to filter clusters with very few domains because they do not yield actionable information. The minimum cluster size was also tested to determine a reasonable setting and the results are shown in Section 5.

Success criteria considered whether candidate clusters demonstrate that a small amount of seed campaign domains can yield a larger set of verification campaign domains with few false positives.

3 FEATURES

The features used in the clustering process are shown in Table 1. Since generating novel features was not the focus of this study, most of the features used have been previously used in other classification studies. While this study shows that these existing features can be very useful in unsupervised methods as well as supervised methods, future work can be done to find features of additional utility for unsupervised efforts.

The IP Address is taken from the passive DNS record and subnet is derived from the address. Any IP addresses in the data set associated with domains classified as malicious by Virus Total [10] are noted and all domains associated with that IP are flagged. The ASN with this IP is taken from BGP, and boolean features flag whether the ASN is known to be "bullet-proof" or rentable. Percentage of digits in the domain and the longest meaningful substring are both derived from the domain name. The numbers of unique IP addresses and TTLs for records in the data set are collected for each domain. The number of unique countries is inferred from the IP addresses. The age of the domain is inferred from its first appearance in historical passive DNS records and the registrar is taken from Whois. Daily similarity, short-lived, and repeated patterns are time-based features taken from analyzing the week of passive DNS records per domain. Daily similarity indicates whether the daily total of passive DNS records for a domain is within 50% of the weekly average on at least 5 days of the week. Short-lived indicates whether the domain was not seen for more than 2 days in a week. Repeated pattern looks at the difference between hourly totals on subsequent days and calculates a total Euclidean distance for the week.

4 ALGORITHMS

The following clustering algorithms were analyzed with this data set:

- DBSCAN
- K-Means
- Birch

Table 1: Features Generated for Clustering

Feature	Source
IP Address	Passive DNS
Subnet (/24)[a]	Passive DNS
ASN	BGP
Known Malicious IP[b]	Virus Total
Bullet Proof ASN	Private company
Rentable ASN	Private company
Percentage of Digits in Domain[5]	Passive DNS
Number of Unique IPs Seen for Domain[1][5]	Passive DNS
Number of Unique TTLs Seen for Domain[5]	Passive DNS
Length of Longest Meaningful Substring[5]	Passive DNS
Number of Unique Countries Seen[1][5]	Passive DNS
Age of Domain[1][c]	Passive DNS
Registrar of Domain[1]	Whois
Daily Similarity of Passive DNS Records[5]	Passive DNS
Short-Lived Passive DNS History[5]	Passive DNS
Repeated Pattern of Passive DNS Records[5]	Passive DNS

[a]While the actual broadcast domain of an individual IP address cannot be determined from passive DNS, it was observed that many campaign domains use IP addresses that are from the same /24 address block. For analysis purposes, the subnet feature is simply the /24 of each IP address.

[b]Using Virus Total as ground truth for malicious domains, when malicious domains are observed, the associated IP address was collected and used as a feature. Any new domain associated with this known malicious IP address was used as a feature.

[c]The age of a domain was determined by historical analysis of passive DNS data rather than relying on the Whois database of this information, since much of the Whois data is missing or unreliable.

- Ward Hierarchical Clustering with and without connectivity constraints[1] and
- Agglomerative Clustering with and without connectivity constraints

These algorithms were chosen primarily for their ability to scale to large sample sets, as well as to provide a broad coverage of available clustering algorithm types.

Each algorithm has various input parameters. For this study, one primary input parameter was chosen per algorithm to tune the performance of the algorithm. For DBSCAN, the input parameter is EPS, the maximum distance for two samples to be considered part of the same cluster. For Birch, the parameter is the Birch Threshold, the radius of the sub-cluster obtained by merging a new sample and the closest sub-cluster. For K-means, Ward Hierarchical and Agglomerative Clustering, the input parameter is number of clusters.

While each of the algorithms were selected in part for their scalability, continuous analysis of passive DNS data for production implementation will require high performance. The execution time for this data set was also evaluated for each algorithm and is documented in the final results.

5 EVALUATION

The algorithms were evaluated based on the maximum cluster campaign percentage, reflecting the least amount of false positives. The

[1]The connectivity constraints were established with a KNN graph with 100 neighbors for each sample.

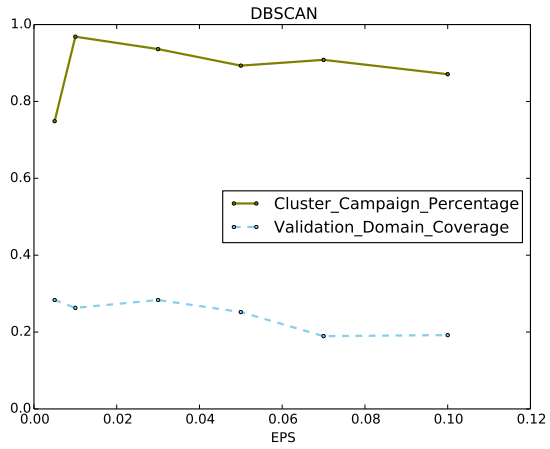

Figure 1: Results of analyzing the data set with DBSCAN for different input values of the EPS parameter. The cluster campaign percentage indicates what percentage of the flagged clusters were malicious. The validation domain coverage shows what percentage of the total validation domains were found in the flagged clusters.

results of the various algorithms are shown in Table 2. Overall, the two best performing algorithms were DBSCAN and Agglomerative Clustering with Connectivity Constraints. To select the best parameters for each algorithm, we adjust the input parameter and compare the coverage and false positives. Figures 1 and 2 show the results for these two best performing algorithms for various input parameters. The results for the rest of the algorithms are shown in Appendix A. The purpose of evaluating multiple algorithms and input parameters is to help determine the relative effectiveness of the algorithms on this data set and identify the optimal tuning parameters for each. The most promising algorithms can then be further analyzed for detailed understanding of the utility of the process.

In each graph, the line labeled Cluster_Campaign_Coverage indicates, in all of the clusters with at least 10% seed domains, what is the total percentage of campaign seed or campaign verification domains in those clusters. The ideal value for this metric is 100%, and any other domain is considered a false positive. Although as we will see in the further evaluation, some of the domains that show up as "false positives" may turn out to be previously unknown campaign domains.

The 1541 ground truth domains were split into a seed group of 308 domains and a validation group of 1233 domains. The line labeled Validation_Domain_Coverage indicates how many of the 1233 campaign validation domains have been identified in this process; that is, how many of the 1233 domains appear in clusters with at least 10% seed domains. Anything less than 100% could be considered a false negative. However, since we are constraining cluster size to keep the results actionable, and since the ground truth domain set is known to be noisy, we are not expecting or intending to eliminate false negatives. For the purposes of this analysis, it is far more important to identify new malicious campaign domains

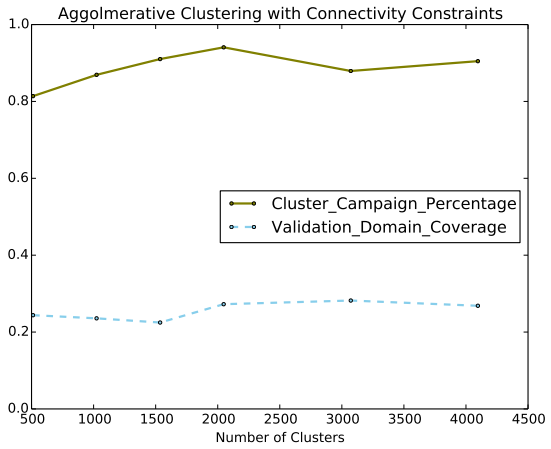

Figure 2: Results of analyzing the data set with Agglomerative Clustering with connectivity constraints for different input values of the number of clusters.

with low false positives than it is to identify all malicious domains because this method is meant to augment existing methods. However, this value does provide guidance of the validity of the process. Results with only a small number of identified domains will not be useful for production purposes.

These results indicate that all of the clustering methods provide benefit, identifying clusters of domains that are on average at least 80% campaign domains. DBSCAN and Agglomerative Clustering with connectivity constraints appear to work the best with this data set identifying clusters with more than 94% campaigns domains.

5.1 Analysis of Results

Looking into the Agglomerative Clustering results with 2048 clusters, there were 15 clusters that had more than 10% seed campaign domains. In these 15 clusters, there were 270 domains in total, 54 of which were seed domains and 200 were validation domains. On average, 94% of the domains in the clusters were seed or validation campaign domains. Eight clusters were 100% campaign domains. Of the 16 potential false positives, manual analysis showed that 4 of the 16 were known malicious sites, according to Virus Total, and three of those four appeared related to the Fake Update campaign but they had not been included in the original campaign list. Six were part of the Equifax campaign, although they were also not on the original ground truth list. More interestingly, these domains were not previously identified as malicious by Virus Total, indicating that this method can identify malicious domains not found by standard methods.

Only six appeared to be unrelated to known campaigns or known to be malicious, making the effective false positive rate 2.22%. Three of the six false positives were in one large cluster with 56 domains related to the Equifax campaign. Two were in a different Equifax campaign cluster, and one was in an Fake Update campaign cluster. Looking at the features of domains in these clusters, all six false positives appeared to have IP addresses and ASNs numerically close to those in the campaign. This does not appear to be anything more than coincidence, and is a limitation of the existing methodology. The standard framework used Euclidean distance for the features, which for a minority of the features, including IP address and ASN, is not ideal. A binary matching comparison should instead be used for those features, and would be part of a custom distance function. This is left for future work, and would likely resolve most or all of these false positives.

The top DBSCAN results identified 253 domains across 16 clusters, with 52 seed domains, 193 validation domains, and 8 false positives. Of those 8 false positives, one appears related to the Fake Update campaign and is known malicious according to Virus Total. One appears related to the Equifax campaign based on the domain name and is not known to Virus Total. Two more do not appear to be related to any campaign but are known malicious. Four do not appear to be campaign related but share the exact IP address of a domain in an Equifax campaign. Of the 8 apparent false positives, on closer inspection, 6 seem to be related to campaigns, 1 is unrelated but known malicious, and 1 is a likely real false positive. The single false positive also has an IP address and ASN that are numerically close to campaign domains in the cluster, and this is also likely a coincidence that a custom distance function would resolve.

5.2 Cluster Threshold

Selecting the proper threshold of seed domains in a cluster will minimize the false positives while still yielding usable results. To determine the proper threshold, the top two algorithms were run with various threshold values. Figures 3 and 4 show the two top performing algorithms with different thresholds set for how many seed campaign domains are found in a cluster for it to be considered a cluster of campaign domains. As the threshold increases, the percentage of campaign domains in the cluster increases, as expected. However, the total number of validation domains also goes down. For example, when the threshold is set at 30% for DBSCAN, 100% of the cluster domains are in the validation campaign list, but this only yields a single domain. Based on this data set, the best threshold for both algorithms that balances total coverage to cluster percentage is 10%.

5.3 Minimum Cluster Size

Another variable is what the minimum cluster size should be to balance usable results and total coverage of the campaign domains. For example, including all clusters with a single domain may increase the total number of seed domains found, but it does not provide additional usable results. Figure 5 shows the results of different minimum cluster sizes for Agglomerative Clustering.

Reducing the minimum cluster size does increase the validation domain coverage a little, but below a minimum of five, there is little additional benefit in this data set.

5.4 Expansion

The primary use case for this process is to expand knowledge of domains being used in malicious campaigns. One way to gauge the effectiveness of the technique is to evaluate the level of expansion achieved from the initial seeds. The previous results used a seed set of 20% of the 1541 campaign domains. To verify what the expansion

Table 2: Evaluation results of the different algorithms with their best input parameter setting. DBSCAN and Agglomerative Clustering yielded the highest cluster campaign percentage.

Algorithm	Best Parameter	Cluster Campaign %	Validation Domain Coverage	Total Domains	Malicious Domains	False Positives	Run Time
DBSCAN	0.01	96.9%	26.3%	253	245	8	87 s
AC w/ Constraints	2048	94.1%	27.2%	270	254	16	270 s
Birch	0.05	90.7%	29.0%	292	265	27	51 s
AC w/o Constraints	2048	87.8%	30.1%	319	280	39	670 s
WC w/ Constraints	3072	84.9%	33.0%	364	309	55	269 s
WC w/o Constraints	3072	84.9%	33.0%	364	309	55	738 s
K-Means	3072	83.1%	33.4%	379	315	64	1465 s

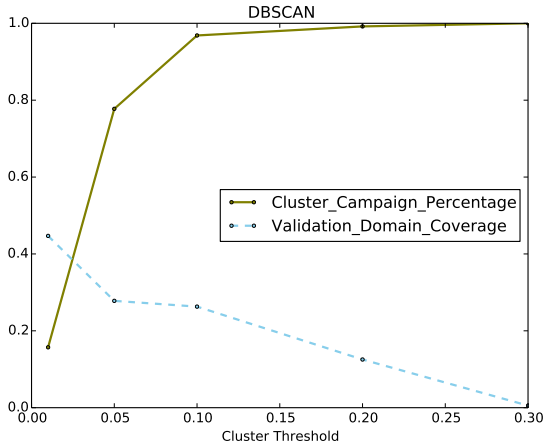

Figure 3: Results of DBSCAN with different cluster threshold values. A lower threshold value will flag more clusters and potentially find more total validation domains, but the clusters will also be more likely to have false positives.

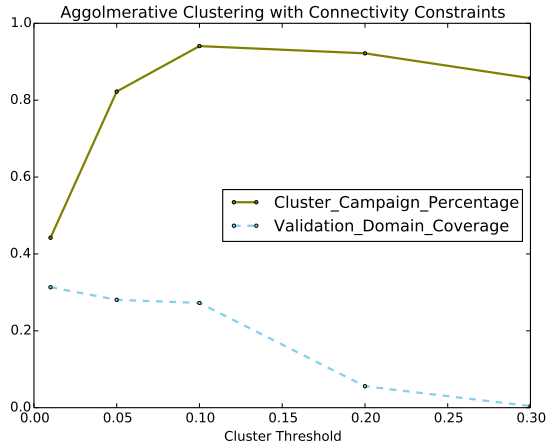

Figure 4: Results of Agglomerative Clustering with different cluster threshold values.

would be for different seeds, various seed sizes ranging from 1% to 90% were analyzed for their expansion. The smallest seeds showed the greatest expansion, but there appears to be a minimum threshold below which there is a trade off with cluster percentage. For this data set, a seed group of 150-300 domains, or 10-20%, provides the greatest expansion while maintaining clusters of 96% malicious domains. The expansion results are shown in table 3.

6 RELATED AND FUTURE WORK

6.1 Detecting malicious domains through DNS

A great deal of work has been done to leverage passive DNS data to detect malicious domains. Antonakakis et al. [1] developed Notos to use features of passive DNS records to determine whether a given domain is malicious. Bilge et al. [5] created EXPOSURE with a different set of unique features to classify domains. Antonakakis et al. [4] followed up with Kopis to monitor traffic at the upper levels of the DNS hierarchy and classify domains. The main difference between this work and these previous studies is that they are focused on classification using supervised methods. This work is testing whether unsupervised clustering methods can be used to expand the identification of known malicious campaigns. Khalil et al. [8] built associations among domains based on passive DNS data and used these associations to identify malicious domains. While they solely relied on domain-IP resolutions to build associations, we leverage more information from passive DNS, Whois and BGP, which can be more accurate in profiling the characteristics of malicious domains.

6.2 Detecting malicious domains through registration

A number of studies have also explored automated detection of malicious domains from information available during registration. Felegyhazi et al. [2] detected malicious domains from registration information found from DNS zone records. Hao et al. [6] studied the registration behavior of spammers to identify malicious domains. Hao et al. [7] developed PREDATOR for early detection of malicious domains with only time-of-registration features. Liu et al. [9] proposed Woodpecker for automated detection of shadowed domains. This work seeks to augment those techniques by expanding the identified domains based on a few seed domains.

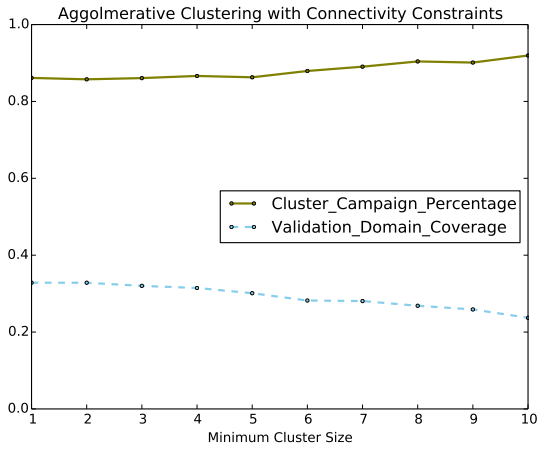

Figure 5: Results of Agglomerative Clustering with different minimum cluster size values.

Table 3: Expansion results of using different seed percentages from the original list of 1541 campaign domains. The cluster percentage begins to degrade with seeds below 10%, yielding expansion of 2.1x.

Seed %	# Seeds	# Found	Expansion	Cluster%
1%	17	108	6.35x	29%
5%	72	197	2.74x	55%
10%	154	325	2.11x	96%
20%	308	509	1.74x	96%
30%	462	633	1.37x	95%
40%	620	766	1.24x	97%
50%	770	886	1.15x	97%
60%	921	1046	1.14x	99%
70%	1079	1171	1.09x	99%
80%	1232	1286	1.04x	99%
90%	1387	1422	1.03x	99%
95%	1469	1480	1.01x	99%
99%	1524	1527	1.00x	99%

6.3 Future Work

Initial analysis indicates that this can be an effective process for identifying domains that are part of topical campaigns. Based on this early testing, follow-on work to improve the system should include further focus on the features used for clustering. Many additional features are possible, and some of the existing features may not provide much current value. The current study focused on expansion of malicious domain campaigns, but future work could investigate identification of new campaigns, and other types of malicious domains.

The limited distance functions used for clustering are not appropriate for all of the features involved. Some of the features, such as percentage of digits in the domain name should have a Euclidean distance function to gauge similarity, but others, like IP address, should have a binary matching function. Future work should include a custom distance function that is appropriate to the final features used.

Finally, Internet scale performance will need to be considered as part of a further implementation. Although all of the tested algorithms were chosen for scalability, the data set was still a fraction of normal daily traffic. Some of the algorithms will not naively scale to the amount of daily passive DNS records seen.

7 CONCLUSION

Malicious campaigns such as the Equifax breach campaigns or the Fake Update campaigns are particularly insidious because they attempt to take advantage of topical crises and can affect large groups people who may not otherwise have been compromised. Comprehensive identification of these domains is critical for broad network protection. The analysis in this study of several clustering algorithms on passive DNS data show that this methodology can be used to identify a substantial amount of previously unknown malicious domains from a small amount of seed domains and and can be an effective tool to combat malicious campaigns.

REFERENCES

[1] Manos Antonakakis, Roberto Perdisci, David Dagon, Wenke Lee, and Nick Feamster. 2010. Building a Dynamic Reputation System for DNS. In Proceedings of the 19th USENIX Conference on Security.
[2] Mark Felegyhazi, Christian Kreibich, and Vern Paxson. 2010. On the Potential of Proactive Domain Blacklisting. In Proceedings of the USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More (LEET).
[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[4] Manos Antonakakis, Roberto Perdisci, Wenke Lee, Nikolaos Vasiloglou, II, and David Dagon. 2011. Detecting Malware Domains at the Upper DNS Hierarchy. In Proceedings of the 20th USENIX Conference on Security.
[5] Leyla Bilge, Engin Kirda, Christopher Kruegel, and Marco Balduzzi. 2011. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis. In Proceedings of the Annual Network and Distributed System Security Symposium (NDSS).
[6] Shuang Hao, Matthew Thomas, Vern Paxson, Nick Feamster, Christian Kreibich, Chris Grier, and Scott Hollenbeck. 2013. Understanding the Domain Registration Behavior of Spammers. In ACM IMC.
[7] Shuang Hao, Alex Kantchelian, Brad Miller, Vern Paxson, and Nick Feamster. 2016. PREDATOR: Proactive Recognition and Elimination of Domain Abuse at Time-Of-Registration. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS).
[8] Issa Khalil, Ting Yu, Bei Guan. Discovering Malicious Domains through Passive DNS Data Graph Analysis. Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security.
[9] Daiping Liu, Zhou Li, Kun Du, Haining Wang, Baojun Liu, Haixin Duan. Donfit Let One Rotten Apple Spoil the Whole Barrel: Towards Automated Detection of Shadowed Domains. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security.
[10] Virus Total. 2017. https://www.virustotal.com/.
[11] Farsight Security. https://www.farsightsecurity.com/solutions/dnsdb/.

A ADDITIONAL ALGORITHM RESULTS

The following graphs show the results for the remaining algorithms tested.

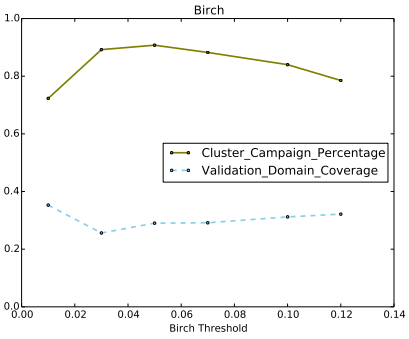

Figure 6: Results of analyzing the data set with Birch for different input values of the threshold parameter.

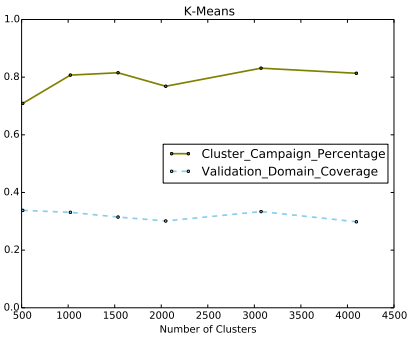

Figure 7: Results of analyzing the data set with K-Means for different input values of the number of clusters.

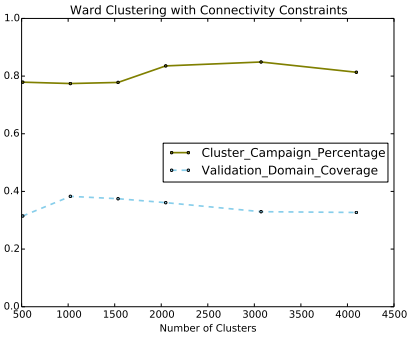

Figure 8: Results of analyzing the data set with Ward Clustering with connectivity constraints for different input values of the number of clusters.

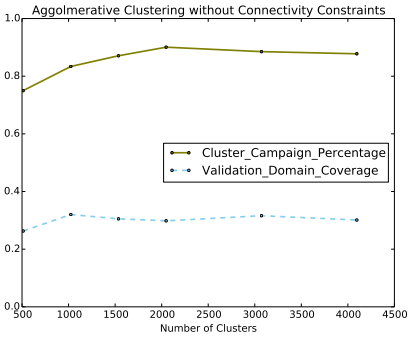

Figure 9: Results of analyzing the data set with Agglomerative Clustering without connectivity constraints for different input values of the number of clusters.

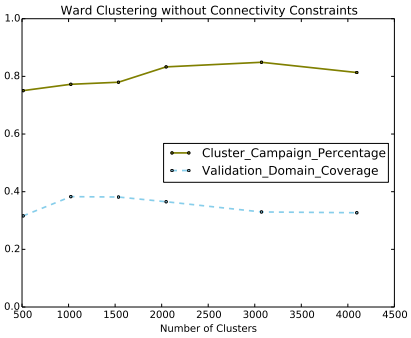

Figure 10: Results of analyzing the data set with Ward Clustering without connectivity constraints for different input values of the number of clusters.