View from Above: Exploring the Malware Ecosystem from the Upper DNS Hierarchy

Aaron Faulkenberry*
Georgia Institute of Technology
afaulken@gatech.edu

Athanasios Avgetidis*
Georgia Institute of Technology
avgetidis@gatech.edu

Zane Ma
Georgia Institute of Technology
zanema@gatech.edu

Omar Alrawi
Georgia Institute of Technology
alrawi@gatech.edu

Charles Lever
Devo
Chaz.lever@devo.com

Panagiotis Kintis
Voreas Laboratories Inc
panos@voreas.io

Fabian Monrose
Georgia Institute of Technology
fabian@ece.gatech.edu

Angelos D. Keromytis
Georgia Institute of Technology
angelos@gatech.edu

Manos Antonakakis
Georgia Institute of Technology
manos@gatech.edu

ABSTRACT

This work explores authoritative DNS (*AuthDNS*) as a new measurement perspective for studying the large-scale epidemiology of the malware ecosystem—when and where infections occur, and what infrastructure spreads and controls malware. Utilizing an *AuthDNS* dataset from a top registrar, we observe malware heterogeneity (202 families), global infrastructure (399,830 IPs in 151 countries) and infection (40,937 querying Autonomous Systems (ASes)) visibility, as well as breadth of temporal coverage (2017–2021). This combination of factors enables an extensive analysis of the malware ecosystem that reinforces prior work on malware infrastructure and also contributes new perspectives on malware infection distribution and lifecycle. We find that malware families re-use infrastructure, especially in cloud hosting countries, but contrary to prior work, we do not detect targeting of clients by countries or industry sector. Furthermore, our 4-year lifecycle analysis of diverse malware families shows that infection analysis is temporally sensitive: over 90% of ASes first query a malicious domain after public detection, and a median of 38.6% ASes only query after domain expiration or takedown. To fit *AuthDNS* into the broader context of malware research, we conclude with a comparison of experimental vantage points on four qualitative aspects and discuss their advantages and limitations. Ultimately, we establish *AuthDNS* as a unique measurement perspective capable of measuring global malware infections.

ACM Reference Format:
Aaron Faulkenberry, Athanasios Avgetidis, Zane Ma, Omar Alrawi, Charles Lever, Panagiotis Kintis, Fabian Monrose, Angelos D. Keromytis, and Manos Antonakakis. 2022. View from Above: Exploring the Malware Ecosystem from the Upper DNS Hierarchy . In *Annual Computer Security Applications Conference (ACSAC '22), December 5–9, 2022, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3564625.3564646

*Authors contributed equally.

ACSAC '22, December 5–9, 2022, Austin, TX, USA

ACM ISBN 978-1-4503-9759-9/22/12.
https://doi.org/10.1145/3564625.3564646

1 INTRODUCTION

Malware is a pervasive and growing problem [21, 22]. To counter this rising tide, the security community has performed extensive research into understanding malware and has devised techniques for detection, mitigation, and prevention. Unfortunately, malware is extremely diverse—it spans potentially unwanted programs (PUPs), ransomware, and rootkits—making it difficult to generalize results and defenses based on individual malware families.

Ecosystem-wide analysis of malware is necessary to understand broad malware characteristics and to enact appropriate high-level protections and policies. For example, Lever et al. [18] noted heavy malicious usage of popular cloud hosting services which introduced the need for stricter vetting and policing by providers. As another example, Kotzias et al. [15] found that different industries have highly variable infection rates (76% versus 16% for Electrical Equipment compare to Banking), which either suggests targeted attacks by malware operators or indicates that security policies for some industries are more effective than others. Macro-level analysis of malware at large can lead to solutions with far-reaching impact.

Although prior work has explored many aspects of the malware ecosystem, existing research perspectives only have partial visibility into when and where malware infections occur. With the exception of peer-to-peer networks, malware sandboxes cannot observe infected hosts in the wild. The visibility of passive recursive DNS [18] is limited to a handful of collaborating networks. Host-based measurement [15] is often biased or dependent on pre-installed software and challenging to scale globally. Sinkholes [3, 38] miss infection phases prior to infrastructure takedown. Studies focused on individual malware families (e.g., Mirai [5], ransomware [14]) may have nearly complete visibility, but the lack of malware heterogeneity precludes broader malware ecosystem insight.

This work explores passively collected authoritative DNS (*AuthDNS*) server logs as a new vantage point for characterizing the broader malware ecosystem. The ubiquity of DNS for network communications and its hierarchical nature create an opportunity to examine malware across four dimensions: malware family diversity, full lifecycle time span, and global visibility into both malware infections and infrastructure. Leveraging data from one of the twenty

largest top-level DNS authority zones[1], we study the extent to which *AuthDNS* can replicate previous research findings and also further expand our understanding of the malware ecosystem.

We perform three case studies from the *AuthDNS* perspective. First, looking at malware infrastructure, we find substantial overlap in the networks utilized by different malware families. In the most extreme case, we observe an AS hosting 715 domains associated with 94 distinct malware families. This observation supports prior work [18, 23, 42], which show malware hosting is often interlaced with legitimate infrastructure. We perform a detailed comparison to understand the nuances of each measurement perspective.

Second, we examine the breadth of global malware infections. Previous works studying a wide set of malware have detailed visibility into a specific subset of affected clients (e.g., enterprise networks protected by a specific AV vendor [15]). The *AuthDNS* vantage point provides a slice of global visibility. After looking at all querying clients, we find that targeted malware infections are not apparent for most malware families. Instead, we find that infection rates per country or sector correlate ($\geq$ 0.95 Spearman's ρ) with overall network activity. The vast majority of clients fall under the *Information & Communication* or *Wholesale & Retail Trade* (due to how Amazon's space is classified) industry sectors.

Third, we examine an under measured aspect of the malware ecosystem: the full lifecycle of malware communications, from domain registration to blocklisting, and ultimately, expiration. We find that most malicious domains are set up and detected quickly, within five days for 50% of new registrations. Furthermore, we observe a multitude of scanners that emerge after a domain's detection, as well as a median 38.6% of new client networks first querying malicious domains *after* their expiration. Two explanations for this phenomenon are persistent infections on mobile clients that migrate ASes, or scanners and security professionals querying expired domains [38]. Estimating malware infections from a network perspective after a domain's expiration should be done with caution.

This study comprises the central pillars for malware epidemiology: the infrastructure that spreads and controls malware, and the location and timing of client infections. To understand how *AuthDNS* supplements existing research, we discuss the advantages and limitations of each vantage point. We then categorize the general types of ecosystem properties (e.g., malware variants, victim targeting, etc.) and provide guidelines for which perspectives will yield meaningful measurements. Ultimately, this work establishes *AuthDNS* as a unique outlook on the malware ecosystem, replicates prior results on malware infrastructure, expands our understanding of malware epidemiology, and introduces a framework to contextualize existing and future research.

2 BACKGROUND

Prior ecosystem-wide studies of malware have investigated a wide swath of properties (e.g., infection mechanism, obfuscation techniques, evasion of malware classifiers, network-based attacks, command and control) that all aim to detect, mitigate, or prevent malware infections. This work focuses on the epidemiological properties of malware, which broadly examines 1) when and where malware infections occur and 2) what infrastructure spreads and controls the malware. We defer the discussion of related works to Section 7 as part of a broader comparison of vantage points and the malware properties they are best suited to capture.

[1]Undisclosed due to data sharing agreements.

2.1 Authoritative DNS

The domain name system (DNS) [24] is one of the core components of the Internet. DNS translates semantic domain names into IP addresses, making it a useful data source for observing Internet communication. While DNS data can be collected at many locations, this work focuses on passive collection at the authoritative nameserver, enabling us to observe all requests for a given domain. We refer the reader to [25, 26] for a broader overview of DNS.

Global visibility via *AuthDNS* has a few limitations. DNS requests to the authoritative nameserver originate from DNS recursives instead of the end-user, and lookup volumes are reduced due to caching at recursive. RFC 7871 [11] introduced a DNS extension, EDNS Client Subnet (ECS), to allow geographically optimized responses when a DNS client's recursive is not located near the client. ECS includes a portion (the first 24 bits, by default) of the client's IP address in DNS requests sent from the recursive to the authoritative nameserver. This enables an authoritative nameserver to approximate the geographic location of a client and reply appropriately. As a result, ECS enables global DNS visibility at the authority with client-level visibility for enabled resolutions.

Prior work utilizing passive authoritative DNS data has focused on detection and measurement in the upper DNS hierarchy [6, 12, 31, 44, 47] and does not tackle the challenge of quantifying the larger malware ecosystem.

3 DATASETS AND METHODOLOGY

This section details *AuthDNS* and supporting datasets, describes our methodology, and discusses the limitations of our approach.

3.1 Datasets

Passive Authoritative DNS (*AuthDNS*). We collaborate with a domain registrar that collects DNS data at the authoritative DNS nameservers used by the top-level zones that it serves. Our DNS data spans 2017-02-09 to 2021-06-30 and includes all DNS packets sent or received by the authority. We extract the IP address of the recursive resolver, the domain name resolved, the response from the authority, and client IP subnet for ECS-enabled queries.
Malware DNS (*MAL*). We collect malware domains from a data partner[43] that executes suspicious Windows binaries in an isolated malware sandbox. The malware executions span from January 2018 to April 2021 and amount to 30,302,106 executions. We obtain the communications in PCAP form and extract the DNS traffic.
VirusTotal (*VT*). We query VirusTotal [2] to collect malware family classification labels for malware samples in our *MAL* dataset, and we use AVClass 2 [40] to identify the most relevant label. While *VT* offers results from a plethora of antivirus engines, we only use AV detection results from 17 popular antivirus (AV) vendors that we have found provide stable labels. Additionally, we utilize *VT* to extract historical data for malware samples, malicious domains, and the dates that they were first labeled as malicious.
IP Whois (*IPWHOIS*). We use the *Prefix-to-AS* dataset available from CAIDA [9] to annotate the networks initiating DNS requests.

We joined this data with the ASN-to-AS organization delegations provided by the Regional Internet Registries (RIRs). When discussing the *IPWHOIS* dataset, we are referring to the union of these datasets. We utilize this dataset to map IPs to the organizations (and countries) that announce their prefixes.

Industries (*IND*). In order to link an IP address to its industry, we use a commercial IP intelligence dataset. While the dataset is imperfect—a portion of Amazon's IP space is labeled as *Wholesale and Retail Trade*, which is partially accurate since Amazon's retail business utilizes its own cloud infrastructure—it represents one of the best labeling available. Open-source solutions such as ASdb [49] provide AS-level granularity that is too coarse for our purposes. *IND* includes organizational property information based on the "International Standard Industrial Classification of All Economic Activities" (ISIC). The Statistics Division of the United Nations (UNSD) [1] provides the mappings of ISIC codes and business categories. We intersect the two datasets to attribute an IP address to a specific business based on the UN standard. We refer to different industries as *sectors* through the rest of the paper.

3.2 Methodology and Validation

Generating Malware Dataset. To obtain a set of malware-related domains, we first find the overlap between our *MAL* and *AuthDNS* datasets. Malware samples may query benign domains to check for network connectivity. Similar to Lever et al. [18], we filter out top-ranked domains in Trancolist [34]. This filtering yielded 12, 212 effective second-level domains (e2LDs), which capture the registrable portion of a domain name. For example, in the fully qualified domain name *www[.]example[.]co[.]uk* the e2LD is *example[.]co[.]uk*, while the second level domain name is *co[.]uk*.

The 12, 212 malicious e2LDs are associated with 174, 112 malware samples from *MAL*. We submit them to *VT* for scanning and find that 98.96% of the samples are known to *VT*, and 99.97% of known samples are marked as malicious by five or more AV vendors. Finally, we expand our dataset by querying *VT* for all malicious samples communicating with the 12, 212 malicious domains. This reflection yields an additional 70, 898 samples, for a total of 245, 010 samples.

Malware Sample Labeling. Different AV vendors offer divergent labels for a malware sample [28]. We use AVClass2 [40] and a malware encyclopedia [33] to resolve these aliases (e.g., *bladabindi* to *njrat*) when possible. We keep the top malware family label by AV vendor agreement and disregard generic labels or labels where the AV vendors cannot agree (`SINGLETON`). Following this methodology, we discard 81,750 samples (33.63%) assigned the label `SINGLETON`. The 161,322 (66.37%) successfully labeled samples represent 202 distinct malware families. No malware families appear to have an outsized representation in our datasets, and we summarize the top 15 malware families by the number of domains in Table 1.

Figure 1 shows the cumulative distribution of the number of malware samples and hosting servers per domain in our dataset. Most domains are associated with only handful of malware samples, with 57% of the domains related to less than three samples. A similar trend holds for the number of servers resolved by a given domain. These distributions are consistent with those in prior large-scale malware measurement studies [18].

	Malware		Server		Client		
Family	Domains	Samples	IPs	CC	Count	CC	Sectors
darkkomet	3,578	16,441	175K	140	2,187K	232	20
njrat	1,924	10,596	195K	129	1,970K	229	21
cybergate	1,181	2,546	38K	100	931K	219	19
xtrat	946	2,801	62K	89	1,108K	222	19
bifrose	700	1,432	11K	62	497K	211	18
razy	667	1,139	107K	110	1,508K	225	18
remcos	563	39,279	61K	103	1,028K	221	18
nanocore	501	2,112	72K	116	1,446K	227	19
ponystealer	450	4,891	49K	93	106K	222	17
gamarue	410	761	53K	97	1,523K	225	19
poison	355	1,018	18K	75	692K	212	18
vobfus	282	3,843	36K	89	936K	219	19
nymeria	279	966	39K	101	838K	215	18
zbot	229	24,736	9K	61	945K	220	20
netwire	228	634	34K	82	859K	223	18

Table 1: Top 15 malware families based on the number of malicious domains in our dataset.

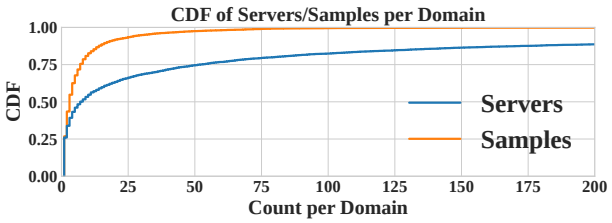


Figure 1: Distribution of the number of malware samples and servers associated with each domain in *AuthDNS*.

Malicious Domain Validation. To validate the maliciousness of the 12, 212 e2LDs, we query *VT* and find that 76.7% of the malicious e2lds have at least one historical URL labeled as malicious. 87.5% of the filtered malware samples only queried domains in our *AuthDNS* and no additional domains. This combination of factors gives us high confidence in our malicious domain dataset.

Figure 2 shows the aggregate daily query volume of malicious domains, as seen in *AuthDNS*. Our vantage point provides a stable view throughout the four years of our study, except for three dips related to collection issues. On average 17.9% of daily requests are ECS-enabled, allowing us to learn the clients' subnets in addition to the IP address of the recursive. Similar to Kountouras et al. [16], we define a client as the client subnet when ECS is enabled and the recursive's IP address when ECS is not enabled. We use this client definition for our experiments in Sections 5 and 6. Finally, we apply *IPWHOIS* and *IND* to servers and clients in order to identify relevant ASNs, organizations, countries, and industry sectors.

3.3 Limitations

Our *AuthDNS* vantage point faces several limitations. We summarize them below and discuss them in more depth in Section 7.

Recursive resolvers. The DNS protocol relies heavily on recursive resolvers, which operate in-between authoritative DNS servers and DNS clients. This indirection makes client estimation difficult. Although many public DNS resolvers support ECS [16, 20], lack of client support for ECS can lead to underestimation. Furthermore, if

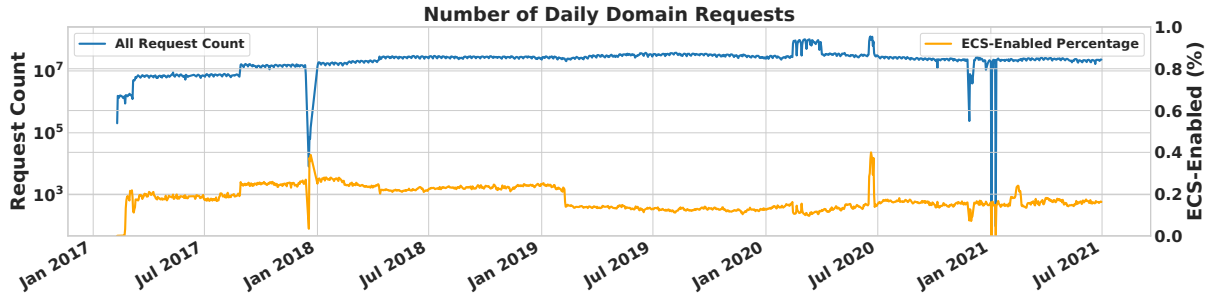


Figure 2: Daily Volume *AuthDNS* for malware domains. ECS-enabled requests (orange) average 17.9% of daily requests.

multiple infected hosts exist within the same ECS network block, *AuthDNS* cannot distinguish between them. Finally, authoritative DNS servers typically see only a portion of the DNS requests issued by individual hosts [30] due to caching by recursives. We do not utilize query volumes. Instead, we focus on the number of unique clients we observe querying for each domain in *AuthDNS* over the four years of our measurement. Due to the aforementioned limitations, our results, even with global perspective, should be viewed as lower bounds on the overall malware ecosystem.

Noisy clients. Not all DNS lookups for malware related domains come from the malware itself. Honeypots and network scanners may query DNS to detect malware-related infrastructure in several cases. This is a common challenge in prior malware ecosystem research that leads to overestimation [13, 38]. We do not address this limitation in Sections 4 and 5, in order to perform meaningful comparisons with established alternative perspectives. However, we begin to address this challenge in Section 6 by examining the different stages of the malware domain lifecycle and identify likely scanners based on signals such as queries that consistently appear after new malware domains are reported/discovered on blocklists.

VPNs and proxies. Clients may utilize VPNs or proxies to hide their true network location. This can skew *AuthDNS*'s geolocation of infected populations. To approximate the presence of proxies and anonymizing networks in our dataset, we measure the prevalence of Tor exit nodes in *AuthDNS* using historical Tor exit node lists[35], accounting for the days that each exit node is active. We find the average daily client percentage and average daily query volume percentage of Tor exit node IPs to be 0.07% and 0.001% respectively: thus, their presence on our dataset is minimal. The low prevalence of anonymizing networks on our dataset does not guarantee the absence of other popular proxy and VPN providers. A lack of well documented historical datasets for proxies/VPNs limits our ability to measure them more thoroughly.

Malware Visibility. The observations in our study are limited by the visibility of our datasets. More specifically, our visibility of malicious domains depends on the *MAL* dataset, which only includes Windows malware. Additionally, we intersect the malicious domains with those registered in our *AuthDNS* dataset, which removes an additional set of malicious domains. Despite these limitations, our study covers more than 200 malware families.

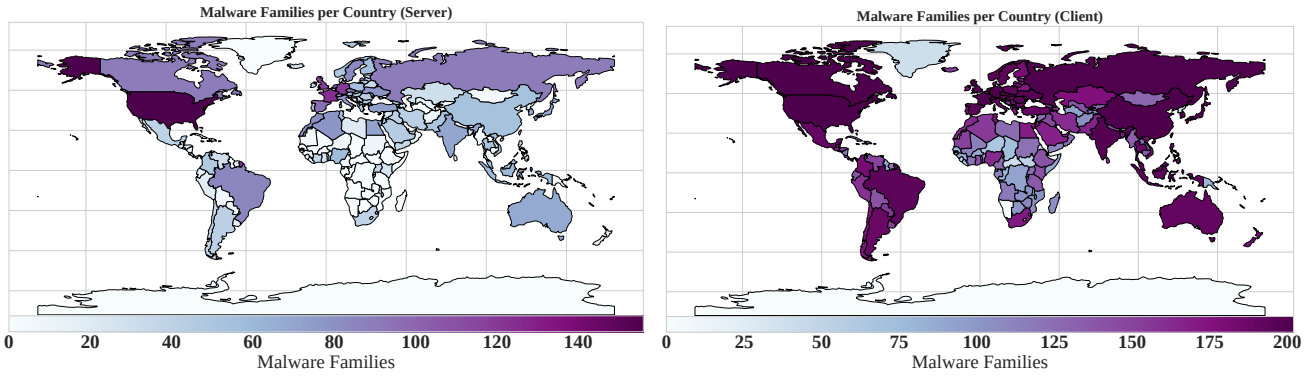
4 HOSTING INFRASTRUCTURE

The hosting infrastructure used by cybercriminals is an essential aspect of malware communication. Understanding how malicious actors distribute and coordinate malware enables the security community to take more effective remediation steps and can focus resources on areas of frequent abuse. To study this infrastructure, we consider a set of 6,400 domains representing the intersection of domains with malware family labels and *IPWHOIS* labels for the IP addresses resolved by those domains. In aggregate, this set of domains point to 399,830 different IP addresses in 151 countries.

First, we consider where malware is hosted. Figure 3a shows a map of all the countries we can associate with infrastructure resolved by malicious domains, with lighter colors indicating fewer malware families. Countries home to large hosting providers—like the United States, France, and Germany—also host large numbers of malware families. Tajalizadehkhoob [42] and Mezzour [23] both found that the distribution of C2 infrastructure on legitimate hosting platforms was strongly correlated with the size of the hosting platform and weakly correlated with their security policies. Our work reiterates that hosting infrastructure may enable malware communication to hide in plain sight. For example, Lever [18] showed that PUP software is often long-lived on legitimate, commercial hosting platforms and found a growing trend of malware samples taking advantage of such hosting.

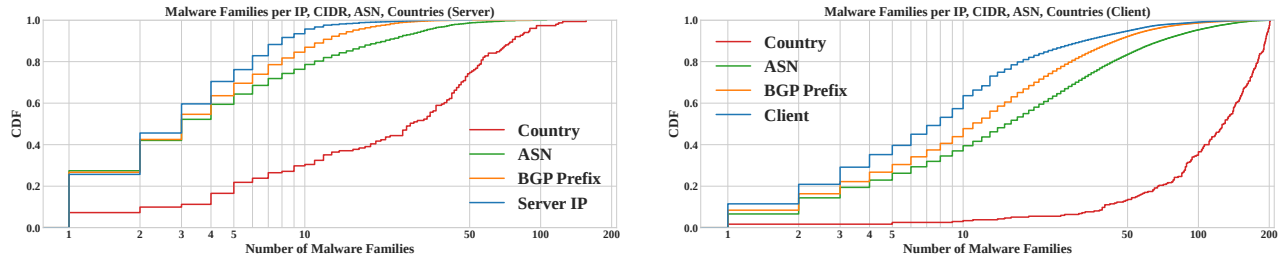
Zooming in, we examine how infrastructure is reused across different malware families. Figure 4a shows the distribution in the number of malware families hosted per country (corresponding to Figure 3a), ASN, network (BGP Prefix), and IP address. We find that only 102,728 (25.7%) of malware-hosting IP addresses were associated with a single malware family. Conversely, 26,226 (6.6%) of IP addresses resolved by malware domains could be tied to ten or more families. In one case, we found that IPs belonging to `AS29075` (`IELO IELO-LIAZO SERVICES SAS`) were pointed to by 715 domains corresponding to 94 malware families. We believe this to be the result of many malicious actors taking advantage of a proxy operated within this French ISP, demonstrating widespread reuse.

Finally, malware families often spread their hosting across multiple countries. We found only 24 malware families have their hosting contained to a single country. To help explain the intra-family diversity of hosting, we looked at the correlation between the number of domains in *AuthDNS* contacted by each malware family and the



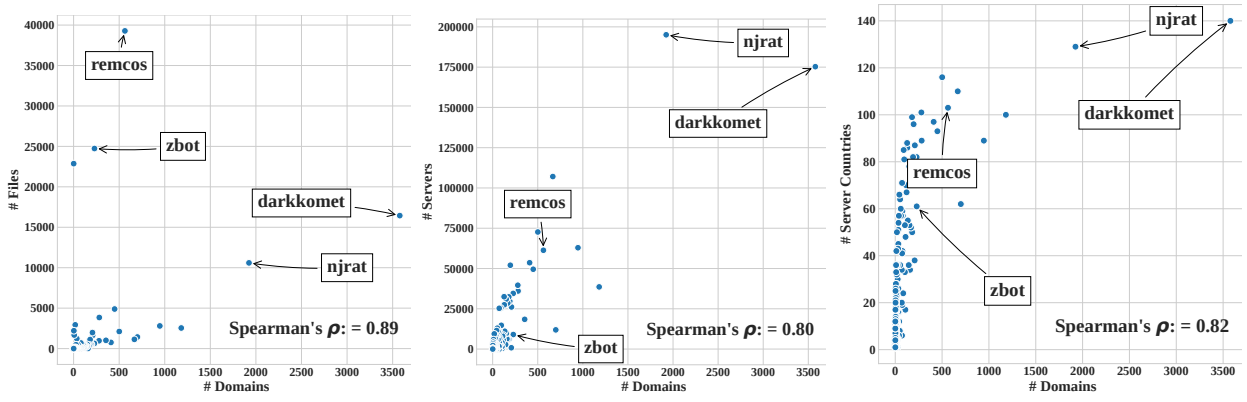
(a) By Hosting Server (b) By Querying Client

Figure 3: Geographic distribution of the IP address of hosting and clients for malware related domains. Darker colors indicate more malware families are associated with the country through the hosting server (left) or querying client (right).



(a) By Hosting Server (b) By Querying Client

Figure 4: Distribution of the number of malware families per IP, Prefix, ASN, and country that have resolved malicious domains in *AuthDNS*. Most network infrastructure and targeted networks are not strongly correlated with a single malware family.



(a) Domains and Files (b) Domains and Server IPs (c) Domains and Server Country Codes

Figure 5: Correlation between different measures of hosting infrastructure. Higher numbers of samples per malware family correlates with more domains. Higher domain utilization correlates with more server IPs and more hosting countries.

number of samples, hosting server IPs, and hosting server countries. Figure 5a shows a strong correlation between the number of domains used by a given malware family and the number of

unique samples in our dataset. Several outliers, such as zbot, for which we observed very few domains but a large number of files, reduced the Pearson correlation. However, the Spearman correlation, which is more tolerant of outliers, still showed a strong correlation. Figure 5b goes on to show a strong correlation between the number of domains contacted by each malware family and the number of hosting servers observed. Further, Figure 5c shows that as the number of domains and hosting IPs increases, so does host country diversity. As the malware family progresses, using more domains or samples, it naturally expands. This breadth of hosting infrastructure contributes significantly to the security community's challenge of attribution and takedowns.

Takeaway-1: *The view of malware-related domain hosting provided by* AuthDNS *largely agrees with prior work which relies upon network data collected at different points in the DNS hierarchy. Infrastructure is reused across different malware families and is often intertwined with legitimate hosting services. Within a malware family, it is common to see many host networks and IPs deployed, often crossing geopolitical boundaries. This agreement between datasets suggests interchangeability; however, a key factor makes* AuthDNS *data superior when available. Non-global vantage points such as RecursiveDNS will only yield snapshots of the hosting infrastructure once customers using that recursive begin querying for a given domain. This may limit visibility during the early stages of a domain's life, particularly before widespread infection by the corresponding malware occurs.* AuthDNS *does not suffer from this limitation.*

5 MALWARE CLIENTS

AuthDNS provides a unique perspective on potential victims who query for malicious domains. In Section 4, many of our findings concerning malicious domain hosting agreed with prior work and could be drawn from other vantage points. The same interchangeability is not valid for studying those infected by malware families. The limitations of previous techniques become apparent when we observe victims through the global perspective of *AuthDNS*.

As discussed in Section 3.2, authoritative nameservers receive DNS requests from recursives rather than individual hosts. This makes the tradeoff of gaining visibility into all querying recursives but gives up visibility into individual endpoints, which recursive DNS provides for a subset of the population. Thus, for non-ECS queries, we consider the IP address of the recursive to be the client, while for ECS-enabled requests, we use the ECS netmask.

Our aim with *AuthDNS* is to study large infected populations as epidemiologists rather than infected individuals as doctors. Figure 3b shows the number of malware families affecting each country, with darker colors representing more malware families. A significant portion of malware families plagues nearly every country. These globally expansive infections contrast with Figure 3a which showed higher concentration levels of malware family hosting in particular countries. Figure 4b zooms in to indicate the number of malware families that are contacted by each network or client.

Viewed from the opposite direction, Figure 6a shows the number of countries each malware family affected with respect to the total number of clients observed. We found that only one family, fosniw, had queries to related domains originating from fewer than ten countries, while 144 (71.3%) of malware families were found to be queried from 100 or more countries. This agrees with Mezzour et al. [23], which also witnessed near-universal affliction by malware in developed countries. Additionally, they found that infection rates correlated strongly with the IT resources of that country. As with the location of hosting infrastructure for malware-related domains, the spread of malware across geopolitical boundaries complicates the security communities' task of identifying the targeting of victims for most malware families. We see that malware families do not generally tend to target specific networks, but rather, many networks appear to be infected by multiple different malware families. Furthermore, malware family infections are not commonly confined by geography as seen from the perspective of *AuthDNS*.

5.1 Industry Sectors

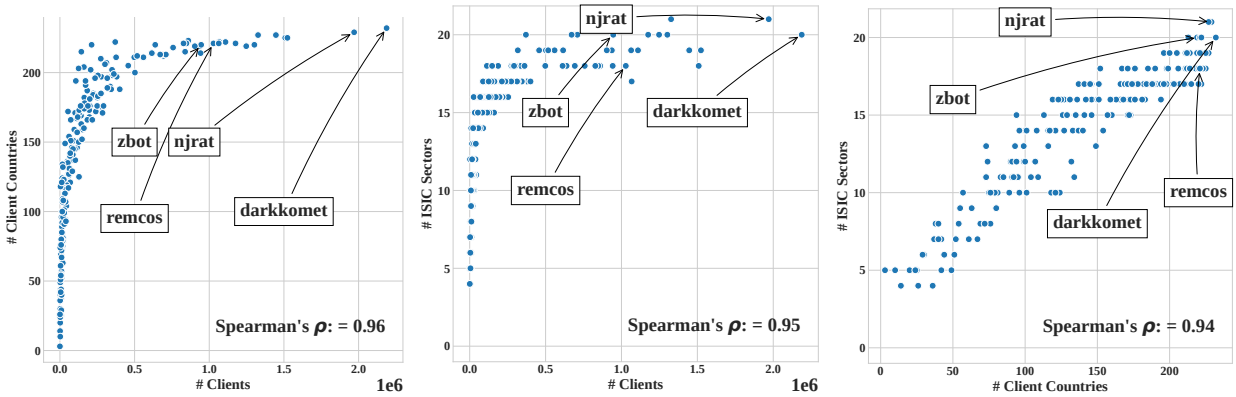
ISIC Section	Clients	Malware Families
Information & Communication	3,108,546	202
Wholesale & Retail Trade	567,729	202
Education	29,741	201
Professional, Scientific & Technical Activities	11,576	196
Manufacturing	4,837	192
Government, Defence	4,697	178
Financial & Insurance Activities	3,670	183
Human Health & Social Work Activities	3,785	172
Accommodation & Food Service Activities	2,785	148
Transportation and Storage	624	155
Arts, Entertainment & Recreation	421	140
Electricity, Gas, Steam & A/C Supply	333	127
Administrative and Support Service Activities	199	141
Extraterritorial Organizations and Bodies	164	120
Other Service Activities	149	149
Real Estate Activities	96	86
Construction	74	38
Mining and Quarrying	17	23
Agriculture, Forestry and Fishing	5	18
Water Supply, Sewerage.	5	8

Table 2: Unique clients querying malicious domains and number of malware families in each industry (ISIC section).

ISIC Section	2017		2018		2019		2020	
	All	Mal	All	Mal	All	Mal	All	Mal
Information & Communication	0.85M	68K	1.4M	121K	1.3M	136K	1.3M	117K
Wholesale & Retail Trade	5.2K	1.0K	14K	3.4K	16K	6.8K	29K	10K
Education	19K	1.5K	27K	2K	25K	2.6K	29K	1.9K
Professional, Scientific & Technical	4.6K	402	8.8K	1.1K	6.5K	1.1K	7.7K	1.1K
Manufacturing	2,109	181	3.2K	387	2.9K	383	2.7K	246
Government, Defence	3.1K	215	5.2K	347	4.3K	384	4.5K	310
Human Health & Social Work	2.1K	129	3.5K	213	3.3K	228	3.5K	163
Financial & Insurance	2.7K	157	4.1K	267	3.6K	271	3.7K	214
Accommodation & Food Service	1.5K	69	2.6K	103	2.2k	131	2.3K	75
Transportation & storage	250	34	379	58	317	59	309	25
Arts, Entertainment & Recreation	342	17	554	32	510	29	493	16
Electricity, Gas, Steam & A/C Supply	224	19	316	37	269	39	299	26
8 remaining sections	494	29	771	52	757	56	773	52
Correlation (Spearman):	0.98	0.99	0.98	0.99	0.98	0.99	0.98	0.98

Table 3: Client distribution across sectors for seven day samples starting 2017-03-01, 2018-03-01, 2019-03-01, and 2020-03-01. *All* represents the complete *AuthDNS* dataset while *Mal* represents only malicious domains.

Another way of grouping clients is by the type of network they query from. Kotzias et al. found evidence that different industries



(a) Client Count and Impacted Countries (b) Client Count and Impacted Sectors (c) Impacted Countries and Impacted Sectors

Figure 6: Spearman correlation between different measures of potential victims. Higher numbers of clients querying for malware family related domains correlate strongly with a more diverse set of impacted countries and economic sectors.

are affected by different amounts of malware samples [15]. In that study, the authors relied on file reputation logs collected from endpoint protection software to study malware affecting customers of a large cybersecurity company. They found that specific industries were affected by a disproportionate number of malware, suggesting targeting by malware families and disparity in security posture across industries. However, their vantage point was limited to customers of the cybersecurity company, which they acknowledge introduces bias. We seek to augment this work by studying how malware families infect industries from a global vantage point.

For this analysis, we use the industries (IND) discussed in Section 3.1. This mapping enables us to group clients by ISIC code. ISIC provides a hierarchical classification of industries that break down into 21 sections, 88 divisions, 238 groups, and 419 classes. We use the ISIC section synonymously with industry sector in the remainder of this work.

We observed around 39.7B requests for malicious domains in our dataset, and were able to assign an industry label for approximately 28.08B (70.7%) requests. We only consider requests from clients to whom we can assign an industry label.

Table 2 shows the number of clients we observed in each sector as well as the number of offending malware families. For clarity, we rename some of the ISIC code classification labels. We can immediately see that each industry contains clients querying domains associated with numerous different malware families. In fact, 15 of the ISIC sections appear to be impacted by over half of the malware families in our dataset. We note that the `Information & Communication` ISIC section, as well as `Wholesale & Retail Trade` section, contains requests from all the malware families and represent more clients than any of the remaining industries by several orders of magnitude. From the more granular ISIC divisions, we see that most queries from the `Information & Communications` industry can be attributed to wired and wireless communications due to the classification `Internet Service Providers (ISP)` and the `Residential & Business Hosting Infrastructure`. A portion of the IP address space controlled by Amazon is labeled as `Wholesale & Retail Trade`, contributing to an overestimation of the effects on this population. This also explains why these two sections contain several orders of magnitude more requests compared to other sections.

To capture the representation of clients querying for malicious domains compared to all clients in *AuthDNS*, we sampled a seven-day window for each year in our study. Table 3 shows the number of clients from top ISIC sections querying for any (malicious or benign) domain during this window and the subset querying for malicious domains. The final row shows the correlation (Spearman) between the sampled datasets and the dataset of malicious domains that spans the complete four-year study.

The next largest ISIC section by number of unique clients is `Education` with roughly half of the requests in this section coming from institutes of higher education such as colleges and universities. These institutional networks typically have a wide variety of users, including students, staff, faculty, and visitors. Many such networks may not have direct control over the devices on their network. A heterogeneous base of infected devices and research related activities provide a sensible explanation for why `Education`, and higher education in particular, accounts for so many malware related queries. While some of the remaining ISIC sections seem like prime candidates for targeted behavior, we note that even sections associated with the government, defense, finance, and infrastructure appear to be impacted by many different malware families.

We find that malware families generally impact multiple ISIC sections, with 72.7% of the malware families found in more than 10 sectors. Our aggregate analysis with global visibilty cannot draw the same conclusions as Kotzias et al. [15], which found that 1,911 (37%) of the malware families in their study were only seen in one enterprise. Instead, Figures 6b and 6c respectively show that the number of industries a malware family impacts is correlated with the overall number of clients impacted by that malware and the geographic diversity of those clients. As malware families grow, so do the diversity of their victims.

Takeaway-2: *In aggregate, we do not see malware families solely affecting individual industries. 72.7% of the malware families in our data are found to affect more than 10 distinct industry sections. While*

datasets derived from RecursiveDNS or host-based security products offer a view into the networking behavior of individual end-users, they can introduce biases by considering customers in a particular geographic region or those already taking steps to mitigate their online risk. Our study of clients affected by a range of malware families highlights how AuthDNS*'s global vantage point can reduce these biases and lead us to draw divergent conclusions when studying malware infections from an epidemiological standpoint. Still,* AuthDNS *leaves ample room for studies such as Kotzias et al. [15] that provide greater visibility into individual infected hosts for a subset of the population once these potential biases are placed in the context of a global view. While* AuthDNS *does not provide visibility into end-users, it does offer a complete view of recursives querying domains under that authority. ECS-enabled requests further narrow this gap when looking at affected clients through the lens of* AuthDNS.

6 MALWARE LIFECYCLE

Utilizing the unique vantage point of *AuthDNS*, we perform a temporal analysis in order to understand the lifecycle of malicious domains. We complement the client visibility of previous studies that observed malicious domains after expiration [38] by considering all clients querying for a malicious domain name during three phases: registration to detection, detection to expiration, and post expiration. We determine the date of detection as the earliest of the following dates: a malicious URL of the domain is detected by more than one vendor in VT, a malicious hash communicating with that domain is detected in VT, a malicious hash communicating with that domain is seen in our malware DNS dataset. In order to fully observe the domain lifecycle, we only consider domains that were registered after the first day of visibility we have in *AuthDNS*. Further, we restrict our analysis to domains that have been registered only once in our *AuthDNS* dataset so that we do not observe noise from previous or subsequent registrations as domain names get repurposed. This filtering leaves us with 2,308 domain names, 18.9% of the total domains in our dataset.

Table 4 summarizes the networks for the lowest 10%, lower quartile, median, upper quartile, 90% and max of querying clients during all phases of the domain lifecycle. The registration to detection window is relatively short, lasting 19 days or less for 75% of domains. Additionally, at the median, only six networks (ASNs) and three countries queried for domains while they were in this initial phase. By comparison, the second temporal window, detection to domain expiration/takedown is significantly longer, with at least 23 days representing the lower quartile. At the median, 54% of ASNs that will ultimately query for a domain do so for the first time during this window. The same observation holds for the countries and industry sectors of these clients. Finally, queries continue during the post-expiration/takedown period, which continues until the end of our four-year *AuthDNS* dataset for domains that are not re-registered. Interestingly, in this period, the median domain observes more than 76 unique ASNs and ten countries querying it for the first time. This represents a long tail of unique clients first seen only after a domain has expired or been taken down.

In order to understand the most popular networks in each lifecycle phase, we look at the top querying ASNs across domains. Table 5 shows the top five unique ASNs as seen by the number of first occurrences in each temporal window. During the registration to detection window, we first observe large hosting networks (Amazon), large recursives (Google), and large telecommunication companies (Vimpelcom and Level3). This window is related to the setup and testing of the domains by the actors and the first potential victim connections, resulting in queries from large recursives and telcos. After the domain's detection, the most common ASNs to be first observed are large scanners (GEORGIA-TECH [17]), AV companies (MFENET - McAfee and PAN0001 - PALO ALTO NETWORKS), and other large hosting networks (WINTEK-CORP and OVH), which can contain other scanners. In this window, AVs, sandboxes, and scanners query malicious domains and map their IP address space. Lastly, in the final window, post-expiration/takedown, we observe large Chinese telcos and business networks from other countries. These post-expiration queries could be due to network mobility of infected clients, new infections, or scanning.

Takeaway-3: *The view provided by* AuthDNS *shows that researchers need to consider a domain's lifecycle to measure infected populations accurately. Most domains in our dataset were detected as malicious soon after registration, with the median time being four days. After detection, domains will receive increased interest from scanners and AV vendors, which can artificially inflate infected population counts if proper filtering is not applied. Notably, there commonly exists a long tail of new client queries after a domains' expiration or takedown. Existing infections on mobile clients generate queries from new networks and may persist into this final phase. However, prior studies further suggest that scanning activity late in the lifecycle of a domain may constitute a significant portion of queries [38]. Researchers and practitioners using network data,* AuthDNS *or otherwise, to estimate client infections risk obscuring malware behavior when they do not distinguish between phases of the domain lifecycle. As a community, there is room for further improvement in identifying scanners and distinguishing the lifecycle phases for domains with multiple registrations. Addressing these challenges will allow researchers to better understand and help infected populations.*

7 VANTAGE POINT COMPARISON

Thus far, we have shown *AuthDNS*'s ability to 1) reproduce previous observations of malware infrastructure (Section 4), 2) add a novel perspective on the distribution of malware infections (Section 5), and 3) introduce a full temporal view of the malware domain lifecycle (Section 6). In this section, we synthesize these findings and contextualize them in the broader landscape of malware ecosystem and epidemiology research. We first enumerate related work and map the relationships between different perspectives. We then compare the perspectives along four different qualitative characteristics and highlight the appropriate role of each perspective, gaps in existing malware visibility, and avenues for future research.

7.1 Measurement Planes

Broadly speaking, malware utilizes three distinct network planes[2] (Figure 7), which we define as a grouping of network components based on their location and functionality within the network.

[2]Unrelated to control/data planes from software defined networking.

Domains	Registration to Detection							Detection to Expiration/Takedown							Post Expiration/Takedown						
	ASNs	(%)	ASCCs	(%)	Sectors	(%)	Days	ASNs	(%)	ASCCs	(%)	Sectors	(%)	Days	ASNs	(%)	ASCCs	(%)	Sectors	(%)	Days
10%	0	(0.00)	0	(0.00)	0	(0.00)	0	19	(11.6)	5	(11.5)	0	(00.0)	1	10	(6.80)	1	(1.02)	0	(0.00)	238
25%	0	(0.00)	0	(0.00)	0	(0.00)	1	59	(36.2)	15	(42.9)	2	(37.5)	23	36	(21.9)	3	(9.09)	0	(0.00)	526
50%	6	(2.63)	3	(7.69)	1	(20.0)	4	101	(54.0)	26	(62.5)	4	(62.5)	30	76	(38.6)	10	(22.7)	1	(16.7)	963
75%	22	(10.0)	9	(21.8)	2	(37.5)	19	164	(70.4)	37	(79.0)	5	(80.0)	100	132	(55.3)	18	(40.0)	2	(30.0)	1,180
90%	54	(23.4)	18	(40.0)	3	(57.1)	79	369	(86.5)	54	(90.4)	7	(100)	419	229	(71.3)	28	(58.6)	3	(50.0)	1,256
max	2,243	(96.6)	136	(100)	14	(100)	1,154	11,650	(100)	187	(100)	15	(100)	1,661	4,644	(100)	95	(100)	10	(100)	1,558

Table 4: Request characterization for three temporal windows. While most clients and sectors are observed between malicious domain detection and expiration/takedown, many client ASNs and countries (ASCCs) first connect after expiration/take down.

Registration to Detection		Detection to Expiration/Takedown		Post Expiration/Takedown	
ASNAME	Domains	ASNAME	Domains	ASNAME	Domains
AMAZON-AES	772	WINTEK-CORP	1,745	CNNIC-ALIBABA-US-NET-AP Alibaba (US) Technology Co., Ltd.	1,387
CORBINA-AS PJSC "Vimpelcom"	756	GEORGIA-TECH	1,738	CNIX-AP China Networks Inter-Exchange	1,363
GOOGLE	645	OVH OVH SAS	1,662	CHINATELECOM-TIANJIN Tianjij,300000	1,226
LEVEL3	611	MFENET	1,649	InterConnect ML Consultancy	1,114
AMAZON-02	602	PAN0001	1,641	FSOL-AS F-Solutions Oy	1,084

Table 5: Most popular ASNs first observed in each temporal window. Scanners and AV vendors appear mostly during and after detection of a malicious domain while hosting networks are most prevalent during the setup of the domain.

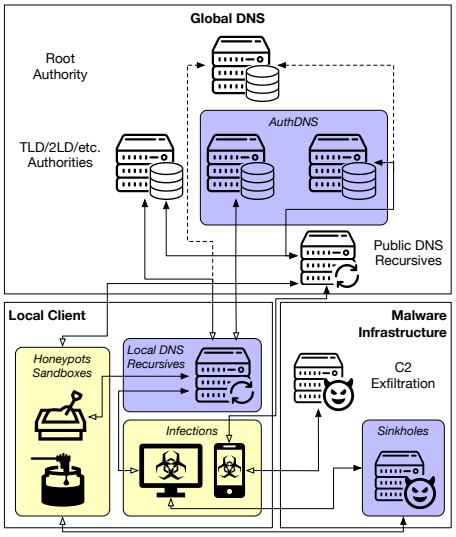


Figure 7: Malware measurement can occur 1) in the global DNS plane, 2) at or near the local client, or 3) at the malware infrastructure. Each location has specific sub-components that interact via request (filled arrow) and response (empty arrow) protocols (e.g., DNS, C2 protocols). Existing research has studied many of the depicted components with host-based (yellow) or network-based (blue) techniques.

Global DNS DNS is the bootstrapping protocol used by most network communication to map domain names to IP locations; malware uses DNS extensively, as evidenced by countless domain blocklists and DGA malware. Global DNS refers to upper-hierarchy authoritative DNS servers (e.g., root, TLD, 2LD) and large public recursives (e.g., Google, CloudFlare), which share a global perspective on domain lookups. Global DNS servers can receive DNS requests from a universal set of clients and have worldwide visibility into domain usage. Prior work [6, 12, 44] utilizing global DNS authorities within the realm of malware has focused on detection, spam measurements, and one work [37] performed a case study on stalkerware based on probing of large public recursives.

Local Client The local client plane consists of malware-infected clients and the local networks in which they reside. In contrast to global DNS, which only has network-based techniques, the local client plane contains both host-based and network-based approaches. Host-based approaches include any measurements that directly observe partial or full execution of malware: interactive honeypots, malware sandboxes, and in-the-wild infections. Existing malware research has skewed heavily towards host-based analysis of the client plane. As a brief example amongst a profusion of works, sandboxes and honeypots have been used to study general Windows malware [8, 29], malware downloaders [39], Android applications [19], malware protocol reverse engineering [32, 48], exploit sites [36, 45], C2 hosting [7, 27, 42], and DDoS [10]. Two works have utilized host-based local client techniques to shed light on the broader malware ecosystem. Kotzias et al. [15] applied host-based infection measurement across 28K enterprises in 67 industries to determine enterprise malware trends. Messour et al. [23] conducted an empirical analysis of Symantec's telemetry data to observe the distribution of different malware types across countries.

The primary network-based approach within the local client plane is the collection of DNS data from local recursives that handle a network's DNS traffic and are often set by default via Dynamic Host Configuration Protocol (DHCP). Alrawi et al. [4] used recursive passive DNS data to estimate infections by IoT malware, while Lever et al. [18] focused on more general malware, including PUPs.

Malware Infrastructure Malware relies on infrastructure most commonly for command and control (C2), but can also use separate infrastructure for hosting, data exfiltration, or other functions. Measurement of malware infrastructure IP addresses can occur from

global DNS and local client planes, but to collect communication data between infections and malware infrastructure, researchers have developed sinkholes. Sinkholes allow a researcher to operate or imitate malware infrastructure and collect richer data about connecting clients. Several works have utilized sinkholes to study specific phenomenon (e.g., remote-access trojans (RAT) [38], botnets [41]); one prior work by Alowaisheq et al. [3] studied sinkhole domain behavior across all types of malware, but did not operate any sinkholes, since they require malware-specific configuration.

7.2 Comparison

Infection Visibility Infection visibility is the capability of a vantage point to assess all infections of a threat globally and temporally. This work shows *AuthDNS* datasets yield high global infection visibility as they provide access to all DNS requests made to a malicious domain, across all locations and time. Datasets that are not based on network infrastructure are limited in infection visibility as they can only observe a subset of clients based on data source (e.g., AV vendors, ISP clients, recursive clients, email clients). Domain sinkholes provide global location visibility, but partial temporal visibility, as they are limited to the post-expiration period of a domain. Infrastructure takeover can provide global and temporal infection visibility guarantees; however, this is difficult to execute and scale.

Infection Precision Infection precision is the capability of the dataset to accurately estimate the validity and type of infections. Passive DNS datasets contain noisy infection data that is muddled with traffic from scanners, malware sandboxes, or security professionals. Thus, users of passive DNS datasets should filter clients based on their behavior and network origin when estimating infections. Additionally, many different malware samples and families can be hosted on the same domain name and the type of infection per client cannot be guaranteed. Client-side antivirus datasets provide higher infection precision for the type of client and the existence of a specific malware sample; infrastructure takeover datasets can provide the highest precision by looking at the collected infected system data. Lastly, domain sinkholing initially provides partial visibility, since a domain will continue to receive queries after its detection period, but sinkholing data can be enhanced for a better infection estimation as shown by Rezaeirand et al. [38]. Email datasets provide the lowest precision as they observe the targeting aspect of an attack rather than the infection.

Client Granularity Client granularity is the capability of the dataset to trace the infections down to single clients or users. Authoritative pDNS datasets are limited in this regard since clients are obscured by recursive DNS servers and caching. However, as shown in this study, researchers can use the ECS field of an ECS-enabled request to obtain higher precision client granularity. Recursive pDNS datasets yield even higher client granularity as they can observe all the clients under the recursive making requests for a malicious domain. Client-side AV datasets, infrastructure takedown datasets, ISP network logs, and email datasets provide high guarantees of client granularity as they can observe a unique client or user.

Malicious Infrastructure Visibility Malicious infrastructure visibility is the capability of the dataset to observe what infrastructure the malware actors have used to perform their campaign. *AuthDNS* is, by definition, the authoritative source of the mapping of domains to IPs for a malicious domain. DNS rewriting, for the profit of the recursive operator [46], or for the protection of customers, may limit the hosting infrastructure visibility provided by a RecursiveDNS dataset. Infrastructure takedowns provide the highest guarantees as they provide direct access to infrastructure; however, it is not scalable. Infrastructure visibility via recursive DNS and ISP network logs depends on the volume and consistency of communications by infected clients within the measured networks. The remaining datasets cannot provide any insights regarding the infrastructure used by the malicious actors.

Takeaway-4: AuthDNS *has several advantages and disadvantages when compared to vantage points used in previously published research. We find global, temporal and infrastructure visibility to be the biggest advantages of our dataset. Thus, we position our measurements along these advantages and we study each aspect in depth in Sections 4, 5 and 6 and report our most insightful results.* AuthDNS *has limited client granularity and limited infection precision. Future work can be aimed to address this issue.*

8 CONCLUSION

Understanding malware lifecycles is vital in the fight against Internet threats. This work presents a longitudinal study analyzing the network communication of 202 different malware families from the perspective of a popular authoritative DNS server. We observed billions of resolutions over four years at our authoritative collection point, enabling temporally complete and global visibility into malicious domain usage. *AuthDNS* simultaneously solidifies prior findings while also shedding new light on the epidemiology of malware. First, different malware families often re-use the same network infrastructure, so threat intelligence needs to label malicious infrastructure cautiously. Second, malware families, when analyzed in aggregate from an *AuthDNS* vantage point, do not appear to target specific networks or industries. Instead, they spread to many different industries with high regularity over time. Third, our temporal analysis shows that newly registered malicious domains are set up and detected quickly. Due to network noise from scanners and AV vendors, both the temporal and organizational properties of network clients should be considered when estimating malware infections from a network perspective. Finally, we introduce a brief taxonomy of malware measurement perspectives and discuss the advantages and disadvantages across four primary measurement goals. By broadening our understanding of global malware infections, this work serves as a stepping stone to making malware characterization more accurate and, ultimately, to make mitigation more effective.

REFERENCES

[1] 2019. UNSD - Statistical Classifications. https://unstats.un.org/unsd/classifications.
[2] 2022. VirusTotal. https://www.virustotal.com.
[3] Eihal Alowaisheq, Peng Wang, Sumayah Alrwais, Xiaojing Liao, XiaoFeng Wang, Tasneem Alowaisheq, Xianghang Mi, Siyuan Tang, and Baojun Liu. 2019. Cracking the wall of confinement: Understanding and analyzing malicious domain take-downs. In *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS 19)*.
[4] Omar Alrawi, Charles Lever, Kevin Valakuzhy, Kevin Snow, Fabian Monrose, Manos Antonakakis, et al. 2021. The Circle of life: A {large-scale} study of the

{IoT} malware lifecycle. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*.

[5] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. 2017. Understanding the Mirai botnet. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security 17)*.

[6] 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 Security Symposium (USENIX Security 11)*.

[7] Athanasios Avgetidis, Omar Alrawi, Kevin Valakuzhy, Charles Lever, Paul Burbage, Angelos Keromytis, Fabian Monrose, and Manos Antonakakis. 2023. Beyond the gates: An empirical analysis of HTTP-managed password stealers and operators. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*.

[8] Ulrich Bayer, Imam Habibi, Davide Balzarotti, Engin Kirda, and Christopher Kruegel. 2009. A view on current malware behaviors. In *Proceedings of the 2nd USENIX Conference on Large-scale Exploits and Emergent Threats (LEET 09)*.

[9] CAIDA. 2022. Routeviews Prefix-to-AS mappings (pfx2as) for IPv4 and IPv6. http://data.caida.org/datasets/routing/routeviews-prefix2as/.

[10] Wentao Chang, Aziz Mohaisen, An Wang, and Songqing Chen. 2015. Measuring botnets in the wild: Some new trends. In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*.

[11] Carlo Contavalli, Wilmer Van Der Gaast, D Lawrence, and Warren Kumari. 2016. Client subnet in DNS queries. RFC 7871 (Informational). http://www.ietf.org/rfc/rfc7875.txt

[12] Shuang Hao, Nick Feamster, and Ramakant Pandrangi. 2011. Monitoring the initial DNS behavior of malicious domains. In *Proceedings of the 2011 ACM Internet Measurement Conference (IMC 11)*.

[13] Chris Kanich, Kirill Levchenko, Brandon Enright, Geoffrey M Voelker, and Stefan Savage. 2008. The heisenbot uncertainty problem: Challenges in separating bots from chaff.. In *Proceedings of the 1st USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET 08)*.

[14] Amin Kharraz, Sajjad Arshad, Collin Mulliner, William Robertson, and Engin Kirda. 2016. UNVEIL: A large-scale, automated approach to detecting ransomware. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security 16)*.

[15] Platon Kotzias, Leyla Bilge, Pierre-Antoine Vervier, and Juan Caballero. 2019. Mind Your Own Business: A longitudinal study of threats and vulnerabilities in enterprises. In *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS 19)*.

[16] Athanasios Kountouras, Panagiotis Kintis, Athanasios Avgetidis, Thomas Papastergiou, Charles Lever, Michalis Polychronakis, and Manos Antonakakis. 2021. Understanding the growth and security considerations of ECS.. In *Proceedings of the 2021 Network and Distributed System Security Symposium (NDSS 21)*.

[17] Athanasios Kountouras, Panagiotis Kintis, Chaz Lever, Yizheng Chen, Yacin Nadji, David Dagon, Manos Antonakakis, and Rodney Joffe. 2016. Enabling network security through active DNS datasets. In *Proceedings of the International Symposium on Research in Attacks, Intrusions, and Defenses*. Springer.

[18] Chaz Lever, Platon Kotzias, Davide Balzarotti, Juan Caballero, and Manos Antonakakis. 2017. A lustrum of malware network communication: Evolution and insights. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (S&P)*.

[19] Martina Lindorfer, Matthias Neugschwandtner, Lukas Weichselbaum, Yanick Fratantonio, Victor Van Der Veen, and Christian Platzer. 2014. Andrubis–1,000,000 apps later: A view on current Android malware behaviors. In *Proceedings of the 3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. IEEE.

[20] Baojun Liu, Chaoyi Lu, Haixin Duan, Ying Liu, Zhou Li, Shuang Hao, and Min Yang. 2018. Who is answering my queries: Understanding and characterizing interception of the DNS resolution path. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*.

[21] Malwarebytes. 2022. 2022 Global Threat Report. https://go.crowdstrike.com/rs/281-OBQ-266/images/Report2022GTR.pdf.

[22] Malwarebytes. 2022. 2022 Threat Review. https://www.malwarebytes.com/resources/malwarebytes-threat-review-2022/mwb_threatreview_2022_ss_v1.pdf.

[23] Ghita Mezzour, Kathleen M Carley, and L Richard Carley. 2017. Global variation in attack encounters and hosting. In *Proceedings of Hot Topics in Science of Security: Symposium and Bootcamp*. ACM.

[24] P.V. Mockapetris. 1987. Domain names - concepts and facilities. RFC 1034 (INTERNET STANDARD). http://www.ietf.org/rfc/rfc1034.txt Updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936.

[25] Paul Mockapetris and Kevin J Dunlap. 1988. Development of the domain name system. In *Symposium Proceedings on Communications Architectures and Protocols*.

[26] Paul V Mockapetris. 1987. Rfc1035: Domain names-implementation and specification.

[27] Abedelaziz Mohaisen and Omar Alrawi. 2013. Unveiling zeus: Automated classification of malware samples. In *Proceedings of the 22nd International Conference on World Wide Web*.

[28] Aziz Mohaisen and Omar Alrawi. 2014. Av-meter: An evaluation of antivirus scans and labels. In *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer.

[29] Aziz Mohaisen, Omar Alrawi, and Manar Mohaisen. 2015. AMAL: High-fidelity, behavior-based automated malware analysis and classification. *computers & security* 52 (2015).

[30] Giovane CM Moura, John Heidemann, Ricardo de O Schmidt, and Wes Hardaker. 2019. Cache me if you can: Effects of DNS time-to-live. In *Proceedings of the 2019 ACM Internet Measurement Conference (IMC 19)*.

[31] Giovane CM Moura, Moritz Müller, and Marco Davids. 2015. Domain names abuse and TLDs: From monetization towards. In *Proceedings of the 2015 ACM Internet Measurement Conference (IMC 15)*.

[32] Antonio Nappa, Zhaoyan Xu, M Zubair Rafique, Juan Caballero, and Guofei Gu. 2014. Cyberprobe: Towards internet-scale active detection of malicious servers. In *Proceedings of the 2014 Network and Distributed System Security Symposium (NDSS 14)*.

[33] Daniel Plohmann, Martin Clauss, Steffen Enders, and Elmar Padilla. 2017. Malpedia: A collaborative effort to inventorize the malware landscape. *Botconf* (2017).

[34] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2018. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156* (2018).

[35] The Tor Project. 2022. TorDNSEL's exit lists. https://metrics.torproject.org/collector/archive/exit-lists/.

[36] N Provos, P Mavrommatis, MA Rajab, and F Monrose. 2008. All your iFRAMEs point to us. In *Proceedings of the 17th USENIX Security Symposium (USENIX Security 08)*.

[37] Audrey Randall, Enze Liu, Gautam Akiwate, Ramakrishna Padmanabhan, Geoffrey M Voelker, Stefan Savage, and Aaron Schulman. 2020. Trufflehunter: Cache snooping rare domains at large public DNS resolvers. In *Proceedings of the 2020 ACM Internet Measurement Conference (IMC 20)*.

[38] Mohammad Rezaeirad, Brown Farinholt, Hitesh Dharmdasani, Paul Pearce, Kirill Levchenko, and Damon McCoy. 2018. {Schrödinger's}{RAT}: Profiling the stakeholders in the remote access trojan ecosystem. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security 18)*.

[39] Christian Rossow, Christian Dietrich, and Herbert Bos. 2012. Large-scale analysis of malware downloaders. In *Proceedings of the International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer.

[40] Silvia Sebastián and Juan Caballero. 2020. Avclass2: Massive malware tag extraction from av labels. In *Proceedings of the 2020 Computer Security Applications Conference*.

[41] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. 2009. Your botnet is my botnet: Analysis of a botnet takeover. In *Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS 09)*.

[42] Samaneh Tajalizadehkhoob, Carlos Gañán, Arman Noroozian, and Michel van Eeten. 2017. The role of hosting providers in fighting command and control infrastructure of financial malware. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*.

[43] Georgia Tech. 2022. GT malware passive DNS data daily feed. https://impactcybertrust.org/dataset_view?idDataset=520.

[44] Matthew Thomas and Aziz Mohaisen. 2014. Kindred domains: Detecting and clustering botnet domains using DNS traffic. In *Proceedings of the 23rd International Conference on World Wide Web*.

[45] Yi-Min Wang, Doug Beck, Xuxian Jiang, and Roussi Roussev. 2006. Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities. In *Proceedings of the 2006 Network and Distributed System Security Symposium (NDSS 06)*.

[46] Nicholas Weaver, Christian Kreibich, and Vern Paxson. 2011. Redirecting {DNS} for Ads and Profit. In *Proceedings of the 2011 USENIX Workshop on Free and Open Communications on the Internet (FOCI 11)*.

[47] Duane Wessels, Marina Fomenkov, Nevil Brownlee, et al. 2004. Measurements and laboratory simulations of the upper DNS hierarchy. In *Proceedings of the 2004 International Workshop on Passive and Active Network Measurement*. Springer.

[48] Zhaoyan Xu, Antonio Nappa, Robert Baykov, Guangliang Yang, Juan Caballero, and Guofei Gu. 2014. Autoprobe: Towards automatic active malicious server probing using dynamic binary analysis. In *Proceedings of the 21st ACM Conference on Computer and Communications Security (CCS 14)*.

[49] Maya Ziv, Liz Izhikevich, Kimberly Ruth, Katherine Izhikevich, and Zakir Durumeric. 2021. ASdb: A system for classifying owners of autonomous systems. In *Proceedings of the 2021 ACM Internet Measurement Conference (IMC 21)*.