Array 14 (2022) 100174

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

Analyzing temporal graphs of malware distribution networks

Jose Andre Morales *, Yang Cai

Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA, 15213, USA

ARTICLE INFO

Keywords:
Malware
Malware distribution networks
Malware attribution
Malware detection
Invasive software

ABSTRACT

This research provides temporal insight on network topological structures, as well as transitional properties and malware attribution of malware distribution networks. This is accomplished with a temporal-based data set created using a novel data fusion of publicly-available data sources. We developed and used a crawler, along with public APIs, to collect publicly-available data of malicious top-level domains and relevant hosted malware from Google Safe Browsing and VirusTotal for an eight-month period between 19 January and 25 September of 2017. We then combined these data sources, which revealed new insight on fully qualified domain name topological network structure properties and temporal transitions not appreciable from the individual data sources. We have provided the technical details of our novel data fusion approach to GSB and VT static data. The result of this data fusion was the creation of new observable knowledge, primarily temporal-based structural changes and malware attribution within the distribution network, which is not available by analyzing the static data of GSB and VT in isolation. Data revealing details of malware-hosting on a domain brought to light topological structures of fully-qualified domain names involved in the distribution of malicious files. Our insights include: 1) malware distribution networks form clusters that follow the Power law, 2) network structure components such as bridges and hubs (both concepts presented and defined in this paper), and URL shortening providers serve significant roles in malware distribution dynamics, 3) persistence of fully-qualified domain names in malware distribution is random and often used only once to host malware, 4) a large number of unique, downloaded malicious files hosted on various nodes in a malware distribution network were found to belong to a much smaller set of malware families. These observed insights revealed continued persistence of surrounding topological structures. These topological structures were streaming malicious data flows to fully-qualified domain names identified as actively hosting malware before and after the date of identification. The insights further suggest large topological structures with data flow distributing malware persist over time with small sub-structural changes. We have provided suggestions on preventing sustained malicious data flows based on our temporal observations of bridge and hub structures. Individual persistent fully-qualified domain names within these large structures repeatedly served as either a source or an intermediate node of malicious data flows. This implies that the continued monitoring of data flows can serve to alert early-stage malware distribution.

1. Introduction

A malware distribution network (MDN) is a set of connected fully qualified domain names (FQDN) used to globally disperse malicious files for the purpose of infecting and compromising systems. In this paper, we perform a temporal topological analysis of an MDN with malware attribution focusing on subsets of connected FQDNs which we refer to as a malicious cluster (M-Cluster). We used FQDNs consisting of second and top-level domains such as example.com. We created a novel data set over two continuous quadrimesters from 19 January to 25 September of 2017. We queried the transparency report repository of Google Safe Browsing (GSB) [1] by submitting information requests for multiple FQDNs. We graphed the daily topological structure of an MDN with the novel approach of crawling static GSB transparency reports. This approach of crawling focused on specific transparency report labels which indicated the direction of data flow of malicious traffic to other FQDNs. We attributed malware by submitting information requests to VirusTotal.com (VT) [2] for FQDNs in our GSB data set, which was determined by GSB to be hosting downloadable malicious files during our collection period. We then analyzed the temporal topological structural evolution and malware hosted on FQDN servers of the three largest M-Clusters across the eight-month data collection period. Our

* Corresponding author.
E-mail addresses: josemora@andrew.cmu.edu (J.A. Morales), ycai@cmu.edu (Y. Cai).

https://doi.org/10.1016/j.array.2022.100174
Received 12 July 2021; Received in revised form 21 April 2022; Accepted 23 April 2022
Available online 14 May 2022

analysis revealed the layout of an M-Cluster as a *bridge* and *hub* construction for which we provide both a definition and detection algorithm. We further observed that the increase in size of an M-Cluster over the collection period aligned with the Power Law, implying that a small number of very large M-Clusters controlled the majority of malicious data flows throughout the MDN. Analysis of malware attribution revealed a high number of downloaded files which were known malware variants stemming from a much smaller number of malware families. Our M-Cluster analysis builds on previous findings [3–6] by revealing a consistent presence of multiple layers of URL shortening services [7] as a bridge and hub, which facilitate the obfuscation of servers hosting malware while simultaneously facilitating the flow of malicious traffic.

Our temporal analysis revealed the persistent existence of large clusters involved in malware distribution. These large clusters preserved their topological structures for extended time periods with only some sub-structural changes which contrast with other work suggesting smaller time period persistence [8]. This implies a potential continuous effort by malicious actors to preserve such structures that could facilitate long-term data flow monitoring upon initial malicious FQDN identification. This critical observation is due to our novel temporal analysis approach, and is likely not appreciable with static analysis methods that do not encompass ongoing structural changes occurring over time.

The contributions of this paper are: 1) identification of large FQDN-based structures used in malware distribution persisting for longer time periods than observed in current literature, 2) the significant role of bridges and hubs in malware distribution dynamics, 3) correlation between size increases of M-Clusters following the Power Law, 4) observation of ongoing dependency of URL shortening services in facilitating transmission of malicious data flows while obscuring the source of the flows. We have also contributed technical details on implementing our novel data fusion approach of two static data sources, namely GSB and VT.

The result of this data fusion was the creation of new observable knowledge. These observations led to unique and novel temporal insights on the structure, tendencies, and data flows of an MDN. In our review of current work, we did not find a previous example of our novel data fusion, and we further discovered our temporal insights presented here would not have been observed without our data fusion implementation, as the insights could not be appreciated from the static data provided by GSB and VT in isolation. Observed bridge and hub structural and temporal insights provided the foundation for suggested prevention approaches to sustained malware distribution in an MDN.

To foster research, we have publicly released the originally collected data set from GSB [9].

2. Related work

A large corpus of publications related to web-based malware delivery which explore their mechanics, structure, and usage context exists in the current literature, we reference here the most relevant to this research. Our temporal graph analysis approach is inspired by the work presented in [10,11]. In these approaches, analyzing a graph includes topological properties such as topological metrics of the degree distribution, connected components, clustering coefficient, rich-club connectivity, all-pairs-shortest paths, betweenness centrality, node stability, dynamic anomalies, and future behavior prediction.

An early analysis of malicious traffic flow through FQDNs described the delivery mechanisms via web browsers and the function of IFRAME redirects as part of a malicious uniform resource locator's (URL) content loading process [12,13]. Using the IFRAME redirects, the authors recorded the IP address from which dynamically-loaded content would appear in a given webpage. Their research discovered the use of multiple IP addresses connected via web content scripts. This sequence of IP addresses traversed the Internet until eventually arriving at a malicious server which was hosting either explicit code or the malware itself.

Some of the co-authors of this work participated in the creation of GSB. Graph-focused analysis from collected data of malware delivery mechanisms has been presented [14–21] including the use of GSB data [22]. Graphs were used in a multitude of ways to map the movement of malicious traffic across multiple servers. Using graph theory, these approaches applied fundamental graph equations, properties, and measurements the networks discovered as distributing malware. Several works describe the process and detection of downloading malicious files in various forms via compromised websites to victim machines [23–32]. The process and detection of malware downloading is very diverse, with most using captured traffic flows as the source data for their research. The source data was analyzed in a multitude of ways, focusing on detecting malicious data flows. Others used virtual environments with specially-instrumented browsers collecting relevant metadata of visited webpages known to download malware onto victim machines. Some used synthetic data, which is researcher-created data representing real malicious traffic as seen in real time. Other relevant papers exploring diverse perspectives including risk, detection, and malicious campaigns have been presented in [33–41].

The majority of these works demonstrated success in analyzing previously-occurred malware related events, and provided rules of risk to prevent future outbreaks. These works also provided insights on the relationships between domains, IP addresses, and other network components used in malware distribution. Our research presented in this document enhances the current literature by using a unique data fusion implementation of public data sources, leading to the creation of new observable temporal-based knowledge, which facilitates insights into temporal evolutions of FQDN topological structures actively hosting and transmitting malware. Most of the related work presented here used either static data sets, synthetic data sets, or data captured in real time during a malicious cyber event from the past. Our data set, collected in real time, represented malicious events happening at that moment. Our daily collecting routine allowed for temporal analysis which enhances the related work. An important enhancement to the related work was our ability to attribute time intervals of active distribution and hosting of multiple malware files in multiple FQDNs, along with subnet structural changes. This provided insight into the operational characteristics and infrastructure of the MDNs.

3. Defining a malware distribution network

MDN is a *dynamic graph* whose vertex (nodes) and edge (links) sets transition over time. Here, we consider a dynamic graph at an initial state $G_0 = (V_0, E_0)$ and its development over time: G_0, G_1, G_2 … The transition between two states G_i and G_{i+1} of the graph can be described by a set of updates T_{i+1}. The evolution of a dynamic graph over time is the result of a sequence of transitions $G_0 \rightarrow G_1 \rightarrow G_2 \rightarrow G_3 \rightarrow \ldots$. Analyzing a dynamic graph includes topological properties at certain states, such as topological metrics of degree distribution and connected components [1]. Given an MDN, we have the following specific infrastructural measurements:

Inbound Hub Node – a node that has more than m inbound links;
Outbound Hub Node - a node that has more than *n* outbound links; m, n $\geq$ 1;
Bridge Node *(Center Node)* – a node that connects to multiple hubs;
Sink Node – a node that has only inbound links.
Root Node – a node that has only outbound links;
Transition Node – a node that has both inbound and outbound links;
Persistent Link - a link that stays active for a period of time *p*.

Fig. 1 shows an example of the infrastructural components of an MDN.

In Fig. 1, every node represents an FQDN and every link represents the flow of malicious data between two nodes. In our MDN graph visualizations, the bridge and hub structure were predominant. This phenomenon illustrated the flow of malicious traffic occurring

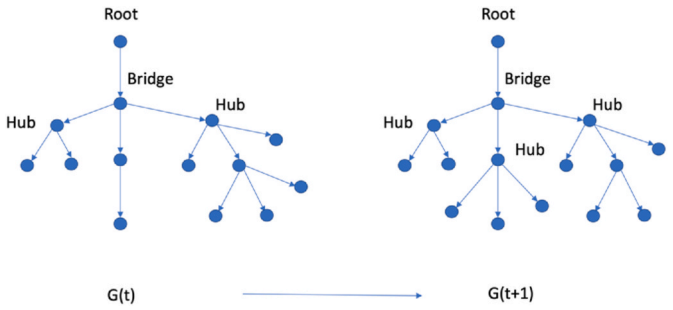

Fig. 1. Infrastructural components of an MDN

asymmetrically. Root nodes served as top-tier distribution starting points. It is assumed that the ingress of malicious data to a root node originates from some source outside of other MDN nodes. This source could be direct human interaction, data transfer methods not captured by GSB, and other data delivery techniques. Due to a lack of data, we were unable to identify root node data ingress sources. Inbound and outbound hub nodes served as middle-tier distribution points. A bridge node is another top-tier distribution point typically receiving malicious data from root nodes. This node type played a critical role in distribution, as their removal would create multiple isolated sub-structures of nodes incapable of malicious data ingress. Sink nodes are bottom-tier landing points of malicious data. Leaf nodes in an MDN graph are labeled as sink nodes. Transition nodes facilitate the flow of malicious data across an MDN. Bridge and hub nodes are considered transition nodes.

4. GSB data collection

The MDN was built from a novel data set which we created by implementing a fusion of static data collected from GSB and VT. The data set spans a period of eight months from 19 January to 25 September of 2017. The end date resulted from the removal of FQDN details relevant to this research in GSB API services. The GSB service is used to warn users not to visit potentially unsafe URLs. A blacklist is provided of URLs determined by GSB to have content related to phishing or malware. Our collected data is the particular class of URLs deemed malicious by GSB. The GSB Transparency Report is an online resource providing statistics from Google's internal collected data repository. An API set was made available to the public by Google to automate the retrieval of data from the repository for any submitted URL. The GSB API required a URL as input and returned a report in JSON format with several fields of labeled data including:

- 'website:name'
- 'sendsToAttackSites' *(AS)*
- 'sendsToIntermediarySites' *(IS)*
- 'receivesTrafficFrom' *(RT)*

These data labels were used to create a graph representation of the MDN's daily topological structure. In the GSB report for an FQDN named in the label *'website:name'*, the labels *AS* and *IS* included a set, delimited in [ ], of FQDNs that was the recipient of malicious traffic sent from this FQDN. An outbound edge was created between this FQDN and each entry in *AS*. The same process was applied to *IS*. The FQDNs listed in the set for RT were the senders of malicious traffic to this FQDN. An inbound edge was created from each FQDN listed in RT to this FQDN.

The label *RT* was used to distinguish between transitional nodes and root nodes. A non-empty set was labeled 'transitional' and an empty set was labeled 'root.' Note that a node labeled 'root' does not imply that it is a root node in the distribution of malware.

The static data collection process was started by querying GSB with a seed FQDN. If in the returned JSON report, one or more FQDNs were listed for AS, IS, or RT, those were added to a queue of pending queries to GSB. This process was exhaustively repeated until GSB reports with no FQDNs listed in any of these labels were received, and all FQDNs in the queue were processed.

A report was always retrieved for each of the FQDNs listed in the sets for each label. We did not have a case where a listed FQDN did not have a GSB report available. Our daily automated data collection would commence at 2 min past midnight EST by submitting the same seed FQDN, *vk.net*, to GSB. We selected vk.net as our seed FQDN through a daily query of the top 100 most visited websites from *Alexa.com* [42] to GSB for a period of 4 months from 10 July to 19 November of 2016. We sorted the results by the most frequently-appearing FQDNs based on the JSON reports. The FQDN vk.net was selected as the seed based on its persistent appearance in GSB reports.

Our daily collection process required between 4 and 11 hours to complete, thus starting and finishing within the same calendar date. We decided on a daily collection process after extensive manual analysis of the listed value for GSB report *'lastVisitDate'* data label over several weeks, which revealed GSB tended to only provide updates of their report details once per day. Once the daily collection process was completed, the data set was stored and labeled with the collection date. The complete eight-month data set consisted of 6,232,304 FQDN occurrences from a pool of 22,801 unique FQDNs.

The top 20 most frequently occurring FQDNs are listed in Table 1. Listed in this table are 7 URL shortening services with a total of 195,835 occurrences, highlighting their extensive use in MDNs [43]. The most frequently occurring URL shortener was bit.ly, which has previously been used for malicious purposes [44].

5. Dynamics of malware distribution clusters

We define an *M*-Cluster to be an interconnected subgraph of an MDN containing no less than five nodes. This number of nodes was arbitrarily chosen. Temporal visualizations of MDN graphs stemming from our daily data collection revealed each graph as a universe of multiple interconnected FQDN structures, as seen in Fig. 2. We refer to each structure as a cluster.

We discovered that the relationship between cluster sizes and rank fits the Power Law in the eight-month data collection time period as seen in Fig. 3 with:

$$N = \alpha X^{\beta} \quad \text{where} \quad \alpha = 218.745, \quad \beta = -0.832 \tag{1}$$

We confirmed that the Power Law fit by visualizing the data set for the first day of each month in our data collection as seen in Fig. 3 and Fig. 4 with the goodness of fit shown in Table 2 with R^2 defined in Equation (1) where y is the true number of nodes and f represents the number of nodes on the curve-fit. The Power Law alignment illustrates that the vast majority of clusters contained a small number of nodes (less than 20), with only a small number of clusters (less than 5) containing a larger count.

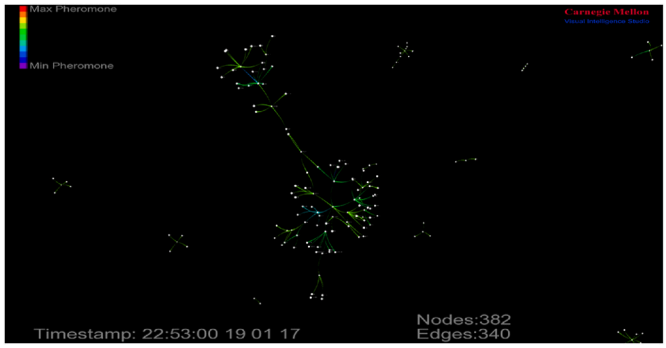

Fig. 2. MDN universe 19 January 2017.

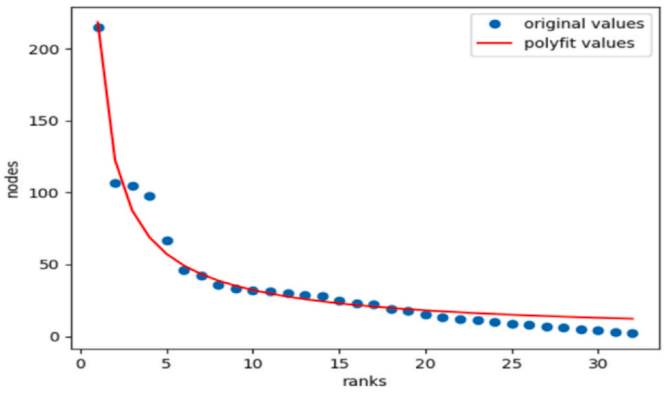

Fig. 3. Power Law over 8 month data collection period.

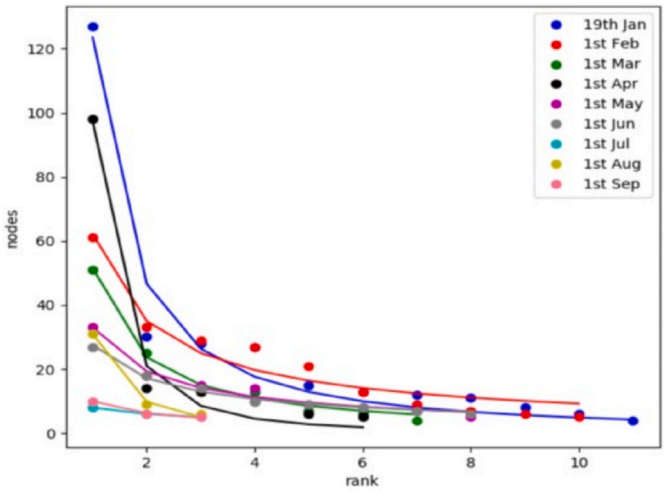

Fig. 4. Power Law applied to M-Cluster data by Month.

For any given collection date in our data set, the MDN universe is a set of multiple graphs, each representing the topological structure of a cluster.

We originally hypothesized the graphing of an MDN for a single day would be one large graph, but our temporal-based graph visualizations revealed many independent clusters. We rationalized that this universe of clusters was correct, since GSB reports covered a multitude of FQDNs likely linked to a wide range of global malicious efforts.

Furthermore, our literature review yielded no evidence to suggest malware distribution for all malicious efforts worldwide occurs via one large distribution network. A more reasonable scenario is that the operators of malicious campaigns incorporate their own networks of FQDNs, some using domain generating algorithms, for various tasks including distribution of malicious traffic [45–48].

The three largest, labeled C1, C2, and C3 respectively, peaked in size between January and April 2017, as seen in Fig. 5. C1 peaked in size in late January, C2 in early to mid-March, and C3 in early April. Visualizations of C1 – C3 during their peak sizes are in Figs. 6–8.

Focusing on C1 – C3, we analyzed the evolution of the number of nodes for each M-Cluster over its lifetime. The evolution of each of these clusters is shown in Figs. 10–12 respectively. The evolution of C1 was found to have sustained over 120 connected nodes in the second half of January and early February. There was a significant drop in the number of nodes on 31 January with a recovery increase in early February. Our research could not determine the cause of the sudden drop. There is a gap in the graph for 3–5 February. This gap resulted from our data collection scripts not running on those days.

Focusing on C1 – C3, we analyzed the evolution of the number of nodes for each M-Cluster over its lifetime. The evolution of each of these clusters is shown in Figs. 10–12 respectively. The evolution of C1 was found to have sustained over 120 connected nodes in the second half of January and early February. There was a significant drop in the number of nodes on 31 January with a recovery increase in early February. Our research could not determine the cause of the sudden drop. There is a gap in the graph for 3–5 February. This gap resulted from our data collection scripts not running on those days.

The bridge and hub structure is clearly visible in C1 – C3, and a further illustration for C1 is shown in Fig. 9. In this specific illustration, the URL shortening service bit.ly served as the bridge. Throughout the entire data set, URL shorteners served as a bridge or hub in many

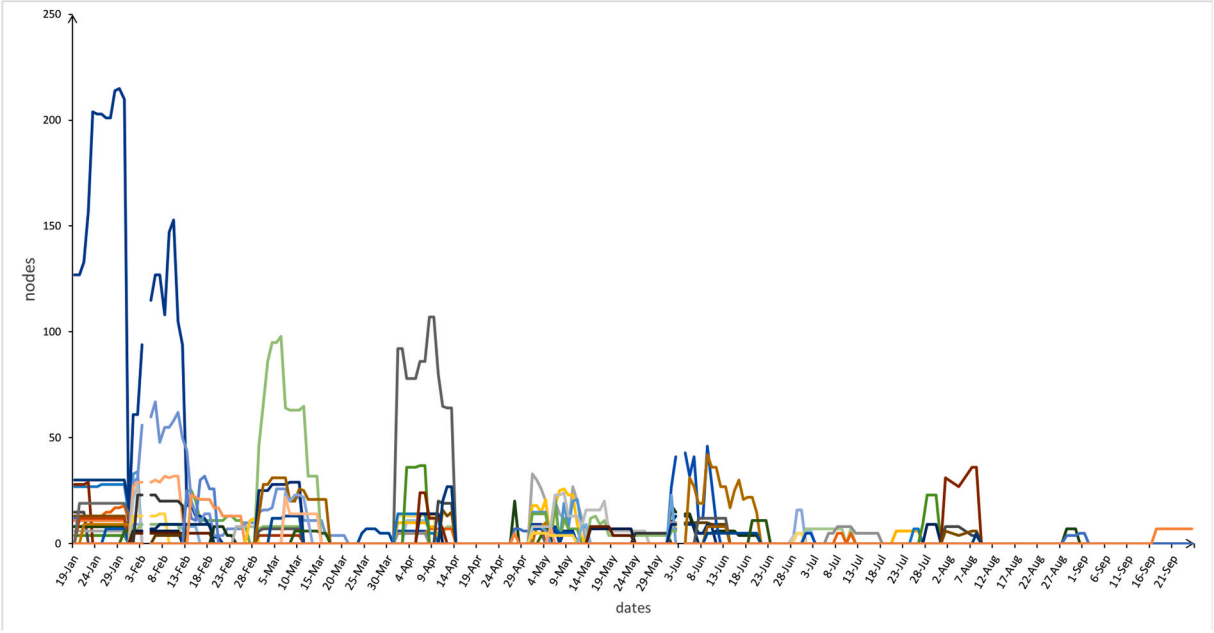

Fig. 5. Evolution of cluster sizes in the dataset.

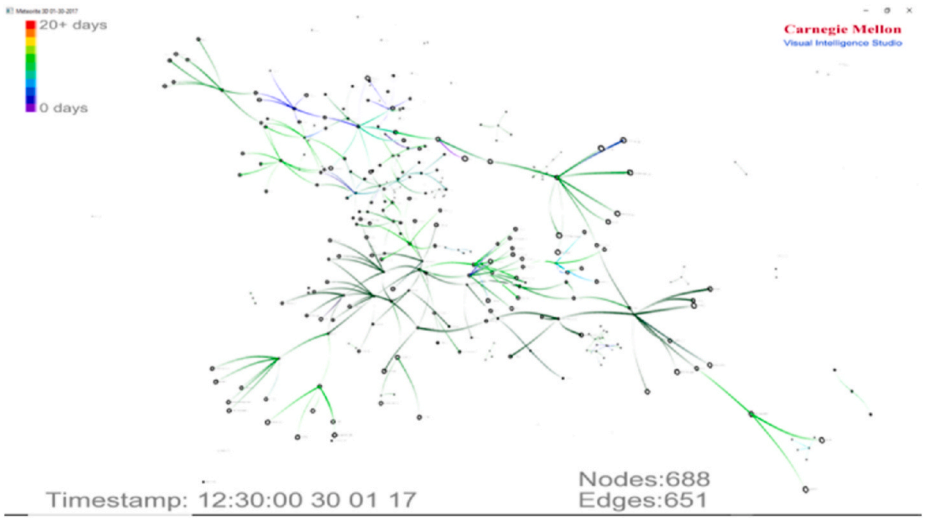

Fig. 6. Cluster C1 - 01/30/2017.

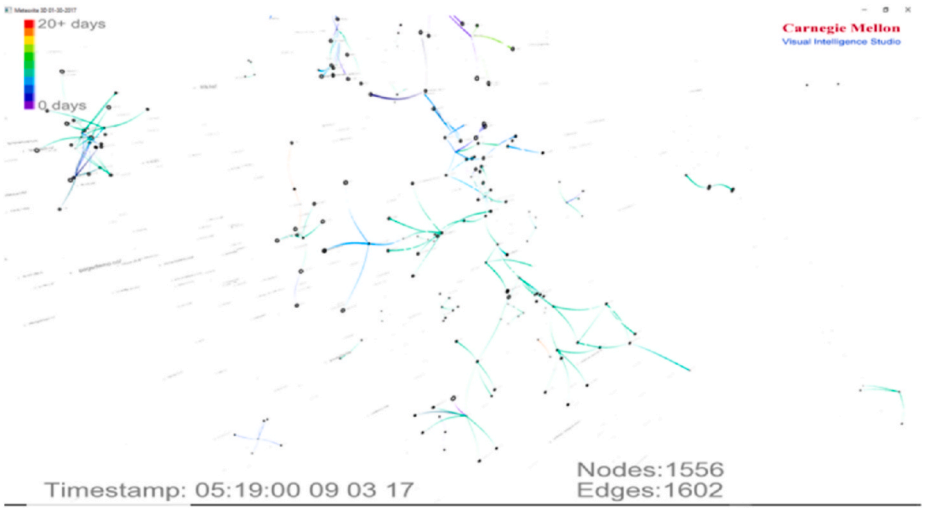

Fig. 7. Cluster C2 - 03/09/20.

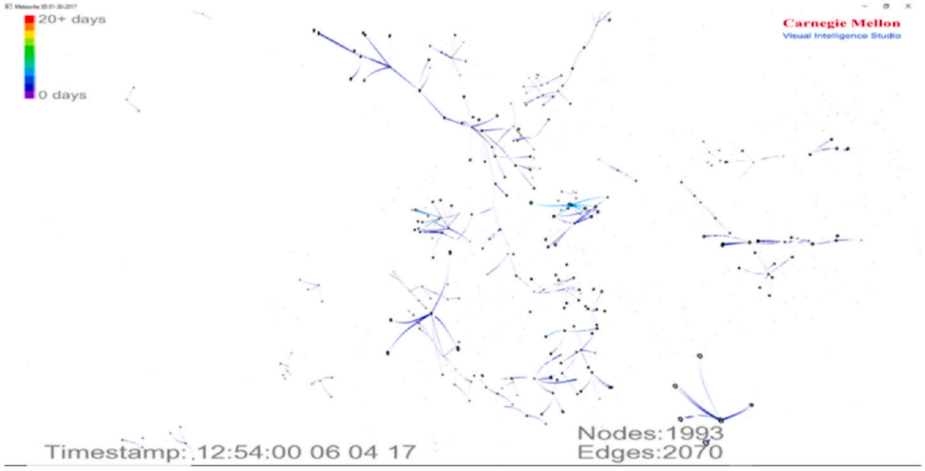

Fig. 8. Cluster C3 - 04/06/2017.

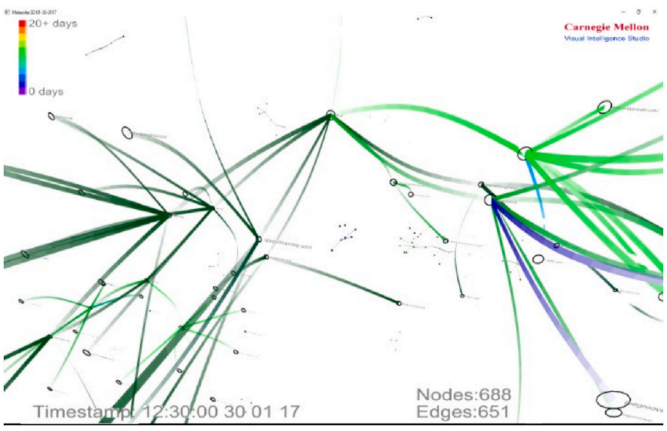

Fig. 9. Bridge - bit.ly of C1 - 01/30/2017.

instances, emphasizing their persistent and extensive use in implementing malware distribution.

The evolution of C2 was seemingly ordinary with a consistent rise in node size upon its first appearance in the MDN. This was followed by a period of node size stability from 3 to 10 March. The node size for C2 started to gradually decrease beginning on 10 March lasting until the end of life on 16 March.

The evolution of C3 occurred similarly to C2 with the exception of two instances of node size persistence during the period of gradual decrease to the end of life. All three clusters achieved sizes of over 100 nodes with C1 increasing past 200 and these large sizes were sustained for a minimum of 2 days for C2 to a maximum of 11 days for C1. Note that C1 appeared at a node size of 120 on our first day of collection. Since our data set starts on 19 January, we cannot definitively determine the creation date of C1 and node size evolution from that creation date to 19 January.

Focusing on the largest cluster, C1, we performed detection of bridge and hub structures using Algorithm 1. An extensive amount of bridge and hub structures were discovered. In Figs. 13–15, we have illustrated three bridge and hub structures contained within C1, and their infrastructural evolution over a period of several days.

In these three figures, each arrow type indicates a new node added to the structure. The structure in Fig. 13 persisted with no change, while those in Figs. 14 and 15 evolved with the addition of new nodes. The following were identified URL shorteners in our M-Clusters: *bit.ly, adf.ly, smarturl.it,* and *wp.me*. Some of these shorteners led to other shorteners, creating multiple layers between root and leaf nodes. The likely reason for the extensive usage, as previously mentioned, is malicious actor awareness of sub-optimal risk analysis techniques that were currently available at the time of our collection period. Our analysis revealed these three specific structures were the basis of the three biggest sub-structures of C1.

Algorithm 1
Hub and Bridge detection algorithm

Input:
The directed network, G1
The node set of the network, N1
The edge set of the network, E1(N_s, N_d)
Output:
Hub nodes, H_n
Bridge nodes, B_n
Pseudocodet:
1: for $N1_1 \rightarrow N1_n$ do
2: if OutDegree ($N1_i$) > 0 & InDegree ($N1_i$) > 0 then
3: if Degree ($N1_i$) > p then
4: $N1_i \in H_n$
5: end if
6: end if
7: end for
8: Create new directed network G2, with nodes set N2
9: for $E1(N_a, N_d)_1 \rightarrow E1(N_a, N_d)_n$ do
10: if $N_s \in H_n$ & $N_d \in H_n$ then

(*continued on next page*)

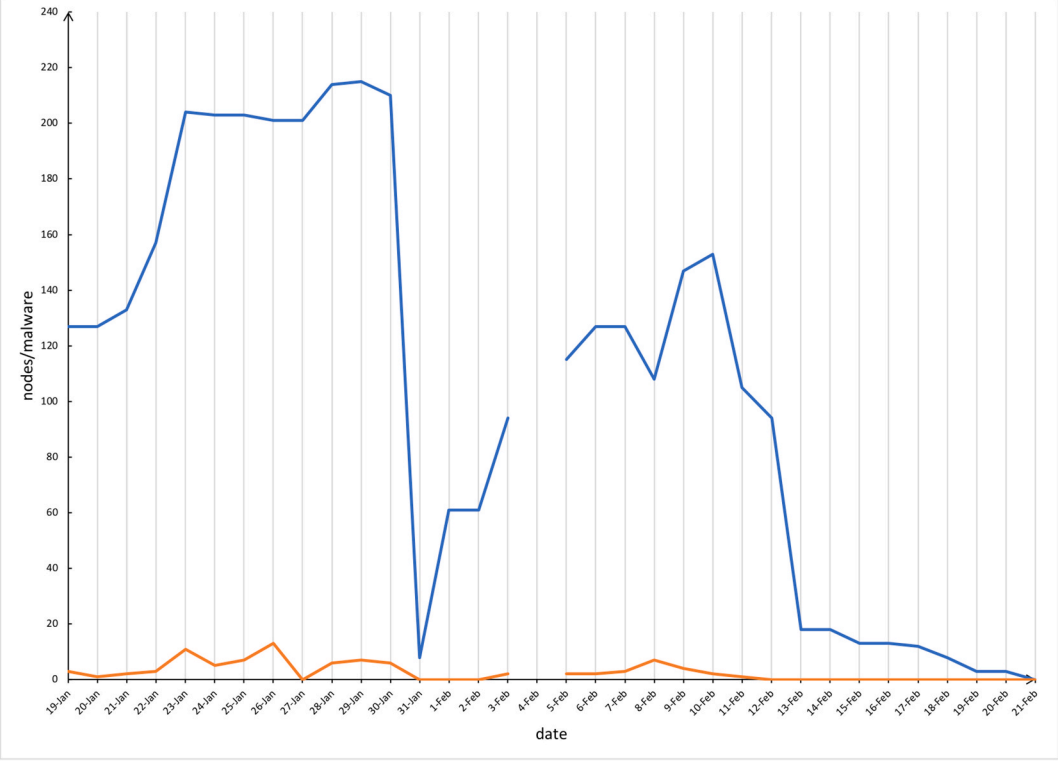

Fig. 10. Node evolution of C1.

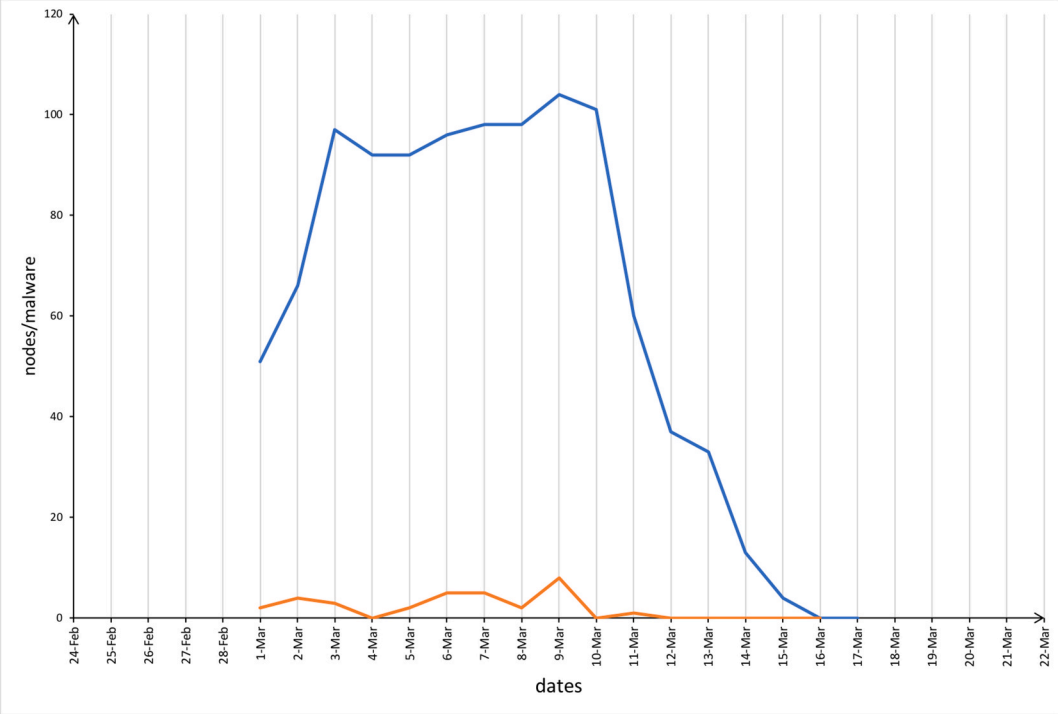

Fig. 11. Node evolution of C2 across its lifetime.

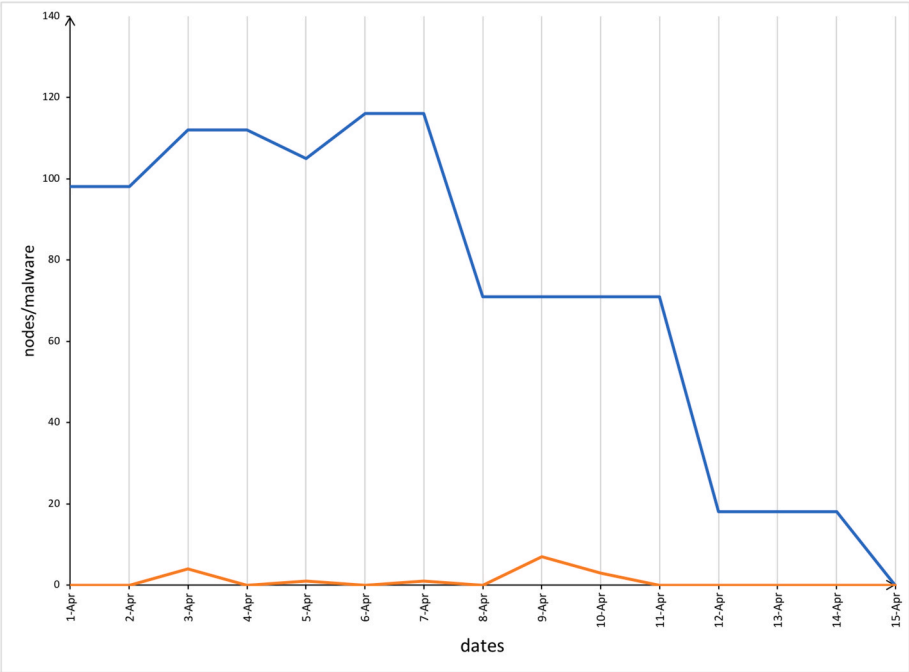

Fig. 12. Node evolution of C3 across its lifetime.

Algorithm 1 (*continued*)

11: $N_s \in N2$
12: $N_d \in N2$
13: end if
14: end for
15: for $N2_1 \rightarrow N2_n$ do
16: if OutDegree ($N2_i$) > 0 & InDegree ($N2_i$) > 0 then
17: if Degree ($N2_i$) > q then
18: $N2_i \in B_n$

(*continued on next column*)

Algorithm 1 (*continued*)

19: end if
20: end if
21: end for

6. Attribution of malware and M-clusters

Malware attribution was added to the data collection via a temporal

Fig. 13. Persistent bridge and hub - 19–30 January 2017.

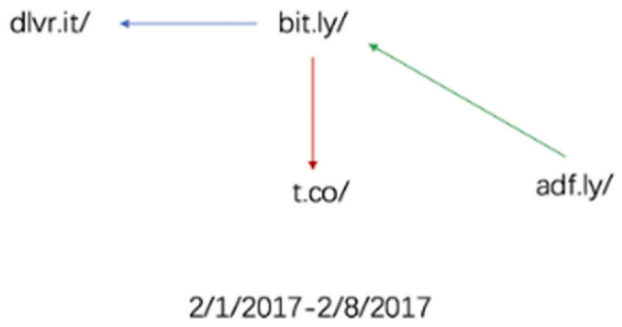

Fig. 14. Evolving hub and bridge - 01–08 February 2017.

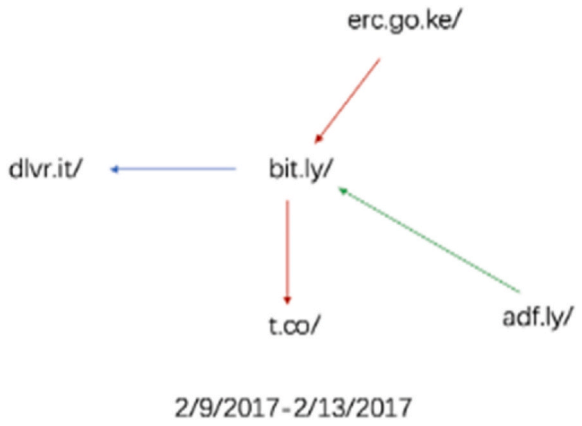

Fig. 15. Evolving hub and bridge - 09–13 February 2017.

correlation of known malware with collected FQDNs. This attribution was part of the resulting new observable knowledge from our data fusion of the GSB and VT collected static data. The correlation was accomplished by leveraging the online malware analysis service VT via an API key for academic use. A list was created enumerating all FQDNs contained in our GSB static data collection. Filtering of FQDNs appearing in M-Clusters C1, C2, and C3 was copied to a new list. Each FQDN was input via an automation script to VT, which returned a report in JSON format. A sample snippet of this report is shown in Fig. 16, highlighting relevant portions to this research. The report provides details of historical scan results of the submitted FQDN by VT. The reports consist of four main sections:

- detected_referrer_samples
- detected_urls
- detected_downloaded_samples
- undetected_downloaded_samples

Correlation between malware and M-Cluster was based on the label "detected_downloaded_samples". This label is defined by VT as a list of files that have been downloaded from the submitted IP address with at least one AV detection [49]. We could not find the definition of "detected_ downloaded_samples" for FQDNs, which is what we submitted to VT. Since VT reports for FQDNs and IP addresses consisted of the same labels, we assumed VT internally resolves or maps FQDNs to IP addresses and produces the relevant report. An MDN provides the network infrastructure facilitating malicious traffic movement across the Internet. It is reasonable to assume that malicious files being transmitted across an MDN would reside on a portion of the servers associated with a malicious FQDN. This reasoning is supported by GSB reports which indicate if an FQDN is determined to be hosting malware at the time of analysis. Each entry listed under "detected_ downloaded_samples" includes a SHA256 hash value of the file which was downloaded from the submitted FQDN, a date stamp of when the file was downloaded and scanned, the number of AVs detecting the file as malicious labeled as "positives," and the total number of AV software which scanned the file. Using the date stamp, we filtered entries that were within the life span of M-Clusters C1, C2, and C3. This range was from 19 January to 15 April of 2017. The results yielded a total of 132 malicious files downloaded from a subset of FQDNs, which were part of C1, C2, or C3 during the designated time period. The malicious files were downloaded from 60 unique FQDNs. The top 5 FQDNs with the highest number of malicious file downloads are listed in Table 3. Of these 5 FQDNs, two of them, namely app *loading.mobi* and expresent. info, had all of the malicious files downloaded in the same month with no downloads occurring thereafter within the M-Cluster lifespan. Three of the FQDNs, namely j.mp, *jmp.sh*, and *migre.me,* had malicious files downloaded from their servers across multiple months. These last three FQDNs are URL shorteners and the first two FQDNs are not.

There were 19 FQDNs with 2 or more downloaded files deemed malicious by VT, listed in Table 4. All downloaded files from the same FQDN had unique SHA256 hash values in the VT reports. Of these 19 FQDNs, 12 had all the malicious files downloaded in a single month with no other occurrence in the months before or after. Malicious file downloading in the remaining eight FQDNs occurred across multiple months. One URL shortener, urlz.fr, had all malicious file download instances occur in one month. The rest of the URL shorteners: *j.mp, jmp.sh, migre.me, sh.st, viid.me,* and *wp.me,* had their instances occur across 2–4 months. The FQDN *viid.me* has been reported as early as 2016 to be conducting browser hijacking as part of adware campaigns [50]. 41 of the 60 unique FQDNs had a single instance of a file downloaded from their associated server being detected as malicious by VT. The first 19 of these FQDNs are listed in alphabetical order in Table 5. Several URL shorteners appear in Table 5 such as *adf.ly, ht.ly*, *grabify.link*, and *fb.me*. Observe that these shorteners had only one instance of a malicious file download.

We derived insight on the lifespan evolution of FQDNs in an M-Cluster which had files downloaded and detected as malicious by VT. We leveraged our novel correlation of GSB and VT malware attributed data to note the following; the dates of the first appearance in an M-Cluster, instances of downloaded files detected as malicious, and last appearance in an M-Cluster. This task was performed on the first 10 entries of Tables 4 and 5. The results for FQDNs with all downloads in a single month are listed in Table 6. The results for FQDNs with downloads occurring in different months are in Table 7. FA is the first appearance of this FQDN in a cluster, MD is the date when files were downloaded by VT and detected as malicious, LA is the last appearance of the FQDN in the

```
submitted URL: reggaesudouest.com/
VirusTotal.com response:
{"detected_downloaded_samples": [
   {"date": "2017-08-08 09:31:28", "positives": 28, "total": 59,
   "sha256": "479c27298bb29f6f28089254f93a7f68fc040fddf60bcba26fc4a0fb29cc898f"},
   {"date": "2017-07-09 16:00:11", "positives": 25, "total": 58,
   "sha256": "b0f3de184d184a2cb502d5f9ca190ff181bc53d52939384dad1bfaf11e7fada0"},
   {"date": "2017-04-05 16:12:21", "positives": 18, "total": 56,
    "sha256": "caec1359075664ee2847b67a5bf49297d86d32b6bf8f90aaa278f28e29cb50d6"},],
 "detected_referrer_samples": [
   {"date": "2018-09-17 06:26:52", "positives": 23, "total": 71,
   "sha256": "7fbb53394e0e3df7b55ef90598eecf476411c165e5ad29fdf43dff0ddd4bdcb6"},],
 "undetected_downloaded_samples": [
   {"date": "2017-03-14 18:44:18", "positives": 0, "total": 63,
   "sha256": "0a6af411d11b53aeab73dc096568530d495f0f848a051304b02839c14c3e7c67"},],
 "detected_urls": [
   {"url": "http://reggaesudouest.com/tag/john-blazz/", "positives": 6,
   "total": 64, "scan_date": "2017-09-19 01:21:56"},]}
```


Fig. 16. VirusTotal.com JSON report snippet for a malicious FQDN.

cluster, and MC is the M-Cluster this FQDN belonged to during the time period of the malicious file downloads.

The lifespans of the FQDNs in one month listed in Table 6 illustrate the dynamic nature of participation in an MDN. There is no consistent persistence of lifespan. Some FQDNs persist in the full existence of the M-Cluster, while others last for a shorter period. This is in contrast to recent findings suggesting an average lifespan of 8 days [8]. There is a dominance of URL shorteners in Table 7, and their lifespan is much more persistent. These FQDNs persisted the entire lifespan, or merely a couple of days short, of their respective M-Cluster.

As previously mentioned, the attribution efforts yielded a total of 133 downloaded files detected as malicious by VT. Of these 133 files, 111 were unique based on the SHA256 hash value. This indicates that 22 malicious files were downloaded more than once. These repeated downloads occurred with seven unique hash values, listed in Table 8. Each SHA256 hash value is described by its malware name provided by either Microsoft, Sophos AV, or ESET-NOD32, in that order of preference, via the free detection service of VT. We chose these three products due to the fact that no single AV software provided malicious detection and identification for all 7 FQDNs at the time of scanning for malware identification in December of 2019. Two hash values did not produce any malicious detection from any AV software used by VT. According to the retrieved VT reports for these FQDNs, the hash values were considered malicious on the date of download and analysis.

Also included in Table 8 are the date and FQDNs for each instance of a malicious file download. Six of the seven hashes were downloaded only twice. The same pair of FQDNs were the source of a download of the same malicious file for three out of seven SHA256 values. The two FQDNs were: *jpm.sh* and jumpshareusercontent.com. As previously noted, *jpm.sh* is a URL shortener. There were three hash values that were downloaded in the same month with the remaining four spanning several months. There were also several download instances from two FQDNs on the same date. Based on our malicious file download statistics, one day was the average lifespan of a malicious file in an M-Cluster. These were the unique malicious files that had one instance of one download. There were two instances spanning between two and seven days, as is noted by the time lapse between download dates in Table 8. Also in this table are two files with the longest lifespan between downloads. The hash ending in 9863 lasted 40 days and 7e83 lasted 70 days. In Tables 4 and 6, the FQDN *apploading.mobi* had the highest number of downloaded malicious files with 24. This occurred in the span of seven days, detailed in Table 9. All of the downloaded malicious files were unique based on their SHA256 hash values. Based on their names, the files are members of one of just three malware families.

In order to appreciate the daily evolution of an MDN, we have illustrated the node transformations and appearances of malware for TechBrolo.A in Fig. 17. This malware is listed in Table 8 as the most downloaded in C1, C2, and C3. The red nodes are those FQDNs that contain this malware and the non-red nodes are FQDNs that send or receive traffic between red nodes. The red nodes are determined from

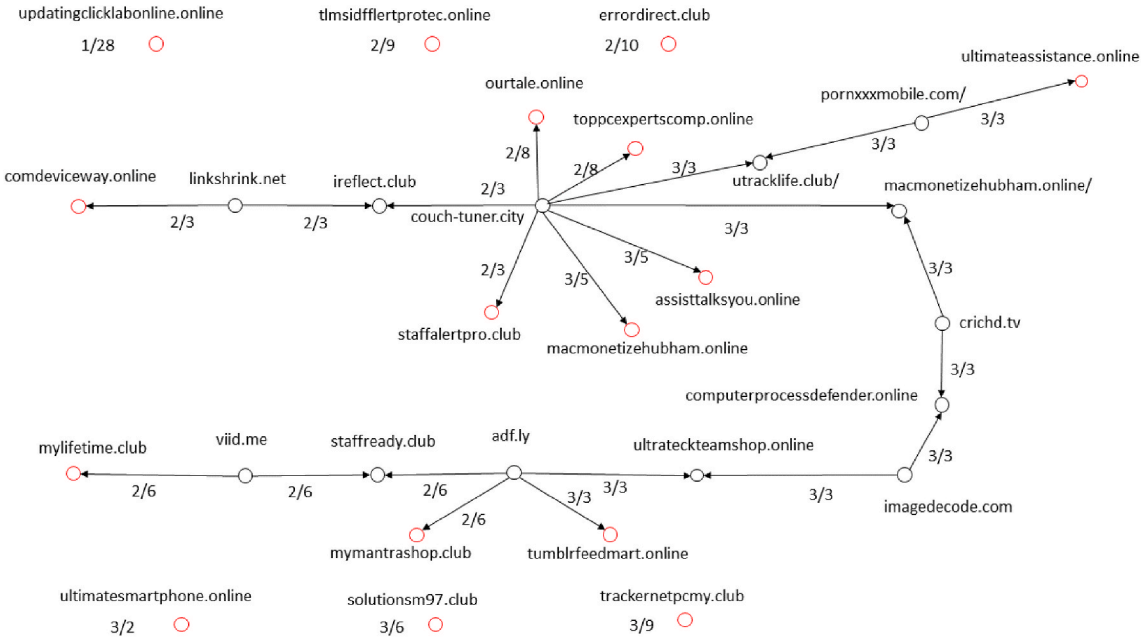

Fig. 17. The appearances of TechBrolo.A within clusters with time.

our collected VT static data and the non-red nodes from the GSB static data. The edges contain date stamps of appearance which indicate when malicious traffic flowed between nodes. Some infected nodes are shown as not directly connected to a subgraph. These nodes were part of the M-Cluster but were not directly connected to this slice of the cluster's topological network structure. The evolution is dynamic, with most nodes appearing and including an edge to one or two other nodes. There is a smaller set of nodes that, over time, acquired edges sending malicious traffic to multiple nodes - the FQDN couch-tuner.city is an example of this. The subgraph in Fig. 17 illustrates the new knowledge that the data fusion we implemented between GSB and VT data revealed from our collected static data primarily offers: temporal evolution of connected nodes which are known to contain malicious files. Leveraging this novel insight, our temporal analysis of daily structural changes in M-clusters revealed connectivity preservation over long periods of time for several bridge and hub components for all three clusters. We observed that for any given FQDN identified as hosting malware, the topological structure implementing its malicious data flow ingress preserved its pre-infection connectivity for days and even weeks after malware detection. In tracing the ingress data flows toward the malware-hosting FQDNs, we observed some first and second node degree sub-structural changes. However, from the third node degree and back, the structures were mostly preserved, with a far lower number of structural modifications. Most of this preservation occurred in structural connectivity stemming from bridge and hub structures composed of URL redirectors. We further observed subsequent malicious data flow ingress to malware-hosting FQDNs traversed circuits within the preserved areas. This suggests a focused effort to sustain, for a long period, the overall FQDN-based malware distribution network. The ability to identify and monitor data flows in the preserved areas could lead to halting malware distribution in its early stages.

7. Discussion

The key insight from this research is the observation of the long-term existence and preservation of FQDN-based topological structures involved in the transmission of multiple malicious data flows. This suggested a focused effort to sustain such infrastructures, and lends to the long-term monitoring of data flows traversing through these structures. This evidentiary insight is in contrast to previous studies suggesting that malware distribution structures exist only for short time periods. The insights of our work may suggest a shift in approach to malware distribution, favoring long-term sustainment of large topological structures by constantly-changing malware hosting and immediate node degree FQDNs while preserving FQDNs in more distanced degrees from a detected leaf node. The majority of FQDNs appear only once as the source of a unique malicious file download. Only a handful of FQDNs appeared in VT reports as the source of two or more unique malicious file downloads. Once VT reported an FQDN as hosting a file that was downloaded and detected malicious on a specific date, the FQDN no longer appeared in a VT report as hosting any malicious file. The implication is that an FQDN, according to VT, is used by malicious actors very few times, or potentially only once, to host downloadable malicious files on its servers. This lends itself to the idea of MDN operators constantly changing FQDNs in the degrees around leaf nodes, while preserving those in the sub-structures connected to root nodes, most notably bridge and hub implementations.

The following insights have also been gleaned from the research:

1) An MDN is composed of several isolated clusters of interconnected FQDNs. As opposed to our initial assumption of one large network, our MDN graph revealed that multiple isolated clusters are prevalent. Cluster survivability beyond termination attempts by security defenders is increased if multiple clusters are in use.
2) A Power Law distribution was established for cluster size and rank across the eight-month data collection period. This suggests that the majority of clusters are very small in size with only a handful being much larger. In our data collection, 20 nodes or less were common, with fewer than five M-Clusters possessing a higher node count.
3) A bridge and hub construct are key distribution points of malicious traffic. Our visualizations clearly showed the critical usage of bridges and hubs to route traffic from root nodes to leaf nodes within a cluster. Termination or use of a DNS sinkhole [59] on FQDNs serving as bridges and hubs can shut down large portions of an MDN.
4) There is a strong reliance on URL shortening services. Malicious actors seem to implement shorteners as bridges and hubs on a regular basis. Most of the well-known, and other not-so-well-known, shortening services were present in all of the clusters.
5) Malware distributed across a cluster seems to be composed of a large number of unique individual samples from a smaller number of malware families. The majority of the malware were HTML pages with JavaScript. The script is accessed when visiting a webpage, and can potentially perform automated tasks such as downloading or installing other malware on host machines. Other discovered malware were HTML pages attempting social engineering and FQDNs with a link to a publicly accessible directory within a storage system which contained malware.

7.1. Approaches facilitating prevention of sustained malware distribution in an MDN

In this work, the role of bridge and hub structures was shown to play a critical part in facilitating the transmission of malicious data flows in an MDN. These structures were always part of the connectivity between: top-level domain names leading to remote servers hosting malware, exploit servers, and compromised machines. The following were observed regarding the structural and data flow tendencies of bridge and hub constructs in an MDN:

- A node, some of which were root nodes, transmitting malicious data flows to a bridge node were often not exclusive to that bridge. These transmitting nodes were oftentimes observed to simultaneously send traffic to multiple bridges sustaining one to many relationships.
- Single bridge nodes were observed receiving malicious data flows from both single and multiple nodes.
- A single bridge node often transmitted malicious data flows simultaneously to multiple hubs with a smaller set of cases where transmission was just to a single hub.
- In cases where a URL redirecting service, such as bit.ly, was used as a bridge node, there were no observed nodes transmitting malicious data flow to that bridge. We inferred, from the structure and data flows, that the redirector obscured the domain name or IP address of a remote server that was a source of the malicious data flows being sent from the bridge to its hubs.
- Some bridge and hub structures were populated with multiple domain names of URL redirectors resulting in an overlaid obscuration of the IP addresses of remote servers transmitting malicious data flows.

With this insight, the following approaches may assist in stopping these malicious data flows:

- Blocking domain names providing ingress data to a detected bridge node. Given the observed one-to-many relationship discussed above, the interruption of data flows from these nodes could reduce the flow of malicious traffic to a larger, and potentially undiscovered, number of bridge nodes beyond the detected node. These nodes serving the malicious traffic are considered the entry points to the MDN by malicious actors. Their disablement results in actors consuming more resources to reestablish the blocked nodes.

- Blocking the domain name of the bridge node itself. The bridge is a connector between sources of malicious traffic and hubs. Blocking this node, for example by blacklisting, stops all traffic to the connected downstream hubs. The nodes acting as sources of malicious traffic to the blocked bridge node could continue streaming to other not-yet-discovered bridge nodes. The impact of blocking the bridge node is the removal of traffic to the subnet of hubs, both discovered and not, that are downstream.
- Blocking all the nodes populating a hub receiving malicious traffic from the bridge node. Domain names of detected hubs can be blocked to avoid continued receipt of malicious traffic from the bridge node. This option requires more resources by security defenders to achieve. The still-active bridge nodes can continue connecting to newly created and accessible hubs. This option is practical in cases where individual nodes of a hub are deemed necessary to block receipt of traffic, while allowing the rest of the hub's nodes to function normally in the MDN.
- Redirecting instead of blocking malicious data flows. Implementing data flow redirection to a security defender -managed network of servers allows long-term persistent observation and attribution of MDN structural evolutions over time. It was not clear in this work if any of these nodes had the ability to detect if their data flows were blocked and not being received by downstream nodes. If that were the case, redirection is a very good option. The data flows continue as expected by malicious actors, and simultaneously allow defenders to analyze and observe over long periods of time.
- Blocking an entire URL redirection service may not be a practical choice. Given their usage in MDNs, security defenders should consider: a security evaluation of the underlying redirected remote server although this can be difficult to achieve and require large amounts of resources to implement, allowing the use of only one or two approved services, or by establishing a proprietary security defender managed redirection service that meets the needs of a given user base.

7.2. Public release of the data set

In this research, we have introduced our novel construction of a unique method of data collection providing insight into the backend infra-structure facilitating malicious traffic. In order to encourage further exploration of the data set by fellow researchers, we have publicly released the collected Google Safe Browsing data set in its original form [9].

7.3. Limitations

The analysis and results are limited to data made publicly-accessible by GSB and VT. This data may be just a slice of much larger internally held data sets. We believe this limitation does not negatively impact our findings discussed in this research due to strong evidence in our data supporting MDN existence and occurrence. However, the internal data analysis and public dissemination practices of GSB and VT may affect our temporal analysis. Our collected data is based primarily on dates of malicious event occurrences publicly released by the aforementioned entities. Our findings could be made more precise with dates of when a malicious event actually occurred as opposed to when it was publicly reported.

8. Conclusions

In this research, we have detailed our novel data fusion approach of creating a unique data collection capturing the temporal evolution and malware attribution of a malware distribution network across an eight-month period. The data fusion of static data collected from GSB and VT resulted in the creation of new knowledge not observable in either static data collection in isolation. The newly created knowledge focused on structural and temporal dynamics of an MDN, as well as temporal malware attribution including hosting and transmission. This newly created knowledge led to key insights guiding this work. By leveraging publicly-available data from Google Safe Browsing and the VirusTotal online service, our data collection illustrated daily evolutions in the network topological structure of interconnected, fully-qualified domain names consisting of second and top-level domains. These evolutions include documented time periods when fully-qualified domain names hosted and facilitated the distribution and hosting of known malware. Insights from our analysis of the data collection produced several interesting findings including; long term persistence of unchanged topological structures facilitating malicious data flows, significant usage and persistence of URL shorteners, and an overall network universe composed of isolated interconnected, fully-qualified domain name clusters facilitating malicious traffic via bridge and hub structures, a Power Law distribution among cluster sizes, and the identification of a large number of unique downloaded malicious files belonging to a smaller number of malware families. Observed structural and temporal bridge and hub insights (stemming from the data fusion's newly created knowledge) facilitated suggesting approaches to prevent the long-term sustainment of malicious data flow transmission in an MDN. The original Google Safe Browsing data collection has been made publicly-available to encourage continued research. Our future work with malware distribution networks will include: further enrichment of the current data set by leveraging more public data, continuing analysis to determine usage of our data as an indicator of malicious cyber events, continued research to enumerate novel discernible insights achievable with our collected data, and the creation of new temporal MDN data sets with publicly available data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Yang Cai's research was in part supported by Northrop Grumman Cybersecurity Consortium, Siemens Technical Center, and Roll and Royce Security Division. Jose Andre Morales's research was unfunded.

Related Tables

Table 1
Most occurring FQDNs in the MDN data set

FQDN	Occurrences
umblr.com	107,986
bit.ly	73,056
sayhi.tw	65,312
ow.ly	33,204

(*continued on next page*)

Table 1 (*continued*)

FQDN	Occurrences
fb.me	31,161
vturl.net	18,834
ift.tt	18,343
jianso.com	17,625
sorbratchakarn.com	16,978
tinyurl.com	16,558
stalker-gsc.ru	15,941
google.com	13,878
yelp.com	13,194
leesou.com	13,130
adhome.biz	13,059
nregulation.bid	12,775
link.tl	12,372
toadgo.com	11,601
baidu.com	10,919
migre.me	10,650
ad2load.com	9921

Table 2
The goodness of Fit for M-Cluster by Month

Time	α	β	R^2
January 19	123.519	−1.406	0.964
February 1	62.093	−0.828	0.942
March 1	51.459	−1.119	0.992
April 1	97.300	−2.215	0.982
May 1	32.936	−0.772	0.977
June 1	27.495	−0.684	0.991
July 1	8.009	−0.424	0.999
August 1	30.906	−1.656	0.996
September 1	9.939	−0.666	0.991

Table 3
FQDNs with Highest Malicious File Downloads

FQDN	Total Downloads				
	Jan	Feb	Mar	Apr	Total
apploading.mobi	24	0	0	0	24
expresent.info	9	0	0	0	9
j.mp	1	6	2	0	9
jmp.sh	1	0	4	0	5
migre.me	2	2	0	1	5

Table 4
FQDNs with multiple malicious file downloads

FQDN	Total Downloads				
	Jan	Feb	Mar	Apr	Total
acessoriapromgaf.com.br	0	0	2	0	2
apploading.mobi	24	0	0	0	24
expresent.info	9	0	0	0	9
hyperline88.com	2	0	0	0	2
j.mp	1	6	2	0	9
jmp.sh	1	0	4	0	5
jumpshareusercontent.com	0	0	3	0	3
kitchencollections.com.sv	2	0	0	0	2
migre.me	2	2	0	1	5
sh.st	1	0	2	0	3
sproutwealth.com	2	1	0	0	3
sugarsync.com	0	0	4	0	4
totalsoft.org	4	0	0	0	4
transtrailers.com	2	0	0	0	2
ultradownloads.com.br	1	1	0	0	2
urlz.fr	0	0	2	0	2

(*continued on next page*)

Table 4 (*continued*)

FQDN	Total Downloads				
	Jan	Feb	Mar	Apr	Total
viid.me	1	0	2	0	3
wata.cc	0	2	0	0	2
wp.me	1	0	0	4	5

Table 5
Single Instance Malicious File Download

FQDN	Total Downloads				
	Jan	Feb	Mar	Apr	Total
adf.ly	1	0	0	0	1
apareco-rdc.com	0	0	0	1	1
assisttalksyou.online	0	0	1	0	1
bigsports.me	0	0	0	1	1
cast4u.tv	0	0	0	1	1
comdeviceway.online	0	1	0	0	1
demoda.es	1	0	0	0	1
equidadparalainfancia.org	0	0	0	1	1
errordirect.club	0	1	0	0	1
fb.me	0	0	0	1	1
freefilesdownloader.com	0	0	0	1	1
giayluoinamdep.com	1	0	0	0	1
grabify.link	0	1	0	0	1
hengyongonline.com	0	1	0	0	1
ht.ly	1	0	0	0	1
inilahbanten.com	0	1	0	0	1
jumpshare.com	1	0	0	0	1
macmonetizehubham.online	0	0	1	0	1
mmlyfah.com	1	0	0	0	1

Table 6
FQDN lifespan - malicious downloads in one month

FQDN	FA	MD	LA	MC
acessoriapromgaf.com.br	28-Feb	9-Mar	16-Mar	C2
adf.ly	20-Jan	30-Jan	21-Feb	C1
apareco-rdc.com	31-Mar	10-Apr	13-Apr	C3
apploading.mobi	20-Jan	23–26 28-30-Jan	21-Feb	C1
assisttalksyou.online	4-Mar	5-Mar	16-Mar	C2
bigsports.me	31-Mar	3-Apr	13-Apr	C3
cast4u.tv		14-Apr		C3
comdeviceway.online	1-Feb	3-Feb	21-Feb	C1
demoda.es	20-Jan	23-Jan	21-Feb	C1
equidadparalainfancia.org	31-Mar	10-Apr	13-Apr	C3
errordirect.club	9-Feb	10-Feb	21-Feb	C1
expresent.info	20-Jan	22,24-26-Jan	21-Feb	C1
hyperline88.com	20-Jan	26-Jan	21-Feb	C1
jumpshareusercontent.com	28-Feb	01,06,09-Mar	16-Mar	C2
kitchencollections.com.sv	20-Jan	19-20-Jan	15-Feb	C1
fb.me	31-Mar	09-Apr	13-Apr	C3

Table 7
FQDNs with Malicious Downloads in multiple months

FQDN	FA	MD	LA	MC
j.mp	19-Jan	25,29-Jan	–	C1
	–	05,07,08-Feb	21-Feb	C1
	28-Feb	06-07-Mar	16-Mar	C2
jmp.sh	20-Jan	29-Jan	21-Feb	C1
	28-Feb	01,06,08-09-Mar	16-Mar	C2
migre.me	20-Jan	25,30-Jan	–	C1
	–	8-Feb	21-Feb	C1
	31-Mar	3-Apr	13-Apr	C3
sh.st	20-Jan	26-Jan	21-Feb	C1
	28-Feb	03,11-Mar	16-Mar	C2

Table 8
Malicious Files with Multiple Downloads

SHA256	Count
2eea543c86312c0fd361c31cba8774d2d6020c5ebcc1ce1a355482de74ed9863 Microsoft: SupportScam:JS/TechBrolo.A 28-Jan: updatingclicklabonline.online 30-Jan: pcpyetutelanetcl.com 03-Feb: staffalertpro.club, comdeviceway.online 06-Feb: mymantrashop.club, mylifetime.club 08-Feb: toppcexpertscomp.online, ourtale.online 09-Feb: tlmsidfflertprotec.online 10-Feb: errordirect.club 02-Mar: ultimatesmartphone.online 03-Mar: tumblrfeedmart.online, ultimateassis, tance.online 05-Mar: macmonetizehubham.online, assisttalk, syou.online 06-Mar: solutionsm97.club 09-Mar: trackernetpcmy.club	17
0ae5a2d24d0e404bb0ed91f317225e3318d7b0348df5e1bbe0ed0022356cac5b Sophos AV: Mal/DrodZp-A 01-Mar: jmp.sh, jumpshareusercontent.com	2
4006bc614ebaa46a6274f3c13ea13b41e0a2e117bd8a2b0b31c65d47ce02c954 ESET-NOD32: A Variant Of Java/QRat.AP 06-Mar: jpm.sh, jumpshareusercontent.com	2
70c65bd0e084398a87baa298c1fafa52afff402096cb350d563d309565c07e83 No Detection in VT 23-Jan: mmlyfah.com 03-Apr: migre.me	2
cc2618a702572a82d54e1d10fc9aa91f99aaf36c966c9673307460d6b33d5211 Sophos AV: Mal/DrodZp-A 09-Mar: jpm.sh, jumpshareusercontent.com	2
ce7127c38e30e92a021ed2bd09287713c6a923db9ffdb43f126e8965d777fbf0 No Detection in VT 19-Jan: ht.ly 25-Jan: migre.me	2
e74099bb07c2b9e21f17aeb5cc01bd9ded54833d8972b02ffea8894b816156a3 ESET-NOD32: A Variant of Win32/Packed.VMProtect.F 28-Jan: viid.me 29-Jan: vfastdownload.com	2

Table 9
Malicious File Downloads for apploading.mobi

1/23/2017	e949d906b915bf1bd550f6cbafb2baba92a570416b87ed86868d804de8e79f6f ESET-NOD32: A Variant Of Win32/Kryptik.FNBK
	5c6032d322493e7ea2bb39bdb4813391e3ce50c527966a14f5ef80832748e9e8 Microsoft: SoftwareBundler:Win32/Prepscram
	60c8deb477839a6f5da7ff9c60fd4ca61c0989e8d9403cd3d33b761596ed377d Microsoft: SoftwareBundler:Win32/Prepscram
	97f5869a30564f560b0e1521ec63dc661f54a4039ea421e771419acc702789ab ESET-NOD32: A Variant Of Win32/Kryptik.FICH
	8fedd14366c352ccdc0fbfbc609fe1094dd1299c38b14dd02f39fee486fd20c1 ESET-NOD32: A Variant Of Win32/Kryptik.FICH
	ef9a08069bcc2c84b08f3b0852f2ac4e16a98cd2f965fcc0d996824c11604c30 ESET-NOD32: A Variant Of Win32/Kryptik.FICH
	d07cbfc5fc9387a559003d1583b83ff0900d6079465f94a00cfa02a0ec4fd07f ESET-NOD32: A Variant Of Win32/Kryptik.FICH
1/24/2017	f47a5f5594eee42be08345071e1baf9c9fa0813fea735a68713e0d8620575f19 Microsoft: SoftwareBundler:Win32/Prepscram
	e27af8ac7bb772dd5e6176b37f1d95beb94c48b4fca4b994e85d6e597bf5494a ESET-NOD32: A Variant Of Win32/Kryptik.FNFJ
	94dfdcbe63f8ecbab0fa0fdc7edaf8034910596e5002e6dce01ae8081c85da1e ESET-NOD32: A Variant Of Win32/Kryptik.FMYN
1/25/2017	8718f36a99246d8a3cd62160a154485d4a09ae981c1892f7c9bdc9639cb656d8 Microsoft: SoftwareBundler:Win32/Prepscram
1/26/2017	7169680f8250658c3f87facf4625fd1796b7f313d4f6e83da992d6dffea2bbef ESET-NOD32: A Variant Of Win32/Kryptik.FNBK
	c1495edb3411a1169ddbd74b6f48993d1ad9825d86db3b83929f198e58890ac4 ESET-NOD32: A Variant Of Win32/Kryptik.FNFJ
	5f2d00d54e5d02f533d4b54c2456f9ca110f8f2faff7034600bf63a643a44200 ESET-NOD32: A Variant Of Win32/Kryptik.FNFJ
	680f8c2285d91e939da37eaf590af10994beffffe24b9ed31177560bd89ea353 A Variant Of Win32/Kryptik.FMPA
	690dacdb6b3720b1f0068af71518af6bab89cde12045b98d9507c7fd5f1c9e7c ESET-NOD32: A Variant Of Win32/Kryptik.FMPA

(*continued on next page*)

Table 9 (*continued*)

	8524f6c34d0b8a96ff8c1508e096fd1f4ec15a5e4779537d1e90912fa73b2e41
	ESET-NOD32: A Variant Of Win32/Kryptik.FMPA
1/28/2017	818be4ef6122ed9af581442248612153347850386 3aaac90253da20fcc155493
	ESET-NOD32: A Variant Of Win32/Kryptik.FNBK
	8004d91e8af6aa2f9988f9574bef997b9a858320c304a2a68483e7762e47008f
	ESET-NOD32: A Variant Of Win32/Kryptik.FNFJ
1/29/2017	59062e40e8f7d8c730e6ac630b5096eb3b6f9de0d1564867c9eef7eac68b5f7b
	ESET-NOD32: A Variant Of Win32/Kryptik.FNPP
	c34cd3d3c27ad8436171b4ec3f25e43feb98a4dbbb1af50c69bb0226d5cde33f
	ESET-NOD32: A Variant Of Win32/InstallMonstr.QJ
	ebe446a4a36d8e1e2951800a7bd9bb71109b310aeca8af21b5fc15936dea1ce1
	ESET-NOD32: A Variant Of Win32/InstallMonstr.QJ
1/30/2017	57abfbde444d336a973019fd6ec797e32be9358c70cba6a8825d5bc3862ebfc9
	Microsoft: SoftwareBundler:Win32/Prepscram
	9e2346cc631dfcd99c91cab3bf5e261be36528ecb773b678567a82273152441e
	ESET-NOD32: A Variant Of Win32/Kryptik.FNPP

References

[1] Google Safe browsing [Online]. Available: https://developers.google.com/safe-browsing/.

[2] VirusTotal [Online]. Available: https://www.virustotal.com/gui/home/upload.

[3] Nikiforakis N, Maggi F, Stringhini G, Rafique MZ, Joosen W, Kruegel C, Piessens F, Vigna G, Zanero S. Stranger danger: exploring the ecosystem of ad-based URL shortening services. In: 23rd international conference on World wide web (WWW '14); 2004. https://doi.org/10.1145/2566486.2567983.

[4] Wang D, Navathe SB, Liu L, Irani D, A T, Pu C. Click traffic analysis of short URL spam on Twitter. In: 9th IEEE international conference on collaborative computing: networking, applications and worksharing; 2013. https://doi.org/10.4108/icst.collaboratecom.2013.254084.

[5] Thomas K, Grier C, Ma J, Paxson V, Song D. Design and evaluation of a real-time URL spam filtering service. In: IEEE symposium on security and privacy; 2011. https://doi.org/10.1109/SP.2011.25.

[6] Nelms T, Perdisci R, Antonakakis M, Ahamad M. WebWitness: investigating, categorizing, and mitigating malware download paths. In: 24th USENIX security symposium (USENIX security 15); 2015. https://dl.acm.org/doi/abs/10.5555/2831143.2831208.

[7] Megiddo N, McCurley KS. Efficient retrieval of uniform resource locators. United States of America Patent US6957224B1; 2000.

[8] Abuse ch. How to take down 100,000 malware sites [Online]. Available: https://abuse.ch/blog/how-to-takedown-100000-malware-sites/; 21 January 2019.

[9] Morales JA, Cai Y. Google Safe browsing MDN data set 2017 [Online]. Available: https://github.com/joemango/GSB-Data-Set.

[10] Schiller B, Deusser C, Castrillon J, Strufe T. Compile- and run-time approaches for the selection of efficient data structures for dynamic graph analysis. J Appl Netw Sci 2016;1(1). https://doi.org/10.1007/s41109-016-0011-2.

[11] Rossi RA, Gallagher B, Neville J, Henderson K. Modeling dynamic behavior in large evolving graphs. In: Sixth ACM international conference on Web search and data mining. WSDM '13); 2013. https://doi.org/10.1145/2433396.2433479.

[12] Provos N, McNamee D, Mavrommatis P, Wang K, Modadugu N. The ghost in the browser analysis of web-based malware. In: First workshop on hot topics in understanding botnets (HotBots'07), Berkeley, CA; 2007. https://dl.acm.org/doi/10.5555/1323128.1323132.

[13] Provos N, Mavrommatis P, Rajab MA, Monrose F. All your iFRAMEs point to Us. In: 17th conference on Security symposium (SS'08), Berkeley, CA; 2008. https://dl.acm.org/doi/10.5555/1496711.1496712.

[14] Li B, Vadrevu P, Lee KH, Perdisci R. JSgraph: enabling reconstruction of web attacks via efficient tracking of live in-browser JavaScript executions. In: Network and distributed system security symposium (NDSS 2018); 2018. https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_07B-4_Li_paper.pdf.

[15] Rahbarinia B, Perdisci R, Antonakakis M. Efficient and accurate behavior-based tracking of malware-control domains in large ISP networks. ACM Trans Priv Secur 2016;9(2). https://doi.org/10.1145/2960409.

[16] Rahbarinia B, Balduzzi M, Perdisci R. Real-time detection of malware downloads via large-scale URL->File->Machine graph mining. In: ACM symposium on InformAtion, computer and communications security (AsiaCCS 2016); 2016. https://doi.org/10.1145/2897845.2897918.

[17] Kwon BJ, Mondal J, Jang J, Bilge L, Dumitraş T. The dropper effect: insights into malware distribution with downloader graph analytics. In: 22nd ACM SIGSAC conference on computer and communications security (CCS '15); 2015. https://doi.org/10.1145/2810103.2813724.

[18] Macdonald M, Frank R. The network structure of malware development, deployment and distribution. Global Crime 2016;18:1–21. https://doi.org/10.1080/17440572.2016.1227707.

[19] Stringhini G, Shen Y, Han Y, Zhang X. Marmite: spreading malicious file reputation through download graphs. In: 33rd annual computer security applications conference (ACSAC 2017); 2017. https://doi.org/10.1145/3134600.3134604.

[20] Huang S, Chuang T, Huang S, Ban T. Comprehensible categorization and visualization of orchestrated malicious domain names using linkage analysis. In: 16th annual conference on privacy, security and trust (PST); 2018. https://doi.org/10.1109/PST.2018.8514178.

[21] Behfarshad Z. Survey of malware distribution networks. Electrical and Computer Engineering, Faculty of Applied Science, UBC, Canada; 2012. https://doi.org/10.1109/ACCESS.2020.2985990.

[22] Peryt S, Morales JA, Casey W, Volkmann A, Mishra B, Cai Y. Visualizing a malware distribution network. In: IEEE symposium on visualization for cyber security (VizSec 2016); 2016. https://doi.org/10.1109/VIZSEC.2016.7739585.

[23] Ife CC, Shen Y, Murdoch SJ, Stringhini G. Waves of malice: a longitudinal measurement of the malicious file delivery ecosystem on the web. In: Asia conference on computer and communications security (asia CCS '19), New York, NY; 2019. https://doi.org/10.1145/3321705.3329807.

[24] Rahbarinia B, Balduzzi M, Perdisci R. Exploring the long tail of (malicious) software downloads. In: IEEE/IFIP international conference on dependable systems and networks (DSN 2017); 2017. https://doi.org/10.1109/DSN.2017.19/.

[25] Gang W, Stokes JW, Herley C, Felstead D. Detecting malicious landing pages in malware distribution networks. In: 43rd annual IEEE/IFIP international conference on dependable systems and networks (DSN 2013); 2013. https://doi.org/10.1109/DSN.2013.6575316.

[26] Choi S-Y, Ik-Seon K, Kim DK, Noh B, Kim Y-m. Multi-level emulation for malware distribution networks analysis. In: International conference on information security and cryptology (inscrypt 2013); 2013. https://doi.org/10.13089/JKIISC.2013.23.6.1121.

[27] Fu C, Liu X, Yang J, Yang L, Yu S, Zhu T. Wormhole: the hidden virus propagation power of the search engine in social networks. IEEE Trans Dependable Secure Comput 2019;16(4):693–710. https://doi.org/10.1109/TDSC.2017.2703887.

[28] Chuang T, Huang S, Mao C, Jeng AB, Lee H, Ziffersystem. A novel malware distribution detection system. In: 2017 IEEE conference on dependable and secure computing; 2017. https://doi.org/10.1109/DESEC.2017.8073834.

[29] Lever C, Kotzias P, Balzarotti D, J C, Antonakakis M. A lustrum of malware network communication: evolution and insights. In: 2017 IEEE symposium on security and privacy (SP); 2017. https://doi.org/10.1109/SP.2017.59.

[30] Tanaka Y, Akiyama M, Goto A. Analysis of malware download sites by focusing on time series variation of malware. Elsevier J Comput Sci 2017;22:301–13. https://doi.org/10.1109/ISCC.2016.7543735.

[31] Thomas K, Elices Crespo JA, Rasti R, Picod J-M, Phillips C, Decoste M-A, Sharp C, Tirelo F, Tofigh A, Courteau M-A, Ballard L, Shield R, Jagpal N, Rajab MA, Mavrommatis P. Investigating commercial pay-per-install and the distribution of unwanted software. In: 25th USENIX security symposium (USENIX security 16); 2016. https://dl.acm.org/doi/10.5555/3241094.3241151.

[32] Nelms T, Perdisci R, Antonakakis M, Ahamad M. Towards measuring and mitigating social engineering software download attacks. In: 25th USENIX security symposium (USENIX security 16); 2016. https://dl.acm.org/doi/10.5555/3241094.3241154.

[33] Kim D. Potential risk analysis method for malware distribution networks. IEEE Access 2019;7:157–67. https://doi.org/10.1109/ACCESS.2019.2960552.

[34] Alabdulmohsin I, Han Y, Shen Y, Zhang X. Content-agnostic malware detection in heterogeneous malicious distribution graph. In: 25th ACM international on conference on information and knowledge management (CIKM '16); 2016. https://doi.org/10.1145/2983323.2983700.

[35] Choi SY, Lim CG, Kim YM. Automated link tracing for classification of malicious websites in malware distribution networks. J. Inf. Proc. Syst. 2019. https://doi.org/10.3745/JIPS.03.0107.

[36] Huang C, Sakib M, Kamhoua C, Kwiat K, Njilla L. A game theoretic approach for inspecting web-based malvertising. In: IEEE international conference on communications (ICC 2017); 2017. https://doi.org/10.1109/ICC.2017.7996807.

[37] Canali D, Cova M, Vigna G, Kruegel C. Prophiler: a fast filter for the large-scale detection of malicious web pages. In: 20th international conference on World wide web. WWW '11; 2011. https://doi.org/10.1145/1963405.1963436.

[38] Singhal M, Levine D. Analysis and categorization of drive-by download malware. In: 4th international conference on computing, communications and security (ICCCS 2019); 2019. https://doi.org/10.1109/CCCS.2019.8888147.

[39] Rossow C, Dietrich C, Bos H. Large-Scale Analysis of malware downloaders. In: Detection of intrusions and malware, and vulnerability assessment (DIMVA 2012); 2012. https://doi.org/10.1007/978-3-642-37300-8_3.
[40] Yu S, Gu G, Barnawi A, Guo S, Stojmenovic I. Malware propagation in large-scale networks. IEEE Trans Knowl Data Eng 2015;27(1):170–9. https://doi.org/10.1109/TKDE.2014.2320725.
[41] Zhang J, Seifert C, Stokes JW, Lee W. ARROW: GenerAting SignatuRes to detect DRive-by DOWnloads. In: 20th international conference on World wide web. WWW '11; 2011. https://doi.org/10.1145/1963405.1963435.
[42] Alexa Internet, Inc.. The top 500 sites on the web [Online]. Available: https://www.alexa.com/topsites.
[43] Gupta N, Aggarwal A, Kumaraguru P. bit.ly/malicious: deep dive into short URL based e-crime detection. In: APWG symposium on electronic crime research. eCrime); 2014. https://doi.org/10.1109/ECRIME.2014.6963161.
[44] Plohmann D, Yakdan K, Klatt M, Bader J, Gerhards-Padilla E. A comprehensive measurement study of domain generating malware. In: 25th USENIX security symposium (USENIX security 16); 2016. https://dl.acm.org/doi/10.5555/3241094.3241115.
[45] Antonakakis M, Perdisci R, Nadji Y, Vasiloglou N, Abu-Nimeh S, Lee W, Dagon D. From throw-away traffic to bots: detecting the rise of DGA-based malware. In: 21st USENIX security symposium (USENIX} security 12); 2012. https://dl.acm.org/doi/10.5555/2362793.2362817.
[46] Grier C, Ballard L, Caballero J, Chachra N, Dietrich CJ, Levchenko K, Mavrommatis P, McCoy D, Nappa A, Pitsillidis A, Provos N, Rafique MZ, Rajab MA, Rossow C, Thomas K. Manufacturing compromise: the emergence of exploit-as-a-service. In: 2012 ACM conference on Computer and communications security (CCS '12); 2012. https://doi.org/10.1145/2382196.2382283.
[47] Yadav S, Reddy AKK, Reddy ALN, Ranjan S. Detecting algorithmically generated malicious domain names. In: 10th ACM SIGCOMM conference on Internet measurement (IMC '10); 2010. https://doi.org/10.1145/1879141.1879148.
[48] VirusTotal.com. API responses [Online]. Available: https://developers.virustotal.com/reference#api-responses.
[49] NightWatcher. How to remove VIID.ME virus from chrome, firefox, Internet explorer? Uninstall VIID.ME guide [Online]. Available: http://regrunreanimator.com/newvirus/howto/remove-uninstall-viid-me-virus-redirect-from-chrome-firefox-internet-explorer.htm.
[50] Mazerik R. Understanding DNS Sinkholes – a weapon against malware [Online]. Available: https://resources.infosecinstitute.com/dns-sinkhole/#gref.