Enhancing Network Security: Host Trustworthiness Estimation

Tomáš Jirsík, Pavel Čeleda

Institute of Computer Science,
Masaryk University, Brno, Czech Republic
{jirsik|celeda}@ics.muni.cz

Abstract. Network connected devices has become inherent part of our lives. These devices have come to be more and more mobile and are target of various malware attacks. An inability to guarantee or check proper security settings of such devices poses a serious risk to network security. In this paper we propose a novel concept of flow based host trustworthiness estimation. The estimated trustworthiness determines a level of the risk to the network security the host posses. This concept enables network operators to identify a potential dangerous host in their network and take an appropriate precautions. Models used for trustworthiness estimation are based on scoring either single events or host characteristics. In order to be able to estimate trustworthiness of a host even in large scale networks, the data used for estimation are reduced only to extended network flows. The research is in its initial phase and will conclude with Ph.D. thesis in three years.

Keywords: network flow, host, model, trustworthiness, scoring

1 Introduction

The development of mobile devices capable of internet connection increases the diversity of hosts connected to a network. The current trend is to grant almost anybody a permission to bring and connect own device. The access is granted even to devices, whose security status is unknown. The unsecured device can then unintentionally serve as an entry point for an attacker. Moreover, it is hard, or even impossible to reach the devices physically in order to check their proper settings. Remote access to them is also undesirable due to privacy issues.

Given the circumstances described above, the network operators need a way, how to asses a level of risk they undertake when they let a device connect into their network. The level of the risk the device poses can be represented by the concept of trustworthiness. The more trustworthy operators find the device, the lower risk the device poses to the inner security of the network. Based on the trustworthiness estimation the operators can consequently take adequate measures when allowing the device to connect to the network like a deployment of extended security precautions.

As in the case of security settings, it is not efficient to collect data needed for trustworthiness estimation directly from the devices, especially in the large scale networks. A better way is to employ passive network monitoring. Monitoring devices like FlowMon [8], nProbe [14] and YAF [7] are capable of collecting aggregated representation of all connections in the network, called network flows. This mean of monitoring enables us to monitor all devices even in high speed networks. Based on acquired flow data we can then estimate *host trustworthiness*. By the term *host* we understand a device connected into the network.

Naturally, we do not intend to estimate the trustworthiness of all hosts in the Internet, we want to focus on hosts located within the monitored network. We stand for the concept of cleaning our house *(network)* first [6]. This approach also ensures sufficient data from flow network monitoring to determine the potentially malicious hosts. The results of our approach will have impact on network monitoring management and reduces amount of resources needed for maintaining network security.

The aim of this paper is to propose a concept for a host trustworthiness estimation as a topic of prospective Ph.D. thesis. In the Section 2 we discuss problem and possible pitfalls of this approach. Section 3 states research questions based on previous discussion and describes proposed approach to problem solution. Section 4 presents final considerations.

2 Problem Disscussion

The aim of this paper is to asses an estimate of host trustworthiness. The estimate should indicate a potentially danger host and must employ flow based monitoring to be able to monitor whole traffic. Using analogy to financial sector and credit scoring [17], we can regard the estimate as a measure, which represents a *trustworthiness of a host*. The greater risk to the network security the host poses, the lower is its trustworthiness. Features chosen to estimate the measure of trustworthiness should properly represent host behavior. There are papers [5,12] and surveys [9,3] which discuss flow-based host behavior classification. The authors classify host behavior into several categories based on predefined features observation. Nevertheless, they do not classify host in terms of security nor use any model to asses individual host trustworthiness.

Considering the fact that we intend to use only flow data, there are certain issues arising that need to be solved in order to be able to estimate the trustworthiness:

1. Host Identification. In the flow data, a host is identified by IP address. Using a MAC address for host identification is meaningless in this case since the monitoring devices collecting flow data are usually located at routers therefore they do not know the MAC address of the host. They know only MAC address of adjacent router. In order to be able to properly collect all hosts characteristics, *the host must be represented unambiguously.* An IP address can serve as unique host identificator in static addressed networks. In dynamic addressed networks, though, a host obtains a different IP address anytime it connects to the network.

Therefore a host cannot be unambiguously identified via its IP address and we need to find a different approach to the host identification.

2. Per Host Characteristics. Flow records were originally developed for network monitoring and they represent a network traffic from the connection point of view. Also measured characteristics reflects this original purpose: majority of them are counted *per flow*. However, a host monitoring requires a rearrangement of the flow records. They should be stored in a *per host* manner. This shift opens possibilities for definition of new host based characteristics obtained from network traffic. Further, we need to develop a suitable type of model for trustworthiness estimation and identify which characteristics should be used as model variables.

We believe it to be a promising approach to focus on the flow monitoring from the host point of view. Solving the above mentioned issues enables us to estimate the host trustworthiness and identify risk. Moreover the results can be used for optimization of privileges granting or improved allocation of resources for traffic monitoring.

3 Research Questions and Proposed Approach

The goal of this research is ***to estimate a host trustworthiness based only on flow information***. Apart from necessary engineering task, which includes a development of tools for storing, processing and querying flow data in host based format, we identify following research questions based on previously described problems and state proposed approach to each question:

1. *Is it possible to unambiguously identify a host employing only flow information?*

Solving research question 1 is crucial to further collection of host characteristics. We intend to identify a set of host specific flow elements, among others OS type, port distribution. We can also employ elements from application layer, e.g., User Agent in HTTP protocol. To the best of our knowledge, there is a lack of literature on flow-based host identification. There is literature on user identification [13] and on means for passive OS fingerprinting or NAT detection [11,1], though. Further, we consider to implement some results of host classification research, such as detecting type of the host [9]. A combination of these elements should be specific enough to provide us a unique host identification. In case we find an unambiguous host identification impossible, we imply a set of assumptions on the host in order to be able model host trustworthiness.

2. *How can the trustworthiness of a host be estimated? What features should be used for the estimation?*

Regarding the research question 2, the first step is to define trustworthiness, then we select monitored features. The selection of monitored features is

dependent on the type of model used for estimation. So far, we considered implementation of two types of models. First, inspired by credit scoring, we propose a model based on scoring host behavior. An example of features representing host behavior is shown in [5]. The features are going to be scored based on their potential security risk and the total score for a particular host is going to be computed. The second model is similar to previous one, just instead of scoring flow features we score flow events. The flow event is an abstraction of sequence of related flows. Typical example of such an event is a visit to a web site - a single event, visiting a web site, generates many flows as additional contents of the website, e.g. banners, are downloaded. Identification of the events is a challenging task since we need to discover relations between particular flows in order to group them into a single event [4].

3. *What methodology should be used to evaluate proposed models for host trustworthiness estimation?*

Answering research question 3 is essential for evaluation of proposed models. Since it is almost impossible to find the ground truth in real-world measured data [15], usual means of model evaluation based on false/true positive/negative rates, like receiver operating characteristic [2], are inapplicable. There exist datasets, which contain labeled traces and we could distill ground truth from them [16]. Many of them are rather old, though, and they do not represent real world traffic properly [2]. To find the ground truth, we are developing a cloud-based security research testbed [10] that is designed for controlled simulation of network attacks. Using this testbed, we intend to carry out controlled experiments that enable us to measure a very-close-to-real-world data. Based on this data we evaluate the models by classical methods of model evaluations. We plan to asses the performance of method for unambiguous host identification in a similar way.

4 Conclusion

To enhance network security we have proposed a concept of host trustworthiness estimation. Since our research is in its early stage, we present a description of the problem and state research questions. Further, we briefly introduce intended approach to the problem. So far, we have developed necessary tools for collecting data and flow information extension, e.g., HTTP monitoring [18], and we have implemented the testbed for the model evaluation [10].

We are painfully aware of privacy issues arising and we want to take strict precautions. We intend to give an proper attention to this area and by no means we do not link a user identity to a host to maintain the privacy of the users.

The main purpose of this proposed research is to asses a security risk the host poses. Nevertheless, results of the research are not limited only to network security management. They can be used to design special tailored services, optimization of the quality of the services or for traffic shaping.

Acknowledgments. This material is based upon work supported by Cybernetic Proving Ground project (VG20132015103) funded by the Ministry of the Interior of the Czech Republic.

References

1. Abt, S., Dietz, C., Baier, H., Petrović, S.: Passive Remote Source NAT Detection Using Behavior Statistics Derived from NetFlow. In: Doyen, G., Waldburger, M., Čeleda, P., Sperotto, A., Stiller, B. (eds.) Emerging Management Mechanisms for the Future Internet, Lecture Notes in Computer Science, vol. 7943, pp. 148–159. Springer Berlin Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-38998-6_18
2. Bhuyan, M., Bhattacharyya, D., Kalita, J.: Network anomaly detection: Methods, systems and tools. Communications Surveys Tutorials, IEEE PP(99), 1–34 (2013)
3. Callado, A.C., Kamienski, C.A., Szabo, G., Gero, B.P., Kelner, J., Fernandes, S.F.L., Sadok, D.F.H.: A survey on internet traffic identification. IEEE Communications Surveys and Tutorials 11(3), 37–52 (2009)
4. Caracas, A., Kind, A., Gantenbein, D., Fussenegger, S., Dechouniotis, D.: Mining semantic relations using NetFlow. In: Business-driven IT Management, 2008. BDIM 2008. 3rd IEEE/IFIP International Workshop on. pp. 110–111 (2008)
5. Dewaele, G., Himura, Y., Borgnat, P., Fukuda, K., Abry, P., Michel, O., Fontugne, R., Cho, K., Esaki, H.: Unsupervised host behavior classification from connection patterns. Int. J. Netw. Manag. 20(5), 317–337 (Sep 2010), http://dx.doi.org/10.1002/nem.750
6. François, J., Moura, G.C.M., Pras, A.: Cleaning your house first: Shifting the paradigm on how to secure networks. In: Proceedings of the 5th International Conference on Autonomous Infrastructure, Management, and Security: Managing the Dynamics of Networks and Services. pp. 1–12. AIMS'11, Springer-Verlag, Berlin, Heidelberg (2011), http://dl.acm.org/citation.cfm?id=2022216.2022218
7. Inacio, C.M., Trammell, B.: YAF: Yet Another Flowmeter. In: Proceedings of the 24th international conference on Large installation system administration. pp. 1–16. LISA'10, USENIX Association, Berkeley, CA, USA (2010), http://dl.acm.org/citation.cfm?id=1924976.1924987
8. INVEA-TECH: FlowMon Exporter – Community Program (2013), http://www.invea-tech.com, [cited 2014-01-23]
9. Kim, H., Claffy, K., Fomenkov, M., Barman, D., Faloutsos, M., Lee, K.: Internet traffic classification demystified: Myths, caveats, and the best practices. In: Proceedings of the 2008 ACM CoNEXT Conference. pp. 11:1–11:12. CoNEXT '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1544012.1544023
10. Kouřil, D., Rebok, T., Jirsík, T., Čegan, J., Drašar, M., Vizváry, M., Vykopal, J.: Cloud-based Testbed for Simulation of Cyber Attacks. In: Proceedings of the 2014 IEEE Network Operations and Management Symposium, NOMS 2014 (2014), to appear
11. Krmicek, V., Vykopal, J., Krejci, R.: Netflow based system for nat detection. In: Proceedings of the 5th International Student Workshop on Emerging Networking Experiments and Technologies. pp. 23–24. Co-Next Student Workshop '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1658997.1659010

12. McHugh, J., McLeod, R., Nagaonkar, V.: Passive network forensics: Behavioural classification of network hosts based on connection patterns. SIGOPS Oper. Syst. Rev. 42(3), 99–111 (Apr 2008), http://doi.acm.org/10.1145/1368506.1368520
13. Melnikov, N., Schönwälder, J.: Cybermetrics: User identification through network flow analysis. In: Stiller, B., Turck, F. (eds.) Mechanisms for Autonomous Management of Networks and Services, Lecture Notes in Computer Science, vol. 6155, pp. 167–170. Springer Berlin Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-13986-4_24
14. ntop: nProbe (2014), http://www.ntop.org/products/nprobe/, [cited 2014-01-23]
15. Ringberg, H., Roughan, M., Rexford, J.: The need for simulation in evaluating anomaly detectors. SIGCOMM Comput. Commun. Rev. 38(1), 55–59 (Jan 2008), http://doi.acm.org/10.1145/1341431.1341443
16. Stolfo, S., Fan, W., Lee, W., Prodromidis, A., Chan, P.: Cost-based modeling for fraud and intrusion detection: results from the jam project. In: DARPA Information Survivability Conference and Exposition, 2000. DISCEX '00. Proceedings. vol. 2, pp. 130–144 vol.2 (2000)
17. Thomas, L.C., Crook, J., Edelman, D.: Credit Scoring and Its Applications. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2002)
18. Velan, P., Jirsík, T., Čeleda, P.: Design and Evaluation of HTTP Protocol Parsers for IPFIX Measurement. In: Bauschert, T. (ed.) Advances in Communication Networking, Lecture Notes in Computer Science, Vol. 8115. pp. 136–147. Springer Berlin / Heidelberg, Heidelberg (2013)