Pheromone Model Based Visualization of Malware Distribution Networks

Yang Cai, Jose Andre Morales, Sihan Wang, Pedro Pimentel, William Casey and Aaron Volkmann

Carnegie Mellon University, 5000 Forbes Ave., PA 15213, USA
Corresponding email: ycai@cmu.edu

Abstract. We present a novel computational pheromone model for describing dynamic network behaviors in terms of transition, persistency, and hosting. The model consists of a three-dimensional force-directed graph with bi-directional pheromone deposit and decay paths. A data compression algorithm is developed to optimize computational performance. We applied the model for visual analysis of a Malware Distribution Network (MDN), a connected set of maliciously compromised domains used to disseminate malicious software to victimize computers and users. The MDN graphs are extracted from datasets from Google Safe Browsing (GSB) reports with malware attributions from VirusTotal. Our research shows that this novel approach reveals patterns of topological changes of the network over time, including the existence of persistent sub-networks and individual top-level domains critical to the successful operation of MDNs, as well as the dynamics of the topological changes on a daily basis. From the visualization, we observed notable clustering effects, and also noticed life span patterns for high-edge-count malware distribution clusters.

Keywords: pheromone, visualization, malware, malware distribution network, force-directed graph, biologically-inspired computing, security, dynamics, 3D graph, graph

1. Introduction

Pheromones in nature are chemical messages that act within a species. They are used widely by insects for communication within a community and within the body by means of hormones. This usage led to these substances also being referred to as "social hormones" [16]. Pheromones are external memories that are physically projected onto the ground or into the air, and are shared within a group. The dynamics of depositing and vaporizing pheromones are very sophisticated processes. These chemical messages have diverse biological effects and differ widely in their modes of action. In practice, the term "pheromone" proves useful in describing behaviors such as trail formation, defensive secretions, and social coherence. For the past several decades, several computational pheromone models have been proposed such as the classic ant colony optimization model [18] and a contemporary online shopping cart product recommendation model patented by Amazon [20]. Pheromone models have been used in visualization of human activities in CCTV footages, social media, and digital forensics [17, 19]. Digital pheromone models not only represent simultaneous localization, navigation, and optimization, but also provide perceptual motion

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

intelligence simulations such as low-pass filters and visual episodic memory. Consequently, digital pheromone models bridge collective intelligence and primitive intuition, enabling dynamic data modeling and motion pattern recognition for humans and machines.

In this study, we explore a novel pheromone model for visualization of topological changes of a large dynamic network. Similar to virus distribution networks in nature, a cyber malware distribution network (MDN) is a connected set of maliciously compromised top-level domains (TLDs) used to facilitate the dissemination of malicious software attempting to victimize computers. It acts like a platform for spreading malwares to other nodes in the network. The challenge here is that MDNs are normally hidden and constantly change over time. The topological structures, such as domains hosting and malware and acting as intermediaries assisting in distribution are not revealed until the network traffic data is collected, attributed from multiple data sources, and systematically plotted.

MDNs have been used in botnets [1-2], spam campaigns [3] and distributed denial of service attacks (DDoS) [4]. In general, MDNs are the essential back end distribution highway fueling underground economies in monetized schemes generating large revenues for malicious actors [3]. In this study, we provide a biologically-inspired approach for the construction and visualization of an MDN's topological structure. We collect and visualize an MDN data over a period of 9 month in order to gain insight on its structure, persistence, and evolution over time. Our MDN graphs were based on the Google Safe Browsing (GSB) transparency report [5] and malware attribution from the VirusTotal website. Our research shows the novel approach of leveraging GSB and VirusTotal reports to graph MDNs reveals deep insight into structural changes over time. The main contributions of this paper include the novel pheromone-based visualization model that reveals the existence of persistent sub-networks and individual TLDs critical to the successful operation of MDNs and use of crowdsourcing data from Google Safe Browsing and VirusTotal for constructing MDN networks.

2. Related Work

A corpus of research [6-7] has proposed various approaches to identify malicious URLs. Research such as [8-9] describe and analyze the use of MDNs and their various components in monetized schemes such as botnets, pay-per-install affiliate programs and traffic direction systems. In the study [10], Behfarshad describes MDNs as a set of landing pages, redirectors, and malware repositories. The authors suggest detecting the presence of an MDN by identifying drive by download attempts via two methods: top-down and bottom-up. The research of Provos, et al [11-12] is part of the early work describing web based malware and the existence and identification of MDNs. Their research provided insight on identifying malicious URLs and explains the critical role of *iframes*, which are a simple and widely accepted mechanism for redirection in http traffic, as the fundamental link binding multiple URLs together in an MDN. The culmination of their work is the Google Safe Browsing service that is the key data source for our research. We enhance the current literature by defining an MDN and visualizing a graphical topological structure including vertex role based

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

node types indicative of a TLD's role in malware distribution. The authors of this paper developed an early version of a 3D interactive visualization system that revealed basic distribution patterns from the GSB data collected in 2014 [21]. In this study, we want to advance visual analytics with a pheromone simulation model to explore the attributed malware distribution network data collected in 2017, in order to gain insight on network structure, evolution, and persistence over time.

3. Malware Distribution Network as a Graph

An MDN is a dynamic structure topologically consisting of interconnected TLDs. An MDN changes over time that is captured with topological structure visualization at different points in time. We formally define a graph capturing the temporal topological structure of an MDN. A representation of an MDN at a given point in time is defined as a directed graph M(t) = {V(t), E(t)} such that V(t) is a set of vertices at time t, $E(t)$ is a set of directed edges: $\{v_i(t), v_j(t)\} \in V(t)$ at time t. In a given graph $M(t)$, a vertex $v \in V(t)$ represents an MDN node in one or more modes. The full range of a node's modality in the MDN is unknown. We suggest, based on GSB reports, three possible roles: in an MDN, a node may act as an intermediary (MI) by facilitating malicious traffic, a malicious host (MH) or root malicious host (RMH) if malware files were hosted from that domain. In our visualizations, MI and MH nodes have incoming edges, but RMH nodes do not. This implies malware in RMH nodes came from some source other than those represented by a node in the MDN. Figure 1 illustrates an MDN structure.

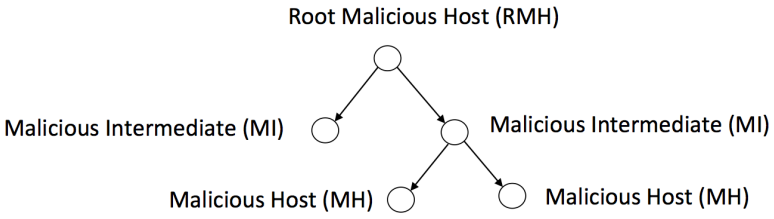

Fig. 1. Definition of a Malware Distribution Network

4. Pheromone Model

We first build a dynamic graph of the malware distribution network. 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 type – MI or MH. 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> where by destination is top-level domain which is reported as being affected by the source. To place all the top-level domains on the visualization, we used a Dynamic Behavioral Graph [22] to incorporate event frequencies, protocol

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

types, packet contents and data flow information into one graph. In contrast to a typical Force-Directed Graph such as D3 [16], 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 formula:

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

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

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 $||x_i - x_j||$ 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. See Figure 2.

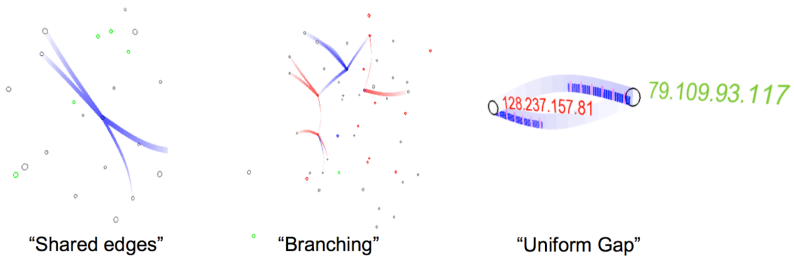

Fig. 2. Examples of the dynamic behavior graph of an MDN

Simple pheromone-based movement can produce sophisticated dynamic patterns. Conventional Ant Colony Optimization (ACO) models assume that insects walk along paths that connect various nodes. Pheromones can be overlaid in multiple layers. Furthermore, pheromones decay at a certain rate. If they did not decay, ants would risk repeating the same route and not respond to a rapidly changing environment. Here, we generalize pheromone deposit and decay on paths of a network. The amount of pheromones at a pixel position at time t is:

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

Decay: $$D(t) = max(\ u_i(t)\ -\ r \cdot t, L\ ) \tag{4}$$

where, $D(t)$ is the current pheromone level at a particular path 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 Figure 3.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

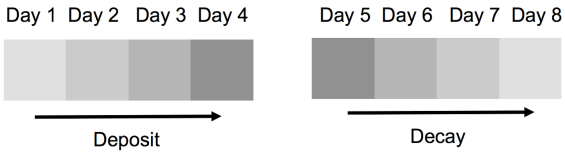

Fig. 3. Pheromone deposit and decay model for representing the persistency of the malware distribution channels (connected edges in the graph).

4. Graph Display Optimization

Whenever there is a new node entering the visualization space, the algorithm has to measure the distance between the new node and all the rest of the nodes nearby to ensure that they won't collide to each other, which is computationally expensive. In order to speed up the process, we optimize the repulsion calculation along the Z-axis. Here is the pseudo code of the method:

```
1.  Calculate repulsive forces
2.  Propagate backwards
3.  Calculate repulsion displacement from search_node to current_node
4.  Propagate forwards
5.  Calculate repulsion displacement from search_node to current_node
```

This enables to speed-up in orders of magnitude and to handle more nodes in the model. For example, before the optimization, it took 30 minutes to process 2,000 nodes. After the optimization, it takes less than 5 minutes to process the same amount of the nodes. See Figure 4.

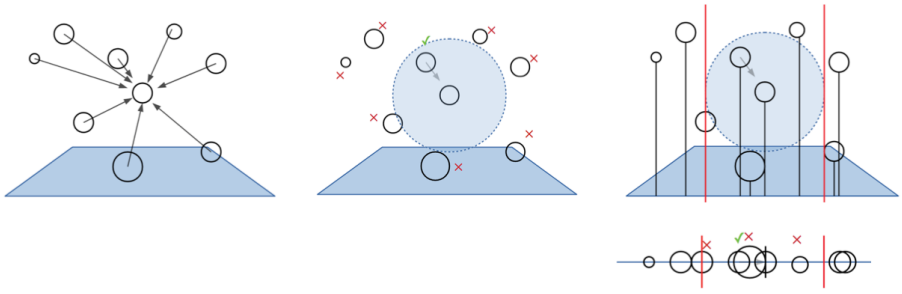

Fig. 4. A compression mapping along Z-axis in order to filter out the nodes that are further than a threshold.

5. Data Collection

We have collected the data for creating MDN graphs for a 9-month period from 19 January 2017 to 25 September 2017. We used the Google Safe Browsing (GSB) Transparency Report as our main source of data. We seeded the process by requesting a GSB report for the known websites such as notorious *vk.net* once every

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

24 hours. 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 as shown in Appendix A. The statistics of interest to us were labeled: *name, sendsToAttackSites, receivesTrafficFrom, sendsToIntermediarySites, lastVisitDate,* and *lastMaliciousDate.* If the returned report listed domains for labels 2, 3, or 4, those domains were added to the queue to be requested to GSB. This was an exhaustive recursive process that continued until a report with no more domains listed for these labels was received from GSB. We labeled nodes for domains listed with nonempty statistic 2 and 4 as an Intermediate Site (MI), and domains with empty statistic 3 as a Malicious Host (MH) or Root Malicious Host (RMH). The data provided in the GSB report only allowed us to create outgoing edges form the node for the current domain under analysis. All nodes are initially labeled as a Malicious Node (MN) and could be relabeled to MH or MI if the GSB report for a domain listed in 3 contained the current domain name in 2 or 4. If this was the case, an incoming edge can be created from the domain listed in 3 to the current domain under analysis. 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. Our collection process occurred daily starting at 9:00 am EST and required from 4 to 11 hours to complete, thus starting and finishing within the same calendar date. We decided on collecting just once a day after extensive manual analysis of the GSB report's value for label 6 over several weeks revealed GSB tended to perform diurnal updates of their report details.

6. Malware Attribution

We constructed a filtered data set because we want to retain significant patterns for effective visualizing and the design of such a filter must prevent over-trivializing the data. We therefore used VirusTotal scans sites that perform malware distribution on submitted URLs and files logs a report for each query. When given the list of malicious domains, VirusTotal gives the set of reports which list malware linked to the sites. Our malware-attributed data set is the set of TLDs that appear in at least one report of a known malware from VirusTotal. However, as reports are generated only when a malware scan query is sent to VirusTotal, there may not be a report corresponding precisely to the last malicious time reported by Google Safe Browsing. Therefore, the filtered dataset is the set of TLDs that have a report that marks the site as containing malware within 10 days of when Google Safe Browsing determined the site to be malicious. The visualization images provided are based upon this malware-attributed set. More specifically, the visualized dataset is the set of edges that either receive or send traffic from a site that has a properly timed report denoting it as malicious. See Appendix B and C for a sample of the VirusTotal report and the statistical summary of the collected data from GSB and VirusTotal. See Figure 5.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

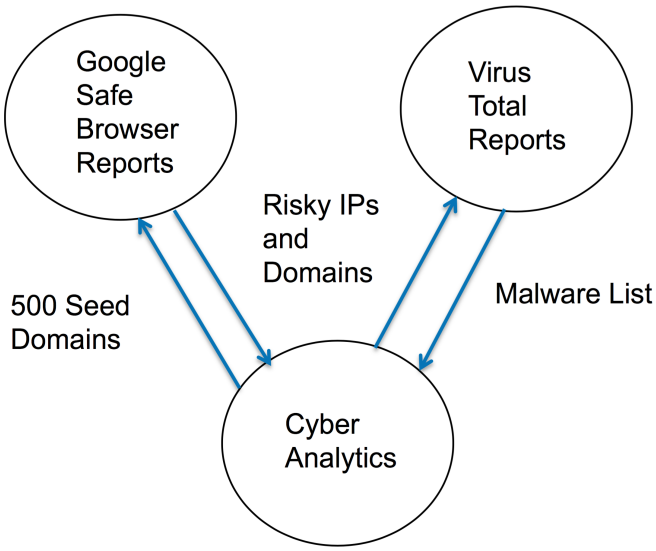

Fig. 5. The malware distribution data collection and attribution process

7. Visualization Results

The data we visualized are the malware-attributed data as described in the prior section, containing all edges that have at least one endpoint being a malware attributed site. Each edge is label with corresponding capture date time stamps, and rendered in chronological order, spanning across the entire 9 months of data. The rendering process iteratively adds each edge supplied to the graph, depending on whether both endpoints of the edge already exists in the graph, the visualization process will add nodes to the graph and then render a cold colored edge. Each day, the color of the existing edge becomes warmer until it reaches red, reflecting the age of each malicious link. If an existing edge is present in a later collection, the edge will be reassigned a cold color. When an edge is red, it will disappear when the visualization reaches the next day if the edge is not present in the data set for the next day. Given that each edge is directed, opaqueness of the edge demonstrates which node is sending traffic and which node is receiving traffic in the link. Figure 6 through 8 show the clusters of MDN evolved on January 19th, 20th, and 21st of 2017. Figure 9 shows a close-up of the MDN on September 25th of 2017.

From our visualization of the malware-attributed dataset, we can observe five key dynamic phenomena: First, having a clear distinction between days, we can observe the clustering effect of the malware nodes. Large clusters do not form rapidly. Instead, a cluster of significant size in general requires one to two weeks to form. It is rare for edges to be continuously added to a singular node in succession to form a super node, instead, smaller chains will first form over time before multiple chains gather together to form a large cluster. In particular, for the months of January, March, and June, we are able to observe a large centralized cluster of similar websites. This highlights potential large-scale malware attacks that may be occurring

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

during those particular periods and the similarities in the distribution network used for those attacks.

Second, the visualization process highlights the quantitative differences in number of malicious nodes that exist for any particular month visually. By examining the time stamps presented and the quantity of edges present on screen we can visually determine the change in quantity of data over time. For example, we can observe that the quantity of malicious nodes peaks in our dataset during February. We can also visually compare the outbound and inbound edge frequencies across different time periods as well as examine whether high-edge-count sites are scattered or closely interconnected.

Third, the visualization is the ease of determining and examining the site with traits of particular interest. For example, root malicious nodes (RMN) are easily identified visually as each edge has a gradient effect where the opaque end marks the source and the transparent end marks the destination, so a node with only outgoing edges can be identified as a root malicious host (RMH) while examining the visualization. Furthermore, because the visualization allows us to traverse the graph in 3D as well as zoom in and out, we can examine the changes in the nodes, such as a malicious root node (RMN), is connected to, as well as observe the chains that the root malicious nodes (RMN) are part of as they dissipated and form. For example, in Figure 9, we can zoom in to a specific node to see the node's address as well as the address of the adjacent nodes in the visualization.

Fourth, given that the data visualized is labeled with time stamps, we can observe changes per specific quantity of time. As mentioned the prior section, the data is filtered to those with a malware-attributed date within 10 days of the capture date. With this time frame in mind, when combined with the timestamps displayed in the visualization, we can view specific changes in whether a link is malicious or not and the time frame which it is malicious. Using this we can view whether past links are deprecated as new links are established and have approximations for overlapping active times in a malware distribution chain.

Finally, the timestamps are further supplemented by the pheromone effect. As describe at the beginning of this section, edges decay over time, with colors representing the age of the edge. The key benefit of this method is the ability to observe the frequency a link is used in malware distribution, at each time in a new collection, when an existing edge is mentioned again, it will be restored to the color that represents very recent. Therefore, in the duration which an edge remains in existence, the frequency of being reset to the coldest color will approximate the frequency the link was malicious. This pheromone method also highlights the order of deprecation in malware distribution chains by showing which links first becomes unused. By decaying edges gradually, we can more easily see changes in chains as malware distribution links shift over time. Overall, the pheromone effect in the visualization give visual feedback on such pattern in a visually identifiable manner.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

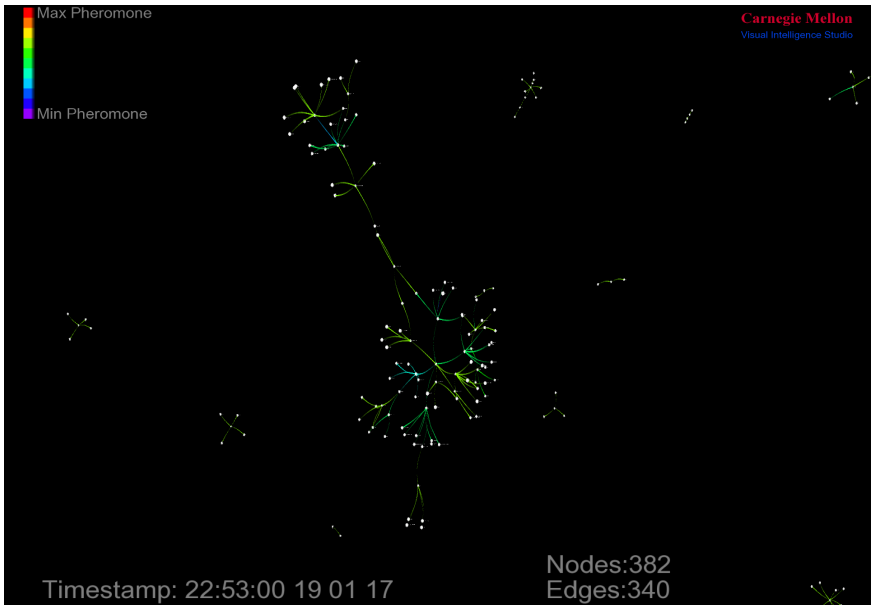

Fig. 6. A cluster of malware distribution network on January 19, 2017.

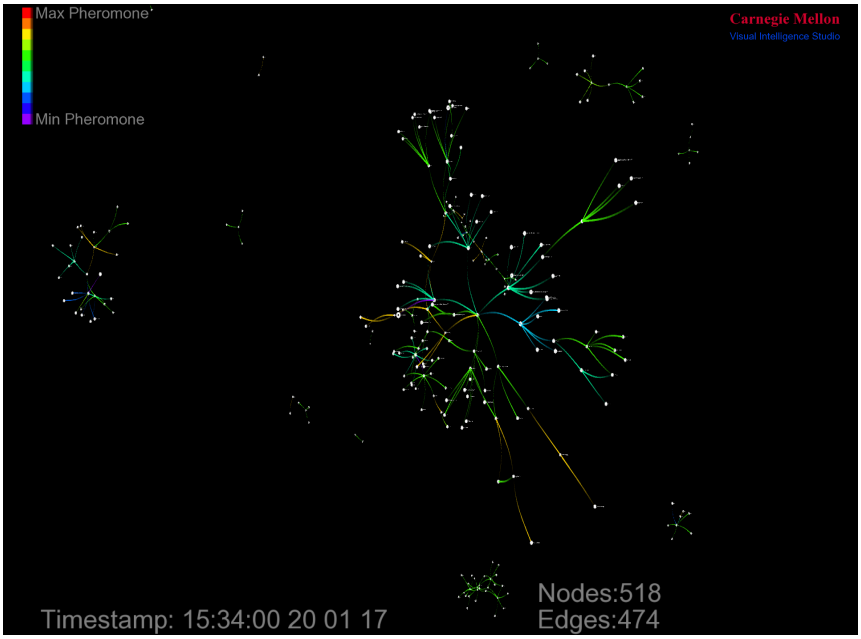

Fig. 7. A cluster of malware distribution network on January 20, 2017.

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

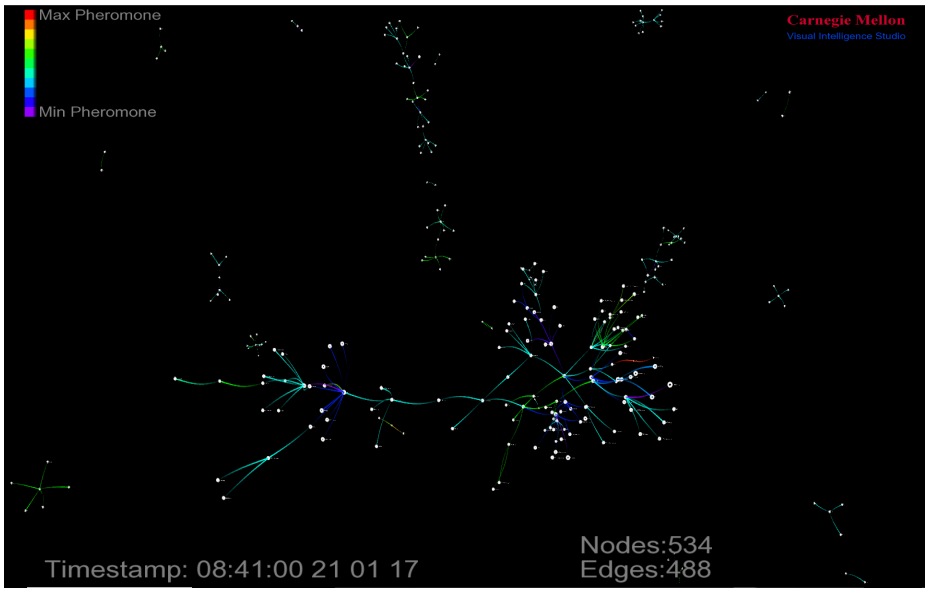

Fig. 8. A cluster of malware distribution network on January 21, 2017.

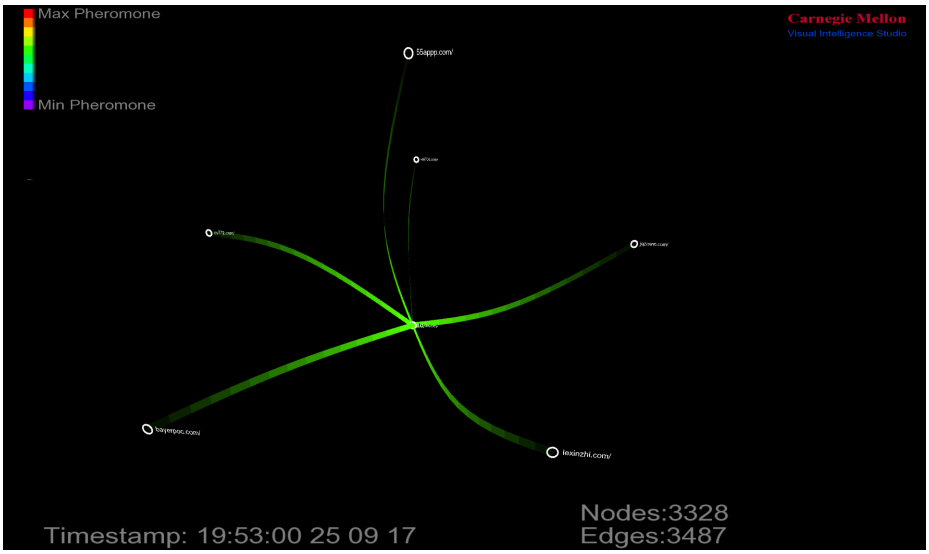

Fig. 9. A close-up of the cluster of malware distribution on September 25, 2017.

8. Conclusions

In this study, we provide a biologically-inspired approach for the construction and visualizing an MDN's topological structure. We collect and visualize an MDN data over a period of 9-month in order to gain insight on its structure, persistence, and

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

evolution over time. Our MDN graphs were based on Google Safe Browsing (GSB) reports and malware attribution from VirusTotal. Our research shows the novel approach of leveraging GSB and VirusTotal reports to graph MDNs reveals deep insight into structural changes over time. We found that the pheromone-based visualization model reveals the existence of persistent subnet works and individual TLDs' critical to the successful operation of MDNs.

The pheromone model shows the evolving MDN dynamics on daily basis, which enables analysts to zoom in and out, pause, replay, and fast-forward the animation interactively. Furthermore, we found use of crowdsourcing data from Google Safe Browsing and VirusTotal are critical for constructing MDN networks. We need to explore the open source space for better availability of data and more accuracy of the attribution of malware types, names, and contents.

From the visualization, we are able to notice significant malware distribution clusters. Specifically we found large malware distribution clusters in the months of January, March, and June. The visualization also highlighted how large distribution clusters require approximately two weeks to form and also an approximate life span of two weeks.

Acknowledgement

The authors would like to thank VIS research assistants Sebastian Peryt for initial 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, Robert Pike, Paul Conoval, and Donald Steiner. This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute. [Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution. Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University. DM-0004676

References

1. G. Gu, R. Perdisci, J. Zhang, and W. Lee, "BotMiner: Clustering analysis of network traffic for protocol- and structure-independent botnet detection," in Proc. of the 17th USENIX Security Symposium (Security'08), 2008.
2. G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting botnet command and control channels in network traffic," in Proc. of the 15th Annual Network and Distributed System Security Symposium (NDSS'08), February 2008.
3. D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko, "Pharmaleaks: understanding the business of online

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

pharmaceutical affiliate programs," in Proc. of the 21st USENIX conference on Security symposium, ser. Security'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 1–1.
4. M. Karami and M. Damon, "Understanding the emerging threat of ddos-as-a-service," in Proc. of the USENIX Workshop on Large-Scale Exploits and Emergent Threats, 2013.
5. "Google safe browsing," https://developers.google.com/safe-browsing/
6. J. Zhang, C. Seifert, J. W. Stokes, and W. Lee, "Arrow: Generating signatures to detect drive-by downloads," in Proc. of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011, S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, Eds. ACM, 2011.
7. C. Rossow, C. Dietrich, and H. Bos, "Large-scale analysis of malware downloaders," in Proc. of the 9th international conference on DIMVA, ser. DIMVA'12. Berlin, Heidelberg: Springer-Verlag, 2013.
8. J. Caballero, C. Grier, C. Kreibich, and V. Paxson, "Measuring pay-per-install: the commoditization of malware distribution," in Proc. of the 20th USENIX conference on Security, ser. SEC'11. Berkeley, CA, USA: USENIX Association, 2011.
9. M. Goncharov, "Traffic direction systems as malware distribution tools," Trend Micro, Tech. Rep., 2011.
10. Z. Behfarshad, "Survey of malware distribution networks," Electrical and Computer Engineering, University of British Columbia, Tech. Rep., 2012.
11. N. Provos, D. McNamee, P. Mavrommatis, K. Wang, and N. Modadugu, "The ghost in the browser analysis of web-based malware," in Proc. of the first conference on First Workshop on Hot Topics in Understanding Botnets, ser. HotBots'07. Berkeley, CA, USA: USENIX Association, 2007.
12. N. Provos, P. Mavrommatis, M. A. Rajab, and F. Monrose, "All your iframes point to us," in Proc. of the 17th conference on Security symposium, ser. SS'08. Berkeley, CA, USA: USENIX Association, 2008.
13. http://www.stachliu.com/2012/08/search-diggity-install/.
14. "Bing linkfromdomain search operator," http://www.bing.com/blogs/site_blogs/b/search/archive/2006/10/16/search-macros-linkfromdomain.aspx
15. http://www.d3.org
16. Wigglesworth VB (1970) Insect Hormones. pp. 134 - pp.141. W.H. Freeman and Company.
17. Cai Y (2016) Instinctive Computing. Springer-London.
18. Bonabeau E. Dorigo M. and Theraulaz G. (1999): Sawrm Intelligence: From Nature to Artificial Systems. Oxford University Press
19. Cai Y (2014) Ambient Diagnostics. CRC Press.
20. Jacobi JA, Benson E.A. and Linden GD. Personalized recommendations of items represented within a database. US Patent. US 7113917 B2
21. Peryt S, Morales JA, Casey W, Volkmann A, and Cai Y (2016): Visualizing Malware Distribution Network, IEEE Conference on Visualization for Security, Baltimore, October, 2016
22. Ryan A. Rossi, Brian Gallagher, Jennifer Neville, and Keith Henderson. 2013. Modeling dynamic behavior in large evolving graphs. In *Proceedings of the sixth ACM international conference on Web search and data mining* (WSDM '13). ACM, New York, NY, USA, 667-676. DOI=http://dx.doi.org/10.1145/2433396.2433479

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5

Appendix A: Sample GSB Data in JSON Format:


```
{"date": 1484884380.0, "website": {"name": "nowcheck247freshandforfree. online/",
"partialUnknownDowHosts": [], "partialMalwareDowHosts": [], "malwareSite":
{"sendsToAttackSites": ["milleniumforum.info/"], "receivesTrafficFrom":
["veryhotmom.com/"], "type": 8, "sendsToIntermediarySites": []},
"uwsDownloadListStatus": "unlisted", "partialUwsDowHosts": [], "uwsListStatus":
"unlisted", "unknownDownloadListStatus": "unlisted", "partialUwsHosts": [],
"malwareDownloadListStatus": "unlisted", "partialSocialEngHosts": [],
"malwareListStatus": "unlisted", "numAses": 1, "asList": ["AS12876 (AS12876)"],
"numListedTimes": 0, "partialMalwareHosts": [], "socialListStatus": "listed"},
"lastVisitDate": 1482084799, "numTested": 7, "as": {}, "dataUpdatedDate":
1484871006, "lastMaliciousDate": 0}
```


Appendix B: Sample VirusTotal Data in CSV Format:


```
ameritag.com,1493672220.0
Trojan.Script.682678,JS.Dropper.JU,JS.Trojan.Iframe.nm,Trojan.Gen.7,
JS/Iframe.MO,JS:Iframe-EPR [Trj], HEUR:Trojan.Script.Generic,
Trojan.Script.682678, Trojan.Script.Iframe.ecnmvw, Troj.Script.Generic!c,
Trojan.Script.682678, Trojan.Script.682678(B), TrojWare.JS.Iframe.MO,
Trojan.Script.682678,SCRIPT.Virus, JS_AXPERGLE.SM, HTML/ExpKit.Gen2,
VirTool:JS/Obfuscator, JS.Z.Agent.4728.I[h],
HEUR:Trojan.Script.Generic,Script.Trojan.IFrame.AL,
Trojan.Script.682678,Win32.Script.Agent.Pgwl, Trojan.HTML.Framer,
JS/Moat.6D1444D0!tr, HTML/Framer,virus.js.qexvmc.1
```


Appendix C: Statistics of the 9-month dataset.

Months	Nodes	Inbound Edges	Outbound Edges	Attributed Nodes
January *	14810	19199	36655	156
February	18660	24830	46896	222
March	15664	20229	36084	173
April	15880	21251	36719	188
May	14620	19953	34129	125
June**	12352	17055	30204	178
July	12057	18111	33202	136
August	8813	12272	22350	88
September***	7731	10699	21122	45

* Started on January 19, 2017
** There were 5 days down time
*** The ending date was September 25, 2017

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93698-7_5