An Uncertain Graph-based Approach for Cyber-security Risk Assessment

Anonymous

ABSTRACT

We proposed a novel risk assessment approach for quantifying the security risk of lateral movement attacks, in which the *attack propagation* is modeled as an uncertain graph and the *attack impact* is a function of the set of compromised devices. We discussed several risk-based security metrics, including the expected loss, survival function, and conditional expectation – the last two measure the low-probability but high-impact events in the right tail of the loss distribution. The model is illustrated with a simple example and several directions for further research are also discussed.

1 INTRODUCTION

Analysis of the risk to cyber-networks of an intruder gaining access, moving laterally, and causing loss (e.g. physical damage or financial loss due to exfiltration of data) is a challenging problem for a number of reasons. One of these is that we cannot know with any certainty of an intruder's ability to find means of moving laterally through the network. That uncertainty can be due to lack of knowledge about the existence of exploitable vulnerabilities, the uncertainty may be due to lack of knowledge of the attacker's capabilities. The addition of security controls such as firewalls can change the underlying network structure. For example a single new firewall rule could block one host from accessing a vulnerable service on another host, changing an uncertain ability to use that connection to move laterally to a certain inability to move laterally. Important applications of risk models include assessing the degree to which one set of controls reduces loss, or the sensitivity of that loss to hardened protections of particular hosts. Even if the model parameters and effective losses are not precisely known, risk models can provide some insight in recognizing when a system is at high risk for loss, and the points in the system where protection or lack of it have the greatest impact on loss.

2 MODELING APPROACH

2.1 Attack propagation

A cyber system is modeled as an *uncertain graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$ where $\mathcal{V} = \{V_1, \ldots, V_n\}$ is the set of vertices, $\mathcal{E} = \{E_1, \ldots, E_m\}$ the set of directed edges, and $p = (p_1, \ldots, p_m)$ where p_i is the probability that edge E_i exists (e.g., see Figure 1.) The vertices of the graph represent networked devices; a directed edge from a to b with weight $p > 0$ means the network's networking and access control infrastructure may allow an intruder resident on a to reach b, exploit some vulnerability and occupy b as well. The vector p encodes the probability that the vulnerability actually exists and that the attacker is able to exploit it. A special vertex $s \in \mathcal{V}$ represents the starting point of the attack, also known as the "initial point of intrusion." The attacker's goal is to propagate from s to devices in the targeted system and to use those already compromised devices to launch a cyber-attack, causing a certain amount of damage and/or financial loss to the targeted system.

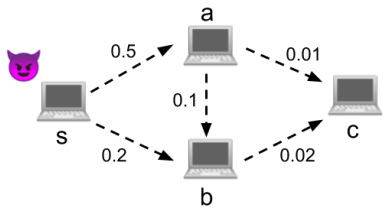

Figure 1: Attack propagation model using uncertain graph.

2.2 Attack impact

The model as stated captures the lateral movement and its inherent complexity but does not express the attack's impact. In some cases, the attack impact depends principally on the functionalities and security ramifications of all the devices that have already been compromised, denoted as $\mathcal{V}_c \subseteq \mathcal{V}$. Moreover, with a static defense one can assume that the attack impact is a non-decreasing function of $\mathcal{V}_c$ – as the attacker compromises more and more devices, the damage they can inflict on the targeted system remains the same if not increases. (A counter-example might be an active defense which is able to disconnect an attacker from previously acquired hosts.) We use $L(V)$ to denote the loss, e.g. financial loss, suffered by host V by being compromised, independent of any other hosts that may be compromised. $L(V)$ might reflect only compromised data at V, it might reflect that V gives the attacker access to a particular critical network. Abusing notation slightly, we apply function L to the set of compromised hosts $\mathcal{V}_c$ as $L(\mathcal{V}_c)$, and use this to give the total loss of all hosts in $\mathcal{V}_c$ being compromised. If V_1 and V_2 both reach the same network we might have $L(V_1) = L(V_2)$, but also $L(\{V_1, V_2\}) = L(V_1) = L(V_2)$ if there is no added value to compromising V_2 if V_1 is already compromised. This observation leads to natural definitions of L being *additive*, i.e. $L(\mathcal{V}_c) = \sum_{V_i \in \mathcal{V}_c} L(V_i)$, *sub-additive*, i.e. $L(\mathcal{V}_{c_1} \cup \mathcal{V}_{c_2}) \leq L(\mathcal{V}_{c_1}) + L(\mathcal{V}_{c_2})$ for all disjoint $\mathcal{V}_{c_1}, \mathcal{V}_{c_2} \subseteq \mathcal{V}_c$, and *super-additive*, i.e. $L(\mathcal{V}_{c_1} \cup \mathcal{V}_{c_2}) \geq L(\mathcal{V}_{c_1}) + L(\mathcal{V}_{c_2})$ for all disjoint $\mathcal{V}_{c_1}, \mathcal{V}_{c_2} \subseteq \mathcal{V}_c$. In general, additive loss functions are natural when the losses are limited to hosts and the data on them.

3 RISK ASSESSMENT

3.1 Risk concepts

With that we can now define the three fundamental concepts in risk [2], the realization (or scenario), the probability (or likelihood), and the impact.

3.1.1 Realization. Let $X = (X_1, \ldots, X_m)$ be the multivariate Bernoulli random variable where $X_i \in \{0, 1\}$ indicates the random event that edge E_i exists. A realization is simply defined as an element $x \in \{0, 1\}^m \stackrel{\text{def.}}{=} \mathbb{X}$ in the set of all possible outcomes of X.

3.1.2 Probability. Assuming that X_i's are mutually independent, for each realization $x \in \mathbb{X}$, its *probability* can be computed as

$$\mathbb{P}(X = x) \equiv \mathbb{P}(X_1 = x_1, \ldots, X_m = x_m) = \prod_{i=1}^{m} (x_i p_i + (1 - x_i)(1 - p_i)).$$

Following the ideas of [4] we believe it is possible to extend the model to one where edge probabilities may be correlated through couple of stochastically independent Boolean random variables, provided that the Boolean expressions on all edges are monotone.

3.1.3 Impact. Given a realization $x \in \mathbb{X}$, we denote $G(x) \stackrel{\text{def.}}{=} (\mathcal{V}, \mathcal{E}(x))$ to be the deterministic graph realized from $\mathcal{G}$ where $\mathcal{E}(x) \stackrel{\text{def.}}{=} \{E_i \in E : x_i = 1\}$. Furthermore, we define $\mathcal{V}(x) \subseteq \mathcal{V}$ to be the set containing all vertices in $G(x)$ that can be reached from s; in other words, $\mathcal{V}(x)$ is the set of compromised devices. The impact of the attack under x, which is denoted as $L(x)$, can be defined as $L(x) \stackrel{\text{def.}}{=} \sum_{V_i \in \mathcal{V}(x)} L(V_i)$. By definition, $L(X)$ is a function of X and therefore is also a random variable. In the risk community, the distribution of $L(X)$ is also known as the *loss distribution* and $L(x)$ is the loss under realization x for $x \in \mathbb{X}$.

3.2 Risk measures

3.2.1 Expected loss. The most intuitive measure that can be used to quantity $L(X)$ is arguably the *expected loss*, defined as

$$\mathbb{E}(L(X)) = \sum_{x \in \mathbb{X}} L(x)\mathbb{P}(X = x) \tag{1}$$

where $\mathbb{E}$ denotes the expectation of functions of X when X is distributed according to $\mathbb{P}$. A common critic of this measure is that it gives little insight about extreme events that show up in the tail of the distribution. Those low-probability high-impact events can be of significant interest to the risk analyst (e.g. most people are not risk-neutral but rather risk-averse to large consequence events [3].) Two risk measures related to extreme events are given below.

3.2.2 Survival function. This risk measure computes the probability that the loss is greater than some given threshold C. Let $\mathcal{X} \in \mathbb{X}$ be the event containing all realizations under which the loss is greater than C, i.e. $\mathcal{X} \stackrel{\text{def.}}{=} \{x \in \mathbb{X} \mid L(x) > C\}$. The probability of $\mathcal{X}$ is the sum of the probabilities of all realizations under $\mathcal{X}$, which can be expressed in a few different ways

$$\mathbb{P}(X \in \mathcal{X}) = \sum_{x \in \mathbb{X}} \mathbf{1}_{\mathcal{X}}(x)\mathbb{P}(X = x) = \mathbb{E}(\mathbf{1}_{\mathcal{X}}(X)) \tag{2}$$

where $\mathbf{1}$ is the indicator function, i.e. $\mathbf{1}_{\mathcal{X}}(x)$ is equal to 1 if $x \in \mathcal{X}$ and 0 otherwise.

3.2.3 Conditional expectation. This risk measure computes the expected loss given that it is greater than some threshold C, formally,

$$\mathbb{E}(L(X) \mid X \in \mathcal{X}) = \frac{\sum_{x \in \mathcal{X}} L(x)\mathbb{P}(X = x)}{\mathbb{P}(X \in \mathcal{X})} = \frac{1}{\mu} \sum_{x \in \mathcal{X}} L(x)\mathbb{P}(X = x) \tag{3}$$

This definition is a generalization of the expected loss in Equation 1. Note that by letting $C = \text{VaR}_p(L(X))$ where $\text{VaR}_p(L(X))$ is the value-at-risk [1] of the loss distribution, the conditional expectation in Equation 3 is equivalent to another well-known risk measure called the *conditional tail expectation* [1]. In plain text, the conditional tail expectation measures the expected loss under the worst $(1 - p) \times 100\%$ of the cases, where p is usually close to 1.

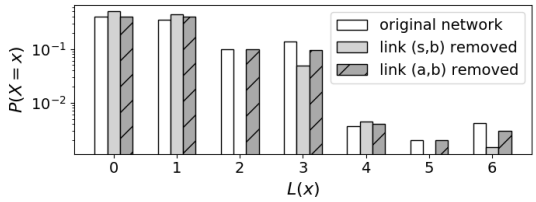

Figure 2: Loss distributions of the studied network without and with two candidate cyber defense solutions.

4 EXAMPLE AND FUTURE RESEARCH

To illustrate the use of the proposed model, we apply it to the network in Figure 1 with $\mathcal{V} = \{s, a, b, c\}$. Assuming that the loss function L is additive with $L(s) = 0$, $L(a) = 1$, $L(b) = 2$, and $L(c) = 3$, the model gives rise to a total of $2^5 = 32$ different realizations with the possible losses ranging between $[0, 6]$ (Figure 2) and an expected loss of 1.009. Using a loss threshold of $C = 3$, the values of the survival function and the conditional expectation are computed as 0.010 and 5.059, respectively. Suppose the network administrator decides to harden the system by implementing one of two candidate cyber defense solutions, one would result in the removal of link (s, b) and the other link (a, b). In either case, the attacker can still perform lateral movement from s to any other host in $\mathcal{V}$. However, using the same computation, the values of the expected loss, survival function, and conditional expectation now become 0.618, 0.006, and 4.497 in the first case (i.e. where link (s, b) is removed), compared with 0.927, 0.009, and 4.886 in the second case. These numbers imply that the first defense solution results in better risk reduction and therefore should be selected.

One may use this model as the starting point to explore several research directions. One direction is to study efficient methods for estimating the proposed risk measures (recall that computing the expected loss, survival function, or conditional expectation can be reduced to the S-T CONNECTEDNESS problem [5], which was proven #P-complete.) Another research direction is to see how the model can be extended to express, including but not limited to, (i) a richer class of the loss functions, particularly those that are non-monotone or monotone but non-additive, (ii) correlations between edge existences (e.g. see method in [4]), (iii) the attacker's preference and decision making process (e.g. repeating successfully tested exploits rather than trying out new ones), and (iv) dynamic interactions between the attacker, the defender, and the targeted system. Lastly, obtaining meaningful and reliable quantitative data that can be used in this model remains at large an open problem.

REFERENCES

[1] Denuit, M., Dhaene, J., Goovaerts, M., and Kaas, R. *Actuarial Theory for Dependent Risks: Measures, Orders and Models.* Wiley, 2005.

[2] Kaplan, S., and Garrick, B. J. On the quantitative definition of risk. *Risk Analysis 1* (1981).

[3] Modarres, M. *Risk Analysis in Engineering: Techniques, Tools, and Trends.* CRC Press, 2006.

[4] Nguyen, H. H., Palani, K., and Nicol, D. M. An approach to incorporating uncertainty in network security analysis. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp* (New York, NY, USA, 2017), HoTSoS, ACM, pp. 74–84.

[5] Valiant, L. G. The Complexity of Enumeration and Reliability Problems. *SIAM Journal on Computing 8*, 3 (1979), 410–421.