Evaluating Reputation of Internet Entities

Václav Bartoš, Jan Kořenek

Evaluating Reputation of Internet Entities

Václav Bartoš[1,2], Jan Kořenek[1]

[1] Faculty of Information Technology, Brno University of Technology, Czech Republic
[2] CESNET a.l.e., Prague, Czech Republic
bartos@cesnet.cz, korenek@fit.vutbr.cz

Abstract. Security monitoring tools, such as honeypots, IDS, behavioral analysis or anomaly detection systems, generate large amounts of security events or alerts. These alerts are often shared within some communities using various alert sharing systems. Our research is focused on analysis of the huge amount of data present in these systems. In this work we focus on summarizing all alerts and other information known about a network entity into a measure called *reputation score* expressing the level of threat the entity poses. Computation of the reputation score is based on estimating probability of future attacks caused by the entity.

1 Introduction

Network operators today often recognize the need to monitor their networks. This includes security monitoring, *i. e.* deployment of various detectors of malicious or unwanted traffic, such as honeypots, IDS, behavioral analysis or anomaly detection. These systems can generate large amounts of *security events* or *alerts*. In large networks with many such detectors, or when alerts are exchanged among several organizations via some alert sharing system (such as Warden, AbuseHelper, n6, *etc.* [6]), the number of such alerts may be very large (millions per day [1]). We believe that analysis of such amount of data can reveal interesting characteristics of sources of malicious traffic. In particular, we want to label them by estimated measure of threat they pose, which we call *reputation score*.

It is known that network attacks are generated mostly by hosts infected with malware allowing attackers to control them remotely. Once a host is compromised, it often stays compromised for some time and therefore many security events can be caused by this single host. This fact is often used for spam mitigation, where lists of known malicious IP addresses (blacklists) are used to block known sources of spam. The same principle can be used to deal with other kinds of malicious traffic as well. For many attack[1] types, it is common to see the same IP address reported as malicious repeatedly for a long time [1, 2, 9].

However, while blacklists are easy to use, the information they provide is very limited. It is a binary information only – an address is either listed or not, surely bad or surely good, nothing in between. Moreover, there is no information about

[1] For simplicity, all kinds of malicious or unwanted traffic, including spam or port scanning, are called *attacks* in this paper.

why it was listed or when. Also, it has been shown that malicious IP addresses are distributed non-uniformly both geographically and in IP space [1, 3, 10]. Some networks, autonomous systems or countries (called *bad neighborhoods* in [10]) host significantly more malicious hosts than others. Therefore, in some cases, the sole fact that an IP address belongs to such a bad neighborhood may be enough for it to be suspicious, although the address itself has never been reported as malicious. This phenomenon cannot be covered by classic blacklists.

We envision a much richer source of information about misbehaving entities, which we call a *reputation database*. It will gather alerts from large number of detectors (via some of the existing alert sharing systems) and keep information about all network entities (not only IP addresses, but also network prefixes, domains, *etc.*) reported as sources of malicious behavior, including information on reasons why the entity was listed, when and by whom. It will further enhance this information by data from external sources, such as geolocation, information from DNS (*e. g.* hostnames assigned to IP addresses), other databases or blacklists. In general, it will gather as much security related information about the reported entities as possible. All this information will be provided to security teams to help them to protect their networks and investigate incidents.

An important part of the system should be an algorithm summarizing all the information known about an entity, *i. e.* its reputation, into a single number representing a measure of threat the entity poses – its *reputation score*. For example, it should allow to easily differentiate between an address which sent a single spam email a week ago and an address which tries to break into password protected services by dictionary attacks every day for last two months. It thus allows to quickly decide on which problems to focus first, or to easily create blocklists by getting a top-n list of IP addresses with the worst reputation score, for example. The goal of our research is to find a method to evaluate the reputation of an entity numerically by computing its reputation score.

2 Proposed approach

The first thing that has to be done is to formally define the meaning of reputation score. In common language, the word "reputation" expresses the common opinion people have about someone or something (an entity) [4, 8]. It is based on shared experience with past behavior of the entity. But although it is based on the past, it is intended to describe the most likely state in the near future and thus to help with current decisions. Similarly, reputation score of a network entity should be based on the history of security incidents caused by the entity, but it should represent the level of threat the entity poses now and in the near future. Therefore, we formally define it as follows:

> *Reputation score* of a network entity (*e. g.* an IP address) represents the *probability* that the entity will perform a *malicious activity* in the *near future* (*e. g.* next 24h), based on its past behavior and other information.[2]

[2] Ideally, the probability should be combined with anticipated severity of the malicious activity. Such variant is much more complex and is not covered by this short paper.

This definition means that evaluation of the reputation score must be based on prediction of future attacks. Our approach to build such a predictor is as follows.

The input of the prediction algorithm will be mainly a summary of all malicious events reported in some past time window. They can be supplemented by various other inputs related to the entity and the threat it potentially poses. For an IP address, it may be the country and the autonomous system it belongs to, whether it is listed on some of the public blacklists, whether its hostname can be resolved using a reverse DNS query, or whether the address is dynamically assigned or there is NAT (which can be sometimes guessed from the hostname).

All this information form the input of an algorithm, whose output should be the probability that the given entity will behave maliciously in a specified future time window. However, due to the number of inputs, their diversity and potential interdependencies, it is unfeasible to design such algorithm by hand. Our approach is to use some of the supervised machine learning methods to infer the algorithm from the data.

Indeed, this task is suitable for supervised learning, since it is easy to get a training set. For example, consider we have a week long sample of alerts. For each malicious IP address in the sample, we can get information about it from the first six days as an input and information whether it behaved maliciously in the last day as the expected output. By repeating this with data from several weeks it is possible to get very large labeled dataset. Moreover, the algorithm can constantly improve itself during operation by comparing its prediction with actually detected attacks.

2.1 Challenges

Besides issues connected to the machine learning itself, we foresee several other non-trivial problems that have to be solved. Some of them are briefly discussed here.

Reputation in a context. The definition of reputation score above implies that there may not be a single score for a given entity. It depends on parameters, such as length of time window for prediction, and context, *i. e.* which kind of malicious activity we are interested in. It should be possible to compute an overall reputation score using some appropriate set of parameters and by combining probabilities of various kinds of predicted attacks. But in many cases a more specific score will be needed, *e. g.* focusing on a specific kind of attack. For example, an IP address may be known as a frequent source of SYN flood attacks, so it would have a bad reputation score in the context of (D)DoS attacks, but its score in other contexts, *e. g.* port scanning or sending spam, might be good.

Information aging. If an IP address is reported to perform some kind of attack, it means it was probably part of a botnet or otherwise controlled by an attacker at the moment of the attack. It is likely that it holds true a minute, an hour or even several days later. However, with increasing time from the last report the probability the address is still malicious is getting lower and we can hardly infer anything from data older than a month. The host behind the

address might get fixed or the address might be assigned to another host in the meantime. This adds a degree of uncertainty to all the data which changes over time. Moreover, it is hard to specify how exactly does it change, since it may depend on various aspects, *e. g.* whether the address is statically or dynamically assigned (more on this later).

Information uncertainty. Many of the pieces of information entering the reputation scoring process may be imprecise, unreliable or otherwise uncertain. They may be, for example, deduced using some heuristic or approximate algorithm, or they may be obtained from an external data source we do not fully trust. Since most of that uncertainty can be described using probability, its incorporation into the reputation scoring process should not a problem in principle, but it will further increase its complexity.

IP to host mapping. The main purpose of having a reputation database is to gain knowledge about malicious *hosts*, but we work with IP addresses instead. And mapping of hosts to IP addresses is far from one-to-one due to dynamic address assignment and extensive use of NAT. However, tracking of individual hosts is practically impossible, especially with alert data only (and even if it would be possible, it would be probably considered very privacy intrusive). Therefore, we will at least try to recognize dynamic address ranges and NATs and adjust the scoring method for them. For example, information about dynamically assigned addresses should expire faster than that about static ones. We will draw from many existing works on this topic, *e. g.* [7, 11, 12].

3 Preliminary results

We started with analysis of alert data from CESNET's alert sharing system Warden [5]. It currently receives data from 16 detectors, mostly in CESNET2 network. We took two month-long datasets gathered in 2015, containing over 70 million alerts in total, and analyzed them from various points of view.

For example, we confirmed that sources of malicious traffic are geographically distributed non-uniformly, but we found that this distribution is very different for different types of malicious traffic. Also, although most IP addresses were reported only once, there were some addresses that were reported repeatedly for a long time. And for example, while only 8.5% of scanning addresses were reported in 5 or more days of a month, they were responsible for 65% of all port scanning events reported in that month. Thus, even knowing only the most active attackers might be very useful. For more information we refer the reader to our technical report [1] which presents more results of the analysis.

Currently we are experimenting with various machine learning methods trying to create the best predictor of network attacks. Also, we continue to gather data from more and more sources, since to successfully capture the global threat landscape by the reputation database, we need to know about a significant portion of all attackers on the Internet. In order to achieve this, we are involved in one national and two international projects about alert sharing.

Acknowledgments This research was supported by Security Research grant no. VI20162019029 in project Shaing and analysis of security events in Czech republic granted by Ministry of the Interior of the Czech republic. It was also partially supported from IT4Innovations excellence in science project (IT4I XS – LQ1602) and by Brno University of Technology grant no. FIT-S-14-2297.

References

1. Bartoš, V.: Analysis of alerts reported to Warden. Tech. Rep. 1/2016, CESNET (Feb 2016)
2. Bartoš, V., Žádník, M.: An Analysis of Correlations of Intrusion Alerts in an NREN. In: 19th International Workshop on Computer-Aided Modeling Analysis and Design of Communication Links and Networks (CAMAD). pp. 305–309. IEEE (Dec 2014)
3. C. A. Shue *et. al.*: Abnormally Malicious Autonomous Systems and Their Internet Connectivity. IEEE/ACM Transaction on Networking 20(1), 220–230 (Feb 2012)
4. Cambridge English Dictionary: reputation, `http://dictionary.cambridge.org/dictionary/english/reputation` (accessed on January 14, 2016)
5. CESNET: Warden – alert sharing system, `https://wardenw.cesnet.cz/`
6. ENISA: Standards and tools for exchange and processing of actionable information (Nov 2014)
7. Gokcen, Y., Foroushani, V., Heywood, A.: Can We Identify NAT Behavior by Analyzing Traffic Flows? In: Security and Privacy Workshops (SPW). pp. 132–139. IEEE (May 2014)
8. Merriam-Webster Dictionary: reputation, `http://www.merriam-webster.com/dictionary/reputation` (accessed on January 14, 2016)
9. Moreira Moura, G.C., Sadre, R., Pras, A.: Internet bad neighborhoods temporal behavior. In: Network Operations and Management Symposium (NOMS). pp. 1–9. IEEE (May 2014)
10. Moreira Moura, G.C.: Internet bad neighborhoods. Ph.D. thesis, University of Twente, Enschede (March 2013), `http://doc.utwente.nl/84507/`
11. Moreira Moura, G.C., *et. al.*: How Dynamic is the ISPs Address Space? Towards Internet-Wide DHCP Churn Estimation. In: 14th International Conference on Networking. IFIP (May 2015)
12. Vu, L., Turaga, D., Parthasarathy, S.: Impact of DHCP Churn on Network Characterization. SIGMETRICS Perform. Eval. Rev. 42(1), 587–588 (Jun 2014)