Detecting Malicious Domains via Graph Inference

[Extended Abstract] *

Pratyusa K. Manadhata
Hewlett-Packard Laboratories
Princeton, NJ
manadhata@hp.com

Sandeep Yadav
Damballa Inc.
Atlanta, GA
sandeepvaday@gmail.com

Prasad Rao
Hewlett-Packard Laboratories
Princeton, NJ
prasad.rao@hp.com

William Horne
Hewlett-Packard Laboratories
Princeton, NJ
william.horne@hp.com

ABSTRACT

Organizations, especially business enterprises, collect and store event logs generated by hardware devices and software applications in their networks. For example, firewalls log information about suspicious network traffic; and hypertext transfer protocol (HTTP) proxy servers log websites (or domains) accessed by hosts in an enterprise. Enterprises collect and store event logs primarily for two reasons: need for regulatory compliance and post-hoc forensic analysis to detect security breaches. Also, availability of cheap storage has facilitated large scale log collection. Such logs generated by both security products and non-security infrastructure elements are a treasure trove of security information. For example, if hosts in an enterprise network are infected with bots, then the bots may contact their command and control server over domain name system (DNS) and may exfiltrate sensitive data over HTTP. Hence both DNS logs and HTTP proxy logs will contain information about bot activities. Developing scalable and accurate techniques for identifying threats from event logs, however, is a challenging problem. In this paper, we introduce a big data analysis approach to identify malicious domains accessed by hosts in an enterprise from the enterprise's event logs.

Malware infections spread via many vectors such as drive-by downloads, removable drives, and social engineering. Malicious domain accesses, however, account for majority of host infections. Hence to contain malware, enterprises must prevent their hosts from accessing malicious domains. Reliable and scalable identification of malicious domains, however, is challenging. Enterprises use both commercial and freely available domain blacklists, i.e., list of known malicious domains, to identify and prevent malicious domain accesses; such lists, however, incur a significant delay in adding new domains as they rely on many manual and automated sources. Moreover, the techniques used to generate the lists are resource intensive. For example, malicious domain inference using DNS and network properties such as a domain's IP addresses and BGP prefixes requires data collection from sources with specialized vantage points. Similarly, machine learning techniques, e.g., analyzing a domain's lexical features, or a related IP address's structural properties, requires large feature data sets and accurately labeled training sets; hence these techniques are computationally expensive and may suffer from increased delay in detection.

In this paper, we present a scalable malicious domain detection approach that uses event logs routinely collected by enterprises, requires no additional data collection, and uses minimal training data. We model the detection problem as an *inference* problem on very large graphs. Our graph inference approach utilizes inherent malware communication structure, e.g., all bots in an enterprise contact the same command and control server. We first construct a *host-domain graph* by adding a node for each host in the enterprise and each domain accessed by the hosts, and then adding edges between each host in the enterprise and the domains visited by the host. We *seed* the graph with ground truth information about a small fraction of domains obtained from domain blacklists and whitelists, i.e., we label a small fraction of domains as malicious and benign, and label the rest of the domains as unknown. Given the host-domain graph and the ground truth information, our goal is to *infer* the states of the unknown domains in the graph. Formally, we would like to compute a node's *marginal probability* of being in a state, i.e., the probability of the node being in a state given the states of other nodes in the graph. We then label the nodes with high marginal probability of being malicious as malicious nodes and benign otherwise.

Marginal probability estimation in graphs is known to be NP-complete [2]. *Belief propagation* is a fast and approximate technique to estimate marginal probabilities [1, 2]. BP's time complexity and space complexity are linear in the number of edges in a graph; hence BP scales well to large graphs. A typical enterprise host-domain graph has millions of nodes and tens of millions of edges. Hence we adapt belief propagation to estimate an unknown domain's likelihood of

[*] A full version of this paper appears in the proceedings of the 19th European Symposium on Research in Computer Security (ESORICS) 2014.

AISec'14, November 7, 2014, Scottsdale, Arizona, USA.
ACM 978-1-4503-2953-1/14/11.
http://dx.doi.org/10.1145/2666652.2666659.

being malicious; if the likelihood is more than a threshold, we identify the domain to be malicious.

An HTTP proxy acts as an intermediary between an enterprise's hosts and the domains accessed by the hosts. Hence we can determine the domains visited by the hosts from proxy logs and construct an enterprise's host-domain graph. We applied our approach to 7 months of HTTP proxy data collected at a large global enterprise. Our results show that with minimal ground truth information, e.g., with only 1.45% nodes in the graph, we achieve high true positive rates, e.g., 95.2%, with low false positive rates, e.g., 0.68%. A benign node labeled as malicious by our approach is a false positive and a correctly identified malicious node is a true positive. Our approach takes the order of minutes to analyze a large-sized enterprise's day long data, and identifies previously unknown malicious domains.

We make the following contributions in this paper.

- We demonstrate that we can extract actionable security information from enterprise event logs in a scalable and reliable manner.
- We model the malicious domain detection problem as a graph inference problem and adapt belief propagation to solve the problem. Our approach does not require additional data beyond event logs, does not compute features, and uses minimal data from existing blacklists and whitelists.
- We apply our approach to event logs collected at a global enterprise over 7 months and show that our approach scales well and identifies new malicious domains not present in the blacklists.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection (e.g., firewalls)*; K.6.5 [**Management Of Computing and Information Systems**]: Security and Protection

General Terms

Security

Keywords

Belief propagation, big data analytics, graph inference, malicious domain detection

1. REFERENCES

[1] J. Pearl. Reverend Bayes on inference engines: a distributed hierarchical approach. In *in Proceedings of the National Conference on Artificial Intelligence*, 1982.

[2] J. Yedida, W. Freeman, and Y. Weiss. Understanding Belief Propagation and its Generalizations. *Exploring Artificial Intelligence in the New Millennium*, 2003.