Cyber Attribution from Topological Patterns

Yang Cai(✉), Jose Andre Morales, and Guoming Sun

Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
ycai@cmu.edu, jose@josemorales.org

Abstract. We developed a crawler to collect live malware distribution network data from publicly available sources including Google Safe Browser and VirusTotal. We then generated a dynamic graph with our visualization tool and performed malware attribution analysis. We found: 1) malware distribution networks form clusters rather than a single network; 2) those cluster sizes follow the Power Law; 3) there is a correlation between cluster size and the number of malware species in the cluster; 4) there is a correlation between the number of malware species and cyber events; and finally, 5) infrastructure components such as bridges, hubs, and persistent links play significant roles in malware distribution dynamics.

Keywords: Cyber attribution · Malware · Malware distribution network · MDN · Dynamics · Graph · Security · Computer virus · Malicious software · Topology

1 Introduction

Similar to an epidemic virus spread, malicious files infect computer systems over a set of globally connected domains or IP addresses, which we call a malware distribution network (MDN) [4–7, 9–15]. In this paper, we study temporal topological structures of an MDN with subsets of connected domains as a malicious cluster (M-Cluster). We created a novel dataset over an eight-month period by crawling the transparency report repository of Google Safe Browsing as well as collected URL and malware file hash scanning results from VirusTotal [8, 17]. We analyzed the topological structural evolution and malware hosted on various domain servers of the three largest M-Clusters in an eight-month period. Our analysis revealed the layout of an M-Cluster as a *hub* and *bridge* structure. We further observed that the increase in size of an M-Cluster occured in parallel to an increase in discovered malware on the domain servers. One scenario in which the manifestation of an M-Cluster may occur is in conjunction with global events, for example, the 2017 Presidential Inauguration of the United States of America. Our M-Cluster analysis also revealed a consistent presence of multiple layers of URL redirection services, which, we believe, serves to obfuscate servers hosting malware. The contributions of this paper are: 1) observation and analysis of malware distribution networks as clusters with a bridge and hub construction; 2) correlation between size increases of M-Clusters and the presence of hosted malware; 3) the significant roles of persistent bridges and hubs in malware distribution dynamics; and 4) development of algorithms to identify hubs and bridges.

V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12142, pp. 58–71, 2020.
https://doi.org/10.1007/978-3-030-50433-5_5

2 Literature Review

Dynamic graphs have been used in software engineering and operation research. Schiller and Strufe developed the framework for the analysis of dynamic graphs with DNA (Dynamic Network Analyzer) [2]. The topological properties of a dynamic graph include topological metrics of degree distribution (DD), connected components (C), assortativity (ASS), clustering coefficient (CC), rich-club connectivity (RCC), all-pairs-shortest paths (SP), and betweenness centrality (BC) [1]. Yu, et al. [26] studied the malware propagation dynamics of a single malware ConFlicker botnet. The authors tracked three top-domain layers and the growth of total compromised hosts by Android malware. The authors used the epidemic dynamics model to interpolate the malware distribution process. They discovered the Power Law distribution of ConFlicker botnet in the top three levers, i.e. ranking in botnet size of the malware versus probability of the distribution. This is perhaps the most comprehensive study of malware distribution at single botnet with a computational distribution model.

Here, we define a malware distribution network (MDN) as a *dynamic graph* whose vertex (nodes) and edge (links) sets change over time. We consider a dynamic graph at an initial state $M_0 = (V_0, E_0)$ and its development over time: M_0, M_1, M_2, … The transition between two states M_i and M_{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.

$$M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow M_3 \rightarrow \ldots$$

Given a malware distribution network (MDN), we have 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; *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; *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*. Figure 1 shows an example of infrastructural components of an MDN.

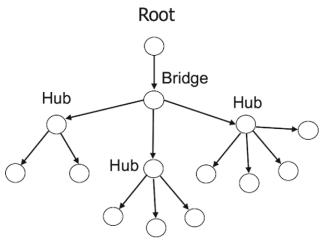

Fig. 1. Infrastructural components of an MDN

3 Semantic Graph Model

In this study, we embed semantic information into the dynamic graph of malware distribution networks. Graphs are represented by an augmented adjacency list data structure that is designed to capture both the dependencies of graph links and the mode of linkage types. We describe this data structure as a list of key–value pairs, whose keys are the top level domain of a website, denoted as a source, and key values are a pair <mode, destination> whereby destination is the top-level domain which is reported as being affected by the source. To place all of the top-level domains on the visualization, we used a Dynamic Behavioral Graph [22–24] to incorporate event frequencies, protocol types, packet contents and data flow information into one graph. In contrast to a typical Force-Directed Graph such as D3 [18], our model goes beyond the aesthetic layout of a graph to reveal the dynamic sequential patterns in a three-dimensional virtual space. In the model, the attraction force between a pair of nodes is calculated using the formula:

$$f_a = \frac{\left|\left|x_j - x_i\right|\right|^2}{\alpha T} \tag{1}$$

$$f_r = \frac{\beta}{\left|\left|x_j - x_i\right|\right|^2} \tag{2}$$

where: i and j are distinct nodes, α is the value of elasticity where a greater value increases the length of the edge. β is the coefficient for repulsion force. T is equal to the average time between each nodes' timestamps and $\left|\left|x_i - x_j\right|\right|$ is the distance between two nodes.

We use a gradient arc for displaying the direction of edges. The decrease of alpha value indicates the direction, with 1 at the source and 0 at the end. This novel visual representation also enables us to add the attributes to the edges [19–21].

Here, we enable digital pheromone deposit and decay on the edges of a network. The digital pheromones are stored on the connected edges over time. The digital pheromones also decay at a certain rate. The amount of pheromones at an edge at time t is:

$$\text{Deposit}: \quad D(t) = min\left(\sum_{i=0}^{N} u_i(t), M\right) \tag{3}$$

$$\text{Decay}: \quad D(t) = max(u_i(t) - rt, L) \tag{4}$$

where, $D(t)$ is the current pheromone level at a particular edge i between two nodes. M and L are the upper and lower bound limits to it. $u_i(t)$ is an individual pheromone deposit at time t, and N is the total number of deposits on that particular edge. 'r' is the linear decay rate. See Fig. 2.

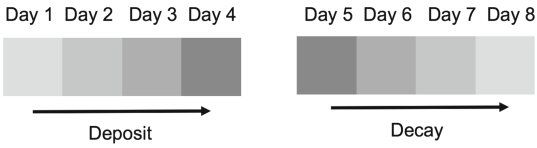

Fig. 2. Pheromone deposit and decay representation of persistency of the malware distribution channels (connected edges in the graph).

4 Data Collection and Malware Attribution

The MDN and M-Clusters were built from our dataset collected from Google Safe Browsing (GSB) and VirusTotal.com (VT). The data set spans a period of eight months from 19 January to 25 September 2017. The collection start date was specifically chosen to capture data related to the 2017 U.S. Presidential Inauguration. The end date, unfortunately, resulted from the unavailability of GSB API services. The GSB service has been used to warn users not to visit potentially unsafe URLs. The GSB Transparency Report is an online resource providing statistics from the collected data repository. An API set was made available to automate the retrieval of data from the repository for any submitted URL. The API requires a URL as input and returns a report including the timestamp of the last visit, the source, and the destination of the transmission. However, the report does not contain specific malware information.

VirusTotal (VT), on the other hand, provides a scanning service to detect the presence of malicious code in files and URLs. VT provides specific malware information. However, it does not contain the source-destination data. Scanning is a combination of multiple commercial anti-malware products providing both static and heuristic-based data analysis. In this study, we used the academic API service to automate submission and result retrieval for large data sets.

The site *vk.net* was selected as the seed website based on a four-month observation of the site reliably appearing on GSB. The report, in JSON format, consisted of various statistics. The statistics of interest to us were labeled: *name*, *sendsToAttackSites*, *receivesTrafficFrom*, *sendsToIntermediary-Sites*, *lastVisitDate*, and *lastMaliciousDate*. An MN with no incoming edges for the current collection was relabeled to a Root Malicious Node (RMN). This node is unique to our MDN graphs as it cannot be determined from the GSB reports alone. It is revealed only if the MDN graph is completed.

5 Topological Dynamic Clusters

The malware distribution network is not a giant web. Instead, there are many clusters of subnetworks. Some are large; others are small. All of the clusters are dynamic. They formed for a period of time and then dissolved gradually. Figures 3, 4 and 5 are the top three clusters in size. Figure 6 shows an overview of the 8-month dataset of cluster sizes (nodes) evolved over time, where each curve represents a cluster whose nodes are more than 5 nodes. The first blue line between 19 January, 2017 and 1 April, 2017 was the biggest cluster.

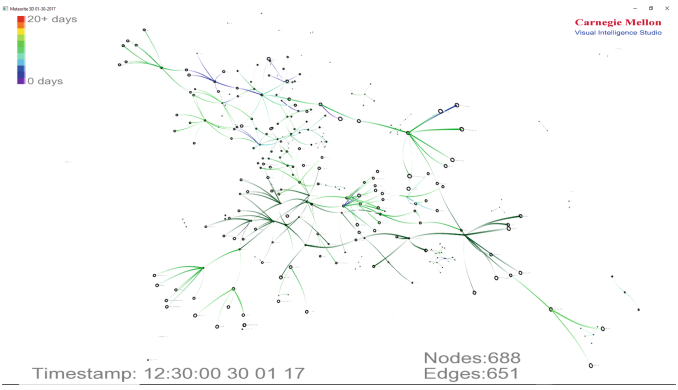

Fig. 3. The biggest cluster on 01/30/2017 from the visualization

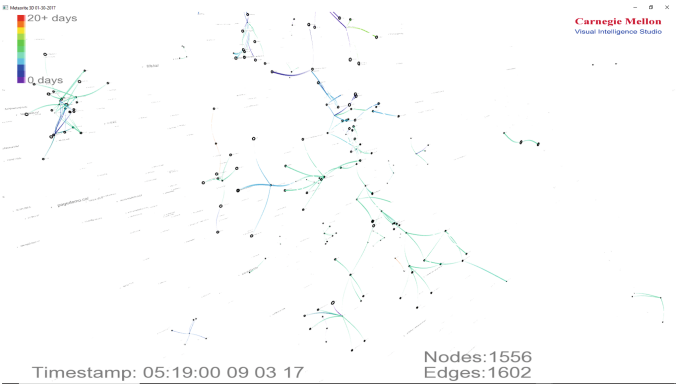

Fig. 4. The second biggest cluster on 03/09/2017

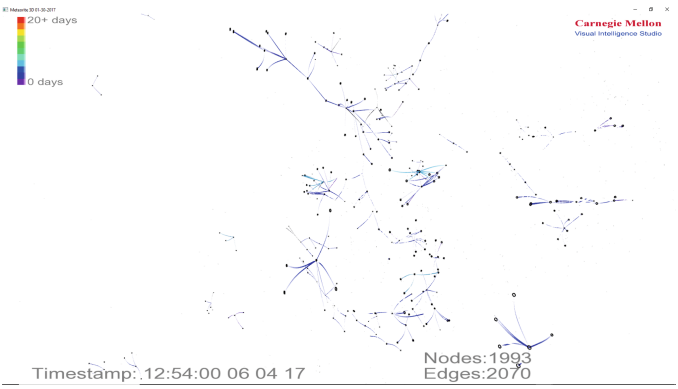

Fig. 5. The third biggest cluster on 04/06/2017

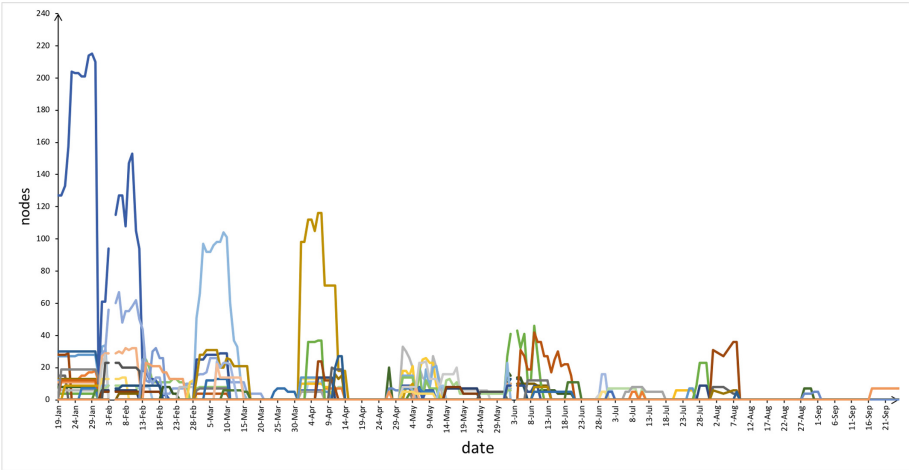

Fig. 6. The overview of the 9-month (1/19/2017–9/25/2017) dataset of cluster sizes (nodes) evolved over time, where each curve represents a cluster whose nodes are more than 5 nodes. The first blue line between 19 January, 2017 and 1 April, 2017 was the biggest cluster (Color figure online)

Statistical data analysis shows that the sizes of the clusters versus their ranks fits Power Law for most months, especially the first two months of 2017. See Fig. 7. This trend indicates that the MDN is a scale-free network: a very small number of nodes have more persistent edges than others. The topological patterns help the analysts to pay attention to the largest clusters, rather than many, many smaller clusters. In our case, this would include the clusters after May. Besides, we found that during volatile cyber attack seasons, the Power Law effect becomes stronger in terms of the slopes of the curves.

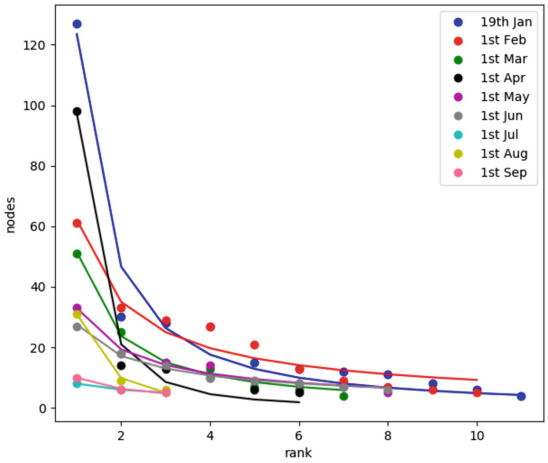

Fig. 7. The relationship between cluster sizes and rank fits the Power Law

6 Correlation of Events and Malware Clusters

Our dataset shows a correlation between major events and surge of malware distribution nodes. For example, the largest cluster formed after US Presidential Inauguration Day, between January 20 and February 13, 2017. Studies show the co-occurrence of bonets on social media and political events, such as national elections, inaugurations, and the controversial "Muslim Ban" [3]. After the election, the active bot accounts continued and increased by a certain amount. After the Inauguration, the active bot accounts increased even more. Our dataset only captured one of the significant events in 2017. The causal relationship between botnets and events is to be further explored. The number of nodes and malware can be fitted by:

$$Y = 9.027X + 125 \tag{5}$$

The correlation coefficient between the number of nodes and malware is 0.60 (Fig. 8). We detected the most popular single malware within our clusters by submitting the domains to VirusTotal. Next, VirusTotal responded to us with all of the malware downloaded from that domain with the last scanned date. We collected all of the malware whose last scanned date was the same as our collection date of the domain. The red nodes are those domains containing the single malware, and the other nodes are domains that send or receive traffic between red nodes. The single malware appears 17 times in the top three biggest clusters.[1] The rest of the detected malware in the three biggest clusters were discovered present on a server no more than two times with several appearing only once. Seven malware events occurred twice and the remaining 102 malware appeared only once (Figs. 9, 10 and 11).

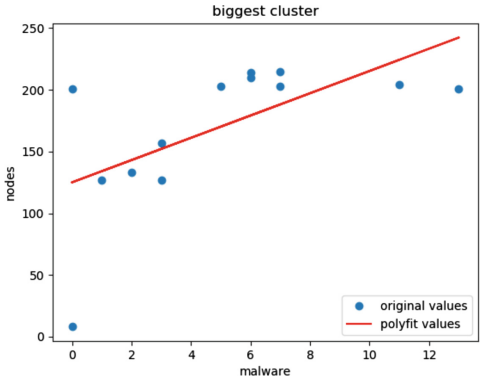

Fig. 8. The linear relation between species of malware and cluster size

[1] The SHA-256 of M is2eea543c86312c0fd361c31cba8774d2d6020c5ebcc1ce1a355482de74ed9863.

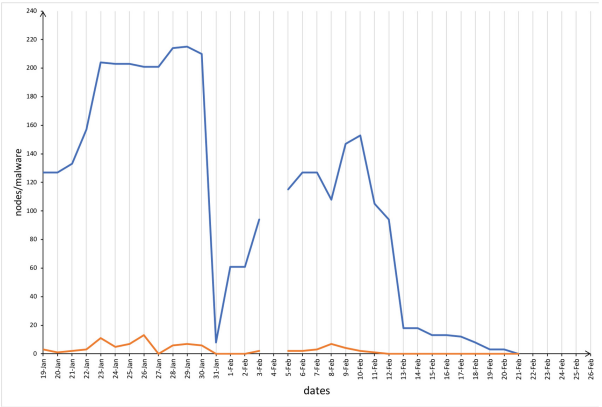

Fig. 9. The biggest cluster evolved over time in terms of size (nodes) and attributed malware. The red line is the number of malware in the cluster. (Color figure online)

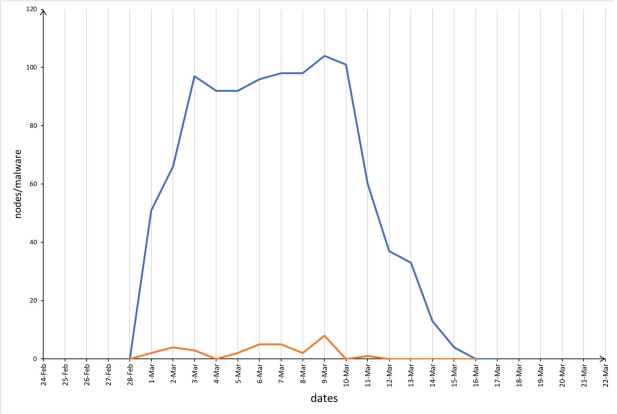

Fig. 10. The second big cluster evolved over time in terms of size (nodes) and attributed malware. The red line is the number of malware in the cluster. (Color figure online)

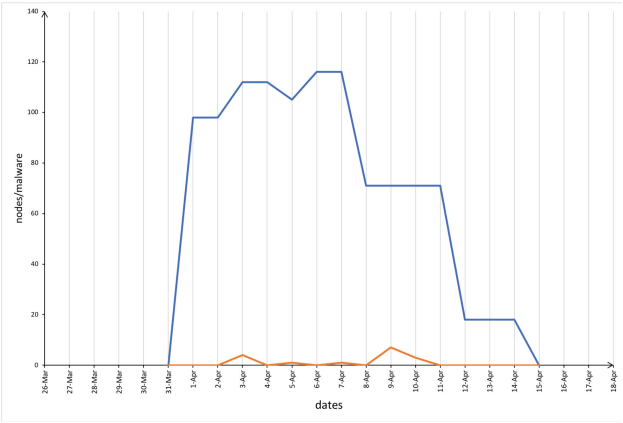

Fig. 11. The third big cluster evolved over time in terms of size (nodes) and attributed malware. The red line is the number of malware in the cluster. (Color figure online)

7 Cyber Attribution from Topological Patterns

The topological attributes help us determine the impact of the nodes in a malware distribution network (MDN). Visualization provides an intuitive tool to find the critical hubs and bridges, which are illustrated in Fig. 1. However, it is not efficient to identify those nodes when the dataset is so large. Here, we present the pseudo code for automatically searching for and labeling hubs and bridges. The algorithm is fast and can be used for tracking particular hubs and bridges over time. Eventually, the visual analytic process would be automated once human analysts have had successful experiences. In addition, humans and machines can always team up to discover new patterns and correlations based on graphic abstraction and visualization.

Algorithm 1. *Hub* and *Bridge* detection algorithm

Input:

The directed network, *G1*

The node set of the network, *M1*

The edge set of the network, $E1(M_s, M_d)$

Output:

Hub nodes, H_n

Bridge nodes, B_n

for $M1_1 \rightarrow M1_n$ **do**

if $OutDegree(M1_i) > 0$ & $InDegree(M1_i) > 0$ **then**

if $Degree(M1_i) > p$ **then**

$M1_i \in H_n$

end if

end if

end for

Create new directed network G2, with nodes set M2

for $E1(M_a, M_d)_1 \rightarrow E1(M_a, M_d)_n$ **do**

if $N_s \in H_n$ & $M_d \in H_n$ **then**

$M_s \in M2$

$M_d \in M2$

end if

end for

for $M2_1 \rightarrow M2_n$ **do**

if $OutDegree(M2_i) > 0$ & $InDegree(M2_i) > 0$ **then**

if $Degree(M2_i) > q$ **then**

$M2_i \in B_n$

end if

end if

end for

Figure 12 shows the infrastructural evolution of the malware distribution network between Jan. 19, 2017 and April 4, 2017. We found that there were several hubs in the biggest cluster, including *bit.ly*, *dlvr.it*, *smarturl.it*, *adf.ly*, *wp.me*, and *zip.net*, a bridge *bit.ly*, and a root node *brandnewbrand.br.* Amazingly, five out of six hubs are utility sites for shortening URL addresses: *bit.ly, adf.ly, smarturl.it,* and *wp.me.* Those sites redirect traffic to the malware host site.

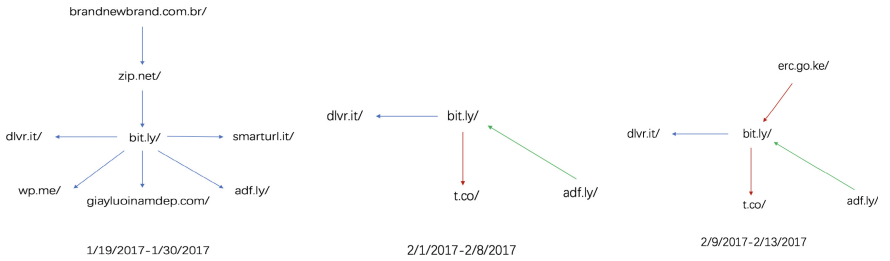

Fig. 12. Dynamic graph of the infrastructure of the biggest cluster between Jan 19, 2017 and Feb 13, 2017

With the visualization and analytic model, we are able to track single Top Level Domain (TLD) nodes and reveal their "life cycle" in the malware distribution network, when the TLD address has been captured by both Google Safe Browsing (GSB) and VirusTotal (VT). Figure 12 shows the dynamics of the TLD *adf.ly* node and its inbound and outbound edges in the 8-months period. The plot shows that the node had persistent malware inbound and outbound traffic before January 19 through May 17. There are multiple recurrences during that period. The malware did not die out until May 17, 2017. It reached its peak between Feb 19 and March 19, in correlation with the cyber activities during that period.

We are also able to track a single malware from Jan 28 through March 9 based on the GSB and VT attributed dataset. Coincidentally, the single malware passed through the popular TLD address node adf.ly during Feb 6 and March 3. The multiple modality tracking enables us to cross-reference, discover new patterns, and ultimately to lead more accurate cyber attributions (Figs. 13 and 14).

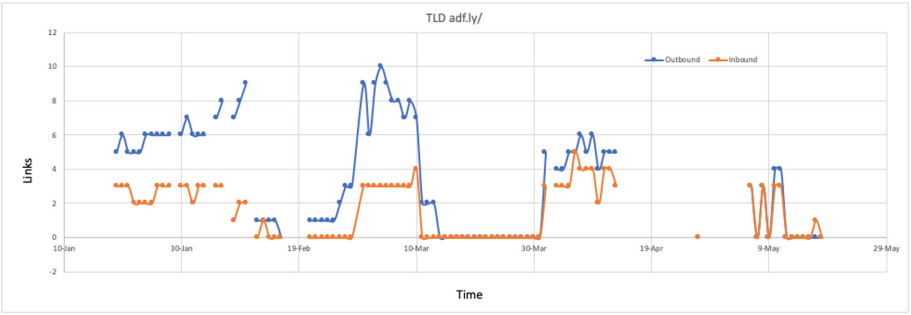

Fig. 13. The dynamics of a single TLD *adf.ly*

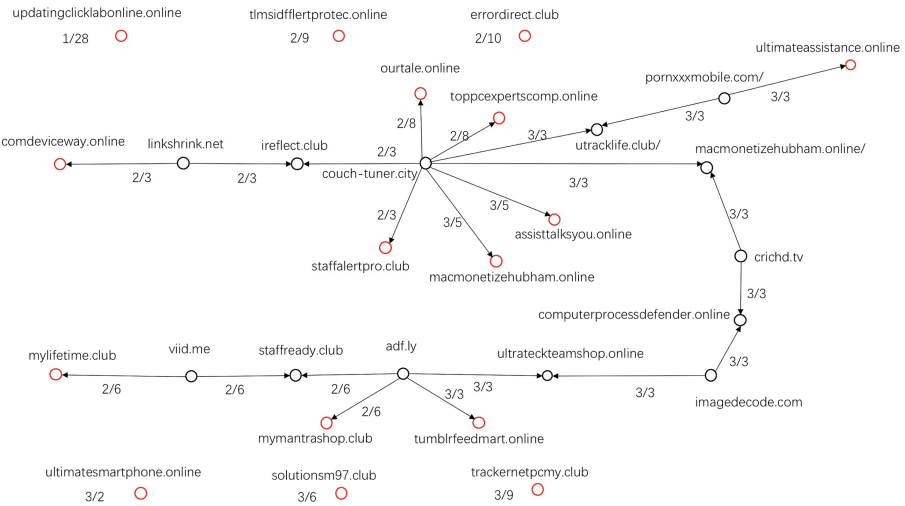

Fig. 14. The development of the single malware within clusters with time.

8 Conclusions

We developed a crawler to collect live malware distribution network data from publicly available sources including Google Safe Browser and VirusTotal. We then generated the graph with our visualization tool and performed malware attribution. We have discovered: 1) malware distribution networks form clusters; 2) those cluster sizes follow the Power Law; 3) there is a correlation between cluster size and the number of malware species in the cluster; 4) there is also a correlation between number of malware species and cyber events; and finally, 5) the infrastructure components such as bridges, hubs, and persistent links play significant roles in malware distribution dynamics.

Acknowledgement. The authors would like to thank VIS research assistants Sebastian Peryt, Pedro Pimentel, and Sihan Wang for participating in 3D model prototyping and data processing. This project is in part funded by Cyber-Security University Consortium of Northrop Grumman Corporation. The authors are grateful to the discussions with Drs. Neta Ezer, Justin King, and Paul Conoval.

References

1. Schiller, B., Deusser, C., Castrillon, J., Strufe, T.: Compile- and run-time approaches for the selection of efficient data structures for dynamic graph analysis. Appl. Network Sci. **1** (2016). Article number: 9 https://link.springer.com/article/10.1007/s41109-016-0011-2
2. DNA at GitHub. https://github.com/BenjaminSchiller/DNA

3. Carey, C.E.: Continued bot infiltration of Trump's Facebook Pages. Data for Democracy, 1 May 2017. https://medium.com/data-for-democracy/continued-bot-infiltration-of-trumps-facebook-pages-2df82ca86b5b
4. Gu, G., Perdisci, R., Zhang, J., Lee, W.: BotMiner: clustering analysis of network traffic for protocol- and structure-independent botnet detection. In: Proceedings of the 17th USENIX Security Symposium (Security 2008), (2008)
5. Gu, G., Zhang, J., Lee, W.: BotSniffer: detecting botnet command and control channels in network traffic. In: Proceedings of the 15th Annual Network and Distributed System Security Symposium (NDSS 2008), February 2008
6. McCoy, D., et al.: Pharmaleaks: understanding the business of online pharmaceutical affiliate programs. In: Proceedings of the 21st USENIX conference on Security symposium, ser. Security 2012, p. 1. USENIX Association, Berkeley (2012)
7. Karami, M., Damon, M.: Understanding the emerging threat of ddos-as-a-service. In: Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (2013)
8. Google safe browsing. https://developers.google.com/safe-browsing/
9. Zhang, J., Seifert, C., Stokes, J.W., Lee, W.: Arrow: Generating signatures to detect drive-by downloads. In: Srinivasan, S., Ramamritham, K., Kumar, A., Ravindra, M.P., Bertino, E., Kumar, R. (eds.) Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, 28 March–1 April 2011. ACM (2011)
10. Rossow, C., Dietrich, C., Bos, H.: Large-scale analysis of malware downloaders. In: Flegel, U., Markatos, E., Robertson, W. (eds.) DIMVA 2012. LNCS, vol. 7591, pp. 42–61. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37300-8_3
11. Caballero, J., Grier, C., Kreibich, C., Paxson, V.: Measuring pay-per-install: the commoditization of malware distribution. In: Proceedings of the 20th USENIX conference on Security, ser. SEC 2011. USENIX Association, Berkeley (2011)
12. Goncharov, M.: Traffic direction systems as malware distribution tools. Trend Micro, Technical report (2011)
13. Behfarshad, Z.: Survey of malware distribution networks. Electrical and Computer Engineering, University of British Columbia, Technical report (2012)
14. Provos, N., McNamee, D., Mavrommatis, P., Wang, K., Modadugu, N.: The ghost in the browser analysis of web-based malware. In: Proceedings of the first Conference on First Workshop on Hot Topics in Understanding Botnets, ser. HotBots 2007. USENIX Association, Berkeley (2007)
15. Provos, N., Mavrommatis, P., Rajab, M.A., Monrose, F.: All your iframes point to us. In: Proceedings of the 17th Conference on Security symposium, ser. SS 2008. USENIX Association, Berkeley (2008)
16. http://www.stachliu.com/2012/08/search-diggity-install/
17. http://virustotal.com
18. http://www.d3.org
19. Wigglesworth, V.B.: Insect Hormones, pp. 134–141. W.H. Freeman and Company (1970)
20. Cai, Y.: Instinctive Computing. Springer, London (2016). https://doi.org/10.1007/978-1-4471-7278-9
21. Bonabeau, E., Dorigo, M., Theraulaz, G.: Sawrm Intelligence: From Nature to Artificial Systems. Oxford University Press, Oxford (1999)
22. Cai, Y.: Ambient Diagnostics. CRC Press, Boca Raton (2014)
23. Jacobi, J.A., Benson, E.A., Linden, G.D.: Personalized recommendations of items represented within a database. US Patent. US 7113917 B2 (2006)

24. Peryt, S., Morales, J.A., Casey, W., Volkmann, A., Cai, Y.: Visualizing malware distribution network. In: IEEE Conference on Visualization for Security, Baltimore, October, 2016 (2016)
25. Rossi, R.A., Gallagher, B., Neville, J., Henderson, K.: Modeling dynamic behavior in large evolving graphs. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining (WSDM 2013), pp. 667–676. ACM, New York (2013). http://dx.doi.org/10.1145/2433396.2433479
26. Yu, S., Gu, G., Barnawi, A., Guo, S., Stojmenovic, I.: Malware propagation in large-scale networks. IEEE Trans. Knowl. Data Eng. **27**, 170–179 (2015)