Game Theoretical Adaptation Model for Intrusion Detection System[*]

(Extended Abstract)

Martin Rehak[†‡], Michal Pechoucek[†‡], Martin Grill[†], Jan Stiborek[†], Karel Bartos[†]
† Department of Cybernetics, Czech Technical University in Prague, Czech Republic
‡ Cognitive Security s.r.o., Prague, Czech Republic
martin.rehak@agents.felk.cvut.cz

ABSTRACT

We present a self-adaptation mechanism for Network Intrusion Detection System which uses a game-theoretical mechanism to increase system robustness against targeted attacks on IDS adaptation. We model the adaptation process as a strategy selection in sequence of single stage, two player games. The key innovation of our approach is a secure runtime game definition and numerical solution and real-time use of game solutions for dynamic system reconfiguration. Our approach is suited for realistic environments where we typically lack any ground truth information regarding traffic legitimacy/maliciousness and where the significant portion of system inputs may be shaped by the attacker in order to render the system ineffective. Therefore, we rely on the concept of challenge insertion: we inject a small sample of simulated attacks into the unknown traffic and use the system response to these attacks to define the game structure and utility functions. This approach is also advantageous from the security perspective, as the manipulation of the adaptive process by the attacker is far more difficult. Our experimental results suggest that the use of game-theoretical mechanism comes with little or no penalty when compared to traditional self-adaptation methods.

Categories and Subject Descriptors

C.2.0 [**COMPUTER-COMMUNICATION NETWORKS**]: General—*Security and protection*

General Terms

Algorithms, Security

Keywords

adaptation, game theory, security, intrusion detection

1. INTRODUCTION

In this paper, we use the game-theoretical models to improve the security of the adaptation process within a distributed, agent-based Intrusion Detection System (IDS). The high-level self-adaptation method that we develop our approach on [2] has been designed for the intrusion detection systems based on the anomaly detection paradigm: these systems observe the past behavior of the monitored network/hosts, predict their future behavior using statistical and other models and identify the behavior diverging from the prediction as anomalous. Adaptation, self-management and self-optimization techniques that are used inside an IDS can significantly improve their performance [2] (i.e. reduce the number of false alarms) in a highly dynamic environment, but are also a potential target for an informed and sophisticated attacker. When the adaptation techniques are deployed improperly, they can alow the attacker to reduce the system performance against one or more critical attacks. This paper presents a game theoretical model of adaptation processes inside an autonomic, self-optimizing IDS, presents an architecture integrating the process with an existing IDS.

We present an **architecture** that integrates the abstract game model into an IDS with self-monitoring capability, in order to simulate the worst case, optimally informed attacker and to optimize the system behavior against such attacker. Such (hypothetical) attacker with full access to system parameters could dynamically identify the best strategy to play against the system. Optimizing the detection performance against the worst case attacker protects the system from more realistic attacks based on long-term probing and adversarial machine learning approaches referenced above.

2. GAME MODEL

We conceptualize the relationship between the attacker and the defender as a *sequence* of *single stage, two player, non-zero sum games*, where the attack/defence actions of both players correspond to strategies in the game-theoretical model of their interaction and the environment evolves between the game. The game model (and utility functions in particular) are based on [1], with additional inputs from the network administrators and actual IDS users. The game model integrates the preferences and strategies of two players (attacker and defender). Their strategy sets are defined as a selection of IDS configurations for the defender and the selection of a particular attack type (e.g. buffer overflow, password brute-force, scan...) for the attacker. The main difference of the utility functions from [1] is the relaxation of the requirement on the identical attacker gain/defender loss and the proportionality of associated costs (alarm processing, monitoring etc.) with the gain/loss value. This requirement was considered as too strong by the system administrators we have questioned.

[*] This material is based upon work supported by the ITC-A of the US Army under Contract No. W911NF-10-1-0070. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the ITC-A of the US Army. Effort is also sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant number FA8655-10-1-3016. Also supported by Czech Ministry of Education grants 6840770038 and ME10051.

Cite as: Game Theoretical Adaptation Model for Intrusion Detection System (Extended Abstract), M. Rehak, M. Pechoucek, M. Grill, et al., *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems – Innovative Applications Track (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 1123-1124.

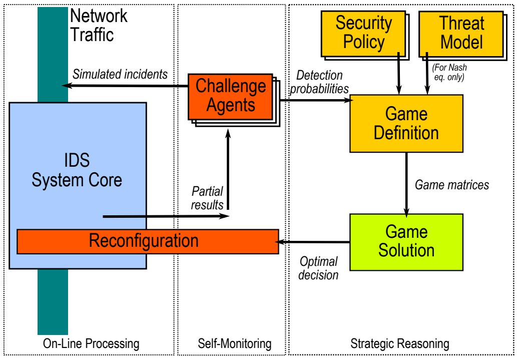

Figure 1: Indirect online variant of game/IDS integration.

The actual utility function values of both players depend principally on the sensitivity of the system using defender's strategies with respect to individual attacker's strategies ($\alpha_{i,j}$) , and the associated rate of false positives (β_i) for each configuration. $\alpha_{i,j}$ denotes the probability that the j-th attack strategy is detected by the IDS when the defender plays the i-th defence strategy and β_i denotes the probability that the i-th defender's strategy will result in a false alert. These parameters shape the utility functions of both players in each game stage. By our experience, these values wary widely with changing characteristics of the background traffic and need to be estimated dynamically for each given game in a sequence, as we will present below.

The gameplay is very simple in our case: both players simultaneously select their strategies from the set S and the combination of these strategies determines the payoffs to attacker and defender, as defined by their respective utility functions. The solution concepts used to solve/analyze the game are **Max-Min** and **Nash** equilibria. We play a sequence of games described above, each corresponding to one time interval. The individual games in the sequence are differentiated by the dynamically evolving parameters of player's utility functions. We consider the individual games to be independent and we don't carry over any information between them.

3. ARCHITECTURE

There are two existing approaches to integration of the game model with an IDS:

◇ *Off-line integration*, when the game is defined in design time, solved analytically, using *a priori* knowledge about expected impacts and success likelihood of the attacks, and the system parameters are fixed to resulting strategies according to game results. Game theory use ensures that the system parameters are set to force the adversary into the selection of less damaging (or more rational) strategies. It is sufficient for systems deployed in stable environments, but most IDS need to cope with dynamic environments, where the background traffic an other factors change frequently. In such environments, the static strategies perform poorly.

◇ *Direct on-line integration*, when the game uses presumed adversary actions in the observed network traffic to define the game is the opposite approach. The game is being defined by the actual actions of real-world attackers executed against the monitored system, elegantly solving the relevance problem. On the other hand, direct interaction between the adversary and the adaptation mechanism makes the system potentially vulnerable to attacks against the adaptation algorithms, creating a new attack surface. Motivated attacker can easily mislead the IDS by insertion of a sequence of attacks that are orthogonal to its actual plan to target its utility.

Our approach, named *indirect online integration* combines the above approaches and provides interesting security properties desirable for real-world deployment. The solution uses the concept of challenges [2] to mix a controlled sample of legitimate and adversarial behavior with actually observed network traffic and is a compromise between the above approaches (see Fig. 1). In this case, the real traffic background (including any possible attacks) is processed in conjunction with simulated hypothetical attacks within the system. We measure the system response to these challenges, drawn from the realistic attack classes, and use them to estimate the system response to the real-world samples from the same classes. In practice, we will define one class for each broadly defined attack/legitimate traffic type and measure the difference between the system response to legitimate traffic and to various classes of malicious traffic. The challenges are then mixed with the real traffic on IDS input and the system response to them is used as an input for game definition, measuring/estimating the current values of: $\alpha_{i,j}$ and β_i. The major advantage is higher robustness w.r.t strategic attacks on adaptation algorithms, and lower system configuration predictability by the adversary, as the simulation runs inside the system itself and its results can not be easily predicted by the attacker.

This approach offers the optimal mix of situation awareness and security against engineered inputs. In this case, we actually play against an abstract opponent model inside the system, and expect that the moves that are effective against this opponent will be as effective against the real attacks. The advantage of this approach is not only in its security, but also in better model characteristics in terms of strategy space coverage (unfrequent, but critical attacks are covered), robustness and relevance – the abstract game can represent the attacks and utility combinations that would be obvious only for insider attackers.

4. CONCLUSIONS

The experiments we have performed with a simplified (and modified) version of commercially available IDS solution clearly showed that the game theoretical models/solvers integrated into an adaptive IDS provide the results more than equivalent to the alternative direct optimization methods, as we have verified on inserted challenges and real-world attacks performed on the monitored network. These methods provide robust performance and reliably converge when using both max-Min or Nash equilibria. The additional benefits, such as increased robustness against an attacker with insider access, therefore build a strong case for their use by the industry. In particular, our results suggest that the max-min solution concept provides very consistent results, does not require an explicit model of opponent's utility function and is computationally trivial, making it an interesting first choice for future proof-of-concept implementations.

5. REFERENCES

[1] L. Chen and J. Leneutre. A game theoretical framework on intrusion detection in heterogeneous networks. *Information Forensics and Security, IEEE Transactions on*, 4(2):165–178, June 2009.

[2] M. Rehak, E. Staab, M. Pechoucek, J. Stiborek, M. Grill, and K. Bartos. Dynamic information source selection for intrusion detection systems. In K. S. Decker, J. S. Sichman, C. Sierra, and C. Castelfranchi, editors, *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '09)*, pages 1009–1016. IFAAMAS, May 2009.