A Framework for Evaluating Deployed Security Systems: Is There a Chink in your ARMOR?

Matthew E. Taylor, Christopher Kiekintveld, Craig Western, and Milind Tambe
http://teamcore.usc.edu
Computer Science Department
The University of Southern California

Keywords: Physical Security, Game Theory, Stackelberg Games, Evaluation

Received: September 7, 2009

A growing number of security applications are being developed and deployed to explicitly reduce risk from adversaries' actions. However, there are many challenges when attempting to evaluate *such systems, both in the lab and in the real world. Traditional evaluations used by computer scientists, such as runtime analysis and optimality proofs, may be largely irrelevant. The primary contribution of this paper is to provide a preliminary framework which can guide the evaluation of such systems and to apply the framework to the evaluation of ARMOR (a system deployed at LAX since August 2007). This framework helps to determine what evaluations could, and should, be run in order to measure a system's overall utility. A secondary contribution of this paper is to help familiarize our community with some of the difficulties inherent in evaluating deployed applications, focusing on those in security domains.*

1 Introduction

Computer scientists possess many tools that are particularly applicable to security-related problems, including game-theoretic reasoning, efficient algorithmic design, and machine learning. However, there are many challenges when attempting to *evaluate* a security system in a lab setting or after it has been deployed. Traditional evaluations used by computer scientists — such as runtime analysis and optimality proofs — often do not consider the relevance of modeling assumptions or account for how a system is actually used by humans. If there is an error "between the keyboard and the chair," it still needs to be addressed, even if such problems are beyond the scope of some computer programs. An additional complication is that no security system is able to provide 100% protection. Instead, systems must be evaluated on basis of risk reduction, often through indirect measures such as increasing adversary cost and uncertainty, or reducing the effectiveness of an adversaries' attack. Despite these challenges, evaluation remains a critical element of the development and deployment of any security system.

An important challenge in security evaluation is that performance necessarily depends on an adversarial human's behavior and decisions. Controlled laboratory studies can be a valuable component of an evaluation, but the population of test subjects is necessarily different that that of actual attackers. Evaluating a system once it is deployed only increases the experimenter's burden. First, while a system could be alternatively enabled and disabled on different days to measure its efficacy, this is at best impractical and at worst unethical. Second, data related to the configuration and performance of the system may be classified or sensitive, and not available to researches. Third, a key component of many security systems is *deterrence*: an effective system will not only identify and prevent successful attacks, but will also dissuade potential attackers. Unfortunately, it is generally impossible to directly measure the deterrence effect.[1]

This paper introduces a general framework for evaluating deployed systems and then presents a case study of one such security system. While computer scientists traditionally prioritize precise, repeatable studies, this is not always possible in the security community; computer scientists are used to quantitative evaluations in controlled studies, whereas security

[1] To measure deterrence, one needs to know how many attacks *did not occur* due to security, a generally unmeasurable counterfactual.

specialists are more accepting of qualitative metrics on deployed systems. For instance, Lazaric [1999] summarized a multi-year airport security initiative by the FAA where the highest ranked evaluation methodology (of seven) relied on averaging *qualitative* expert evaluations.

The primary advantage of quantitative evaluations, rather than qualitative, is that they can be integrated into a cost-benefit analysis. A particular security measure may be effective but prohibitively expensive — consider two extremes in the domain of airport security. Hand searching every passenger who enters an airport and disallowing all luggage would likely increase the security of plane flights, but the costs from extra security personnel, increased time in the airport, and lost revenue makes such a draconian policy infeasible. On the other hand, removing all airport screenings and restrictions would reduce costs and delays, but also significantly increase security risks. By carefully weighing costs and benefits, including non-monetary effects like privacy loss, security experts and policy makers can better decide which measures are appropriate in a particular context.

Our ultimate goal is to provide a framework for comprehensive evaluation of deployed systems along multiple attributes, in absolute or relative terms, to facilitate cost-benefit analysis. We examine existing evaluations of the ARMOR system [Pita *et al.*, 2008] as a case study. Several different kinds of evaluation of this system support the claim that it significantly improves over the previous best practices of uniform randomization or hand-constructed schedules and is cost effective.

The primary contribution of this paper is to provide a framework to evaluate such deployed systems and apply it to ARMOR. This framework helps to determine what to measure, how to measure it, and how such metrics can determine the system's overall utility. A secondary contribution of this paper is to help familiarize our community with some of the difficulties inherent in evaluating deployed applications, particularly for security domains.

2 Case Study: ARMOR

The Los Angeles World Airports (LAWA) police at the Los Angeles International Airport (LAX) operate security for the fifth busiest airport in the United States (and largest destination), serving 70–80 million passengers per year. LAX is considered a primary terrorist target on the West Coast and multiple individuals have been arrested for plotting or attempting to attack LAX [Stevens *et al.*, 2009]. Police have designed multiple rings of protection for LAX, including vehicular checkpoints, police patrols of roads and inside terminals (some with bomb-sniffing canine units, also known as K9 units), passenger screening, and baggage screening.

Figure 1: A LAX checkpoint scheduled by ARMOR

Figure 2: A K9 patrol

There are not enough resources (police and K9 units) to monitor every event at the airport due to the large physical area and the number of passengers served. ARMOR addresses two specific security problems by increasing the unpredictability of security schedules and weighting defensive strategy based on targets' importance. First, there are many roads that are entry points to LAX. When and where should vehicle checkpoints (Figure 1) be set up on these roads? Pertinent information includes typical traffic patterns on inbound roads, the areas each road accesses within LAX, and areas of LAX which may have more or less importance as terrorist targets. Second, how and when should the K9 units (Figure 2) patrol the eight terminals at LAX? Here it is important to consider the time-dependant passenger volumes per terminal, as well as the attractiveness of different terminals. In both cases a predictable pattern can be exploited by an observant attacker.

To address the two security problems above, we use game theory to model and analyze the two domains. The police and attackers play a Bayesian Stackelberg game [Conitzer and Sandholm, 2006], with the police first committing to a (randomized) security policy. Multiple attacher types are modeled. Each attacker type observes this policy and then selects the optimal attack strategy (depending on the defense strategy). Solving this game for a Strong Stackelberg Equilibrium finds an optimal randomized policy for the police, which is sampled as necessary to give specific schedules. ARMOR (*Assistant for Randomized Monitoring Over Routes*) is the software tool that assists police with randomized scheduling using this game-theoretic analysis [Pita *et al.*, 2008]. The software uses an optimized algorithm for solving Bayesian Stackelberg games called DOBSS [Paruchuri *et al.*, 2008].

The randomized schedules account for three key factors: (1) attackers are able to observe the security policy using surveillance, (2) attackers change their behavior in response to the security policy, and (3) the risk/consequence of an attack varies depending on the target. The end result is a randomized police schedule that is unpredictable, but weighted towards high-valued targets. ARMOR has been in use at LAX since August 2007, marking an important transition from theoretical to practical application. The system has received very positive feedback and is considered an important element of security at the airport.

3 Current ARMOR Evaluations

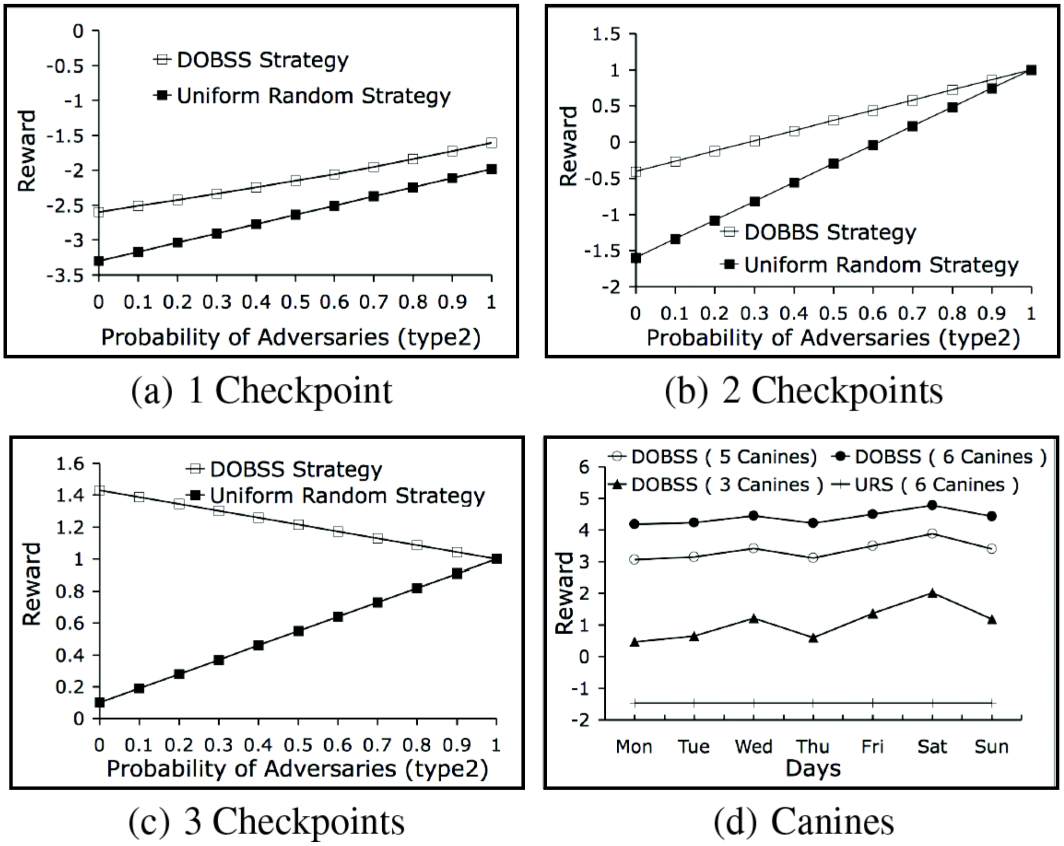
The ARMOR system has undergone multiple evaluations before and after deployment. We summarize the current evaluations below, both from existing publications and novel to this article, grouped by category. It is not difficult to argue that ARMOR is a significant improvement over previous practices: it saves time for human schedulers, it is inexpensive to implement, and humans are known to have difficulty randomizing effectively [Wagenaar, 1972]. However, our goal is to take steps towards a more comprehensive understanding of ARMOR that provides as much insight as possible into the value of the system.

3.1 Mathematical

The first category of analyses are mathematical evaluations that use our game-theoretic model to evaluate ARMOR's security policies against other baseline policies. In particular, if we assume attackers act optimally and have the utilities specified in the model, we can predict how they will react to any schedule and therefore compare the expected utility of these schedules. ARMOR uses a game theoretic optimal schedule. Comparing against benchmark uniform random and hand-crafted schedules show that ARMOR's schedule is substantially better than these benchmarks across a variety of different settings. For example, Figure 3(d) shows the expected reward for the police using ARMOR's schedule (calculated using DOBSS) compared with a uniform random benchmark strategy in the canines domain. ARMOR is able to make such effective use of resources that using three canines scheduled with DOBSS yields higher utility than using six canines with uniform random scheduling!

Sensitivity analysis is another important class of evaluations that can be performed using only the mathematical models. In this type of evaluation, important parameters of the model are varied to test how sensitive the output of the model is to the input. One important input to our models is the distribution of different types of attackers. For example, some attackers may be highly organized and motivated, while others are amateurish and more likely to surrender. Different types of attackers can be modeled as having different payoff matrices. Changing the percentages of each attacker can help show the system's sensitivity to assumptions regarding the composition of likely attackers, and (indirectly) the system's dependence on precise utility elicitation. In Figure 3(a)–3(c), there are two adversary types with different reward matrices. Figure 3(a) demonstrates that DOBSS has a higher expected utility than that of a uniform random strategy on a single checkpoint, regardless of the percentage of "type one" and "type two" adversaries. Figures 3(b) and (c) shows that DOBSS again dominates uniform random for two and three checkpoints, respectively.

Further sensitivity analysis can be applied to measure how the optimal strategy computed by DOBSS changes as payoffs are modified. Since the payoff functions are determined through preference elicitation sessions with experts, these payoffs are estimates of true utilities. Game-theoretic models can be quite sensitive to payoff noise, and arbitrary changes in the payoffs can lead to arbitrary changes in the optimal



(a) 1 Checkpoint

(b) 2 Checkpoints

(c) 3 Checkpoints

(d) Canines

Figure 3: Comparisons of ARMOR's schedules with a uniform random baseline schedule. Figures a–c show the utility of schedules for 1–3 vehicle checkpoints varying the relative probability of two different attacker types. The x-axes show the probability of the two attacker types (where 0 corresponds to 0% attack type 2, and 100% attack type 1) and y-axes show the expected utility of ARMOR (using the DOBBS solver) and a uniform random defense strategy. Figure d shows that DOBSS can outperform the baseline, even using many fewer K9 units. The x-axis shows the results from seven different days, and the y-axis shows the expected utility for the different scheduling methods.

schedule. However, there is some evidence that ARMOR is robust to certain types of variations. In one experiment, we multiplied all of the defender's negative payoffs for successful attacks by a factor of four, essentially increasing the impact of a successful attack. We found that in the one and three checkpoint case, the strategies were unchanged. In the two checkpoint case the actions were slightly different, but the overall strategy and utility were unchanged.

As with any game theoretic analysis, the assumptions regarding the opponent's behavior may dramatically change the outputs and evaluated performance. Figure 4 examines an assumption typically made by Stackelberg solvers. Specifically, such solvers assume that if an adversary is given a set of actions with equivalent payoffs, the attacker will select the action that maximizes the defender's payoff (the Strong Stackelberg Equilibrium, or SSE). We compare this potentially optimistic behavior with two other reasonable choices: the attacker selects randomly from the set of actions with the maximum (equivalent) attacker utility, and the attacker selects the action that minimizes the defender utility from the set of equivalent actions with the maximum attacker utility. The similarity in payoffs of these three ways for attackers to break ties show that this assumption is not critical for ARMOR's success.

Additionally, note that Figure 4 has a roughly linear trend. Resource graphs that have a "knee," or location where the marginal utility improvement sharply decreases, suggest a natural resource allocation. In the case of a linear utility curve, adding an extra resource will return the same marginal expected utility. One benefit of such an analysis is that in budget meetings, security experts can show the expected impact to safety as the budget changes.

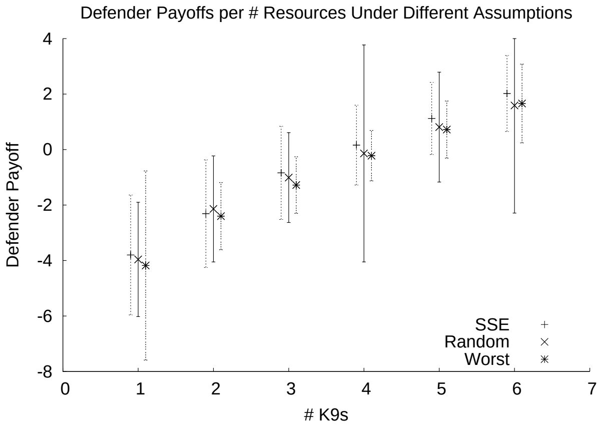


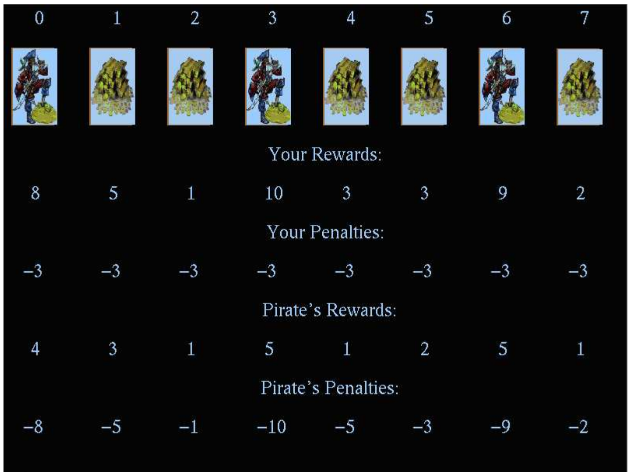
Figure 4: This graph shows game theoretic evaluations of the K9 scheduling program for different numbers of resources, each averaged over 161 trials (error bars show standard errors). First, the SSE assumption is reasonable, as two different defender action selection mechanisms yield little change to the defender payoff. Second, note that the defender utility of additional resources appears approximately linear.

Lastly we also note that significant work has gone into speeding up the Bayesian Stackelberg solver. While detailed timing analysis often features prominently in computer science papers, for the purposes of evaluating ARMOR it is sufficient that the system runs "quickly enough" to meet the needs of the LAWA police on the size of problem instances they face on a daily basis. Other speedup techniques may be necessary in much larger domains, such as when scheduling over hundreds of thousands of different targets [Kiekintveld *et al.*, 2009].

3.2 Human Behavioral Experiments

ARMOR's game-theoretic model uses strong assumptions about the attacker's rationality to predict how they will behave and optimize accordingly. Humans often do not always conform to the predictions of strict equilibrium models (though some other models offer better predictions of behavior [Erev *et al.*, 2002]). In addition, ARMOR assumes that an attacker can perfectly observe the security policy, which may not be possible in reality.

We have run controlled laboratory experiments with human subjects to address both of these concerns [Pita *et al.*, 2009]. In these experiments, subjects play a "pirates and treasure" game designed to simulate an adversary planning an attack on an LAX terminal, shown in Figure 5. Subjects are given information about the payoffs for different actions and the pirates' strategy for defending their gold (analogous to the security policy for defending airport terminals). Subjects receive payments based on their performance in the game.



0	1	2	3	4	5	6	7
Your Rewards:							
8	5	1	10	3	3	9	2
Your Penalties:							
−3	−3	−3	−3	−3	−3	−3	−3
Pirate's Rewards:							
4	3	1	5	1	2	5	1
Pirate's Penalties:							
−8	−5	−1	−10	−5	−3	−9	−2

Figure 5: Screenshot of the "pirates and treasure" game

These experiments have provided additional support for quality of ARMOR's schedules against human opponents. First, they suggest that the assumptions imposed by the game-theoretic model are reasonable. Second, we have tested many conditions, varying both the payoff structure and the observation ability, ranging from no observation of the defense strategy to perfect observation. The results show that ARMOR's schedules achieve higher payoffs than the uniform random benchmark across all of the experimental conditions tested, often by a large margin.[2] These results demonstrate that ARMOR schedules outperform competing methods when humans are trying to defeat the defender.

3.3 Operational Record

A potentially useful test of ARMOR would be to compare the risk level at the airport with and without the system in place. This is problematic for several reasons that we discuss in more depth later on, including the sensitivity of the relevant data and the impossibility of controlling for many important variables. However, there is some public information that can be of

[2] New defense strategies developed in this work show even better performance against some (suboptimal) human adversaries by explicitly exploiting the attacker's weaknesses.

use in evaluating the performance of the system, including arrest records. There have been many success stories, prompting significant media coverage. For example, in the month of January this year, the following seven stops discovered one or more firearms, resulting in five arrests:

1. January 3, 2009: Loaded 9/mm pistol discovered
2. January 3, 2009: Loaded 9/mm handgun discovered (no arrest)
3. January 9, 2009: 16 handguns, 4 rifles, 1 pistol, and 1 assault rifle discovered — some loaded
4. January 10, 2009: 2 unloaded shotguns discovered (no arrest)
5. January 12, 2009: Loaded 22/cal rifle discovered
6. January 17, 2009: Loaded 9mm pistol discovered
7. January 22, 2009: Unloaded 9/mm pistol discovered (no arrest)

This data, while not conclusive, is encouraging. It appears that potential attackers are being caught at a high rate at ARMOR-scheduled checkpoints.

3.4 Qualitative Expert Evaluations

Security procedures at LAX are subject to numerous internal and external security reviews (not all of which are public). The available qualitative reviews indicate ARMOR is both effective and highly visible. Director James Butts of the LAWA police reported that ARMOR “makes travelers safer,” and Erroll Southers, Assistant Chief of LAWA police, told a Congressional hearing that “LAX is safer today than it was eighteen months ago,” due in part to ARMOR. A recent external study by Israeli transportation security experts concluded that ARMOR was a key component of the LAX defensive setup.

ARMOR was designed as a mixed initiative system that allows police to override the recommended policies. In practice, users have not chosen to modify the recommended schedules, suggesting that users are confident in the outputs. While such studies are not very useful for directly quantifying ARMOR’s benefit, it would be very hard to deploy the system without the support of such experts. Furthermore, if there were an “obvious” problem with the system, such experts would likely identify it quickly.

We have also compared ARMOR-enabled scheduling with previous LAWA practices [Cruz, 2009]. First, all checkpoints previously remained in place for an entire day, whereas checkpoints are now are moved throughout the day according to ARMOR’s schedule (adding to the adversary’s uncertainty). Second, before ARMOR only a single checkpoint was manned on any given day; multiple checkpoints are now used (due to an increased security budget). Third, a fixed sequence of checkpoints was defined (i.e., checkpoints 2, 3, 1, etc.), to create a static mapping from date to checkpoint. This sequence was not optimized according to the importance of different targets and the sequence would repeat (allowing the attacker to anticipate which checkpoint would be manned on any given day).

Expert opinions have said that an important benefit of the system is its transparency and visibility which contribute to deterrence. ARMOR assumes that adversaries are intelligent and have the ability to observe the security policy: knowing about the system does not reduce its effectiveness. The deployment of ARMOR has been quite visible: ARMOR has been covered on local TV stations (including FOX and NBC), in newspapers (including the LA Times and the International Herald Tribune), and in a national magazine (Newsweek).

4 Dimensions of Comparison

Evaluating deployed security systems poses many challenges and there is currently no “gold standard” that can be applied in all cases. Our general approach is based on cost-benefit analysis, with the goal of maximizing the utility of the deployed system. A key challenge in applying this methodology to security domains is that many costs and benefits are difficult, or even impossible, to measure directly. For this reason, it is important to carefully consider which metrics of costs and benefits are desirable, and what sources of data are available to estimate these metrics. We thus categorize representative tests in terms of the assumptions they make, relative accuracy, and the cost of running the test. We first discuss three general dimensions of evaluation (Section 4.1) and then a fourth security-specific dimension (Section 4.2). Each type of test has inherent limitations and it is important to draw on as many different categories as possible to provide a compelling validation of a deployed system.

4.1 Test Categories

Possible evaluations cover a broad spectrum of evaluation methods. At one end, mathematical analysis is relatively convenient, but requires strong and sometimes questionable assumptions [Lazarick, 1999; Bier, 2007]. At the other, situated tests using the actual personnel and equipment are very realistic, but also very costly and may not be able to directly measure desired variables. Along this spectrum, we group tests according to their type, their accuracy, and cost:

Test Types:

- Mathematical: Formal reasoning using a precise model
- Computational simulation: Computational simulations of varying degrees of abstraction/realism
- Controlled laboratory studies: Testing systems with human subjects can account for human decision making, which may be suboptimal or irrational
- Natural experiments: Observe the behavior the the deployed system by gathering data without intervention
- Situated studies: Testing a deployed system provides the most realistic data, but at high cost
- Qualitative expert studies: Domain experts can examine a system and give a holistic evaluation

Accuracy: Different categories of evaluation offer different tradeoffs in the realism of their assumptions, as well as the precision and repeatability of the results. A mathematical model is typically precise, but dependent on modeling assumptions. On the other hand, real-world tests make fewer assumptions and simplifications, but it may not be possible to draw strong conclusions from a small number of trials and repeatability is often low.

Cost: Test vary dramatically in cost. In addition to monetary costs, situated tests require the time of domain experts and personnel. A special concern for security domains is that simulated attacks where security personnel are not informed before the event may be quite dangerous to participants.

4.2 Quantitative Metrics

We now shift our attention to the variety of different metrics that different tests can measure. The fundamental goal of a security system is to maximize utility, which can be decomposed into minimizing deployment cost, attack frequency, and expected damage of attacks. These *primary* metrics are not directly measurable in all types of tests, so we must often fall back on *secondary* metrics that are correlated with one or more primary metrics (and therefore, overall utility). Here we describe a representative set of such secondary metrics, commenting on their benefits and detriments.

- **Attacks Prevented:** How many attacks in progress are interdicted? *Pro:* This metric directly measures the benefit of reduced attack damage/frequency. *Con:* The total number of attempted attacks may be unobservable (for instance, it is not known how many weapons have been smuggled past ARMOR checkpoints) and quite rare.
- **Attacks Deterred:** How many planned attacks are abandoned due to security measures? *Pro:* Attack deterrence may be a primary benefit of security [Jacobson *et al.*, 2005; Bier, 2007]. *Con:* Deterrence is generally impossible to measure directly.
- **Planning Cost:** How much time and cost is necessary to plan an attack? *Pro:* Increased planning costs provides deterrence and opportunities to detect terrorist activities before an attack. *Con:* This cost is difficult to measure directly, and motivated attackers may have significant planning resources.[3]
- **Attacking Resources Required:** Can a single attacker with simple equipment cause significant damage? Or is sophisticated equipment and/or multiple attackers required? *Pro:* Like increasing planning cost, increased resources require larger attacker efforts, improving the chance of detection or infiltration. *Con:* Attackers may have sufficient resources, regardless.
- **Attack Damage:** What is the expected consequence of a successful attack? *Pro:* Possible consequences are relatively easy to estimate, as they are less dependent on human decisions. *Con:* Determining which attacks are most likely is still difficult, and there may be high variance.

[3] For instance, see `http://www.globalsecurity.org/security/profiles/dhiren_barot.htm`

Multiple assumptions must hold about the attackers' behavior and preferences for the reasoning to be correct [Erev *et al.*, 2002].

- **Implementation Cost:** What are the implementation and maintenance costs for a particular measure, including detrimental effects such as inconvenience to passengers, lower cargo throughput, etc.? *Pro:* Such a measurement can help decide which security measurements to implement. *Con:* All effects, positive and negative, must be quantified.

- **Expert Evaluation:** Are domain experts satisfied with the system? *Pro:* Security experts, who spend their career addressing such issues, have well informed opinions about what works and what does not. *Con:* Expert evaluations may identify security flaws but generally are not quantitative nor consistent across different experts.

5 Evaluation Options

The previous sections introduced a classification system for different types of tests and metrics that be useful to measure. We now list and discuss possible evaluations that can be conducted in a security domain, in the context of the above discussion. The evaluation options are situated within the proposed framework and categorized according to the type of test, relative accuracy, cost, and which metric(s) can be measured. The decision of which test(s) to run requires weighing each of these factors.

1. **Game Theoretic Analysis:** Given assumptions about the attacker (e.g., the payoff matrix is known), game theoretic tools can be used to determine the attacker's expected payoff. Additionally, deterrence can be measured by including a "stay home" action, returning neutral reward.[4]

   (a) **Attacker Resources vs. Damage:** A game theoretic analysis can evaluate how attacker observation, equipment, and attack vectors can change the expected attacker payoff. Only defensive measures known by the researcher can be considered, but such an analysis will provide an estimate of attack difficulty, an indirect measure of deterrence.

   (b) **Defense Dollars vs. Successful Attack:** A game theoretic analysis can measure how attacker success varies as security measures are added (e.g., implementing a new baggage screening process), or increasing the strength of an existing measure (e.g., adding checkpoints). Such an analysis may help ensure that resources are not overcommitted and provide organizations with quantitative data to assist with budgeting.

2. **Simulated Attacks:** A simulator with more or less detail can be constructed to model a specific security scenario. Such modeling may be more realistic than a game theoretic analysis because structure layout, simulated guard capabilities, and agent-level policies[5] may be incorporated.

3. **Human Studies:** Human psychological studies can help to better simulate attackers in the real world. Evaluations on an abstract version of the game may test base assumptions, or a detailed rendition of the target in a virtual reality setting with physiological stress factors could test situated behavior. Human subjects may allow researchers to better simulate the actions of attackers, who may not be fully rational. Human tests suffer from the fact that participants are not drawn from the same population as the actual attackers.

4. **Foiled Attacks:** The number of attacks disrupted by a security system can provide a sanity check (i.e., it disrupts a non-zero number of attacks). If the metric is correlated with an estimated number of attacks, it may help estimate of the attacker percentage captured. Enabling and disabling the security system and observing how the number of foiled attacks changes would be more accurate, but this methodology is likely unethical in many real world settings.

5. **Red Team:** Tests in which a "Red Team" of qualified security personnel attempt to probe se-

[4] Some attackers may be set on attacking at any cost and may be modeled with a "stay home" action returning a large negative reward.

[5] One exciting direction, as yet unexplored, is to incorporate machine learning into such policies. Such an extension would allow attackers to potentially discover flaws in the system, in addition to modeling known attacker behaviors.

curity defenses provides realistic information in life-like situations using the true defenses (including those that are not visible). However, such a test is very difficult to conduct as some security must be alerted (so that the team is not endangered) while remaining realistic, the tests are often not repeatable, and a single test is likely unrepresentative.

6. **Expert Evaluation:** Security experts — internal or external — may holistically evaluate a target's defenses, including both visible and non-visible, and provide a high-level security assessment.

7. **Deterrence Measurement:** Different methods for directly estimating deterrence can be used, such as estimating how likely an attacker is to know about a security precaution and how that knowledge will affect the likelihood of attack. A more quantitative approach would allow attackers to choose between different actions that attack the defended target and actions that attack a different target.[6]

8. **Cost Study:** A cost estimate for an entire location may examine multiple security measures and different levels of staffing, as well as measuring each resource's total cost. Some intangible factors may be very difficult to determine, such as quantifying a decrease in civil liberties.

6 ARMOR Evaluation, Revisited

This section first re-examines the current evaluations presented in Section 3 to summarize the state of the system's evaluation and then discusses what additional experiments could/should be performed based on the framework presented above.

Existing evaluations of the deployed ARMOR system fall into the Mathematical, Controlled Laboratory, Natural Experiments, and Qualitative categories. These represent a fairly broad range of types of evaluations, showing that ARMOR works well in theory, and that security experts agree it is beneficial. The controlled laboratory experiments, qualitative evaluations, and (sparse) data from natural experiments are particularly interesting in that they go beyond the framework of the game-theoretic model to test it's key assumptions. In many ways, this level of evaluation goes beyond what is typical of applications, even those deployed in real-world settings. Overall, we find strong evidence to support the use of ARMOR over previous methods (notably, hand-crafted or uniform random schedules).

Nevertheless, our framework also suggests new directions that could fill in gaps in the existing evaluation of ARMOR. This is particularly important as we move forward and wish to evaluate ARMOR against more sophisticated alternatives than the hand-crafted and uniform random baselines. In cases where the comparison is less clear-cut, we may need additional metrics to make a compelling argument for one approach or another. Based on our analysis above, we suggest several possible directions for future evaluations of ARMOR and similar systems:

1. None of the current evaluations effectively measure the cost of deploying ARMOR. New analysis should estimate the cost of deploying ARMOR at a new location, both in monetary terms and in side effects. For example, does using ARMOR result in increased congestion or wait times for travelers? It would also be useful to quantify the time required to create hand-crafted schedules instead of using ARMOR.

2. Any additional data we can gather about the effects of ARMOR on risk at LAX would be incredibly valuable for evaluation. Hard numbers are quite difficult to obtain due to security concerns, but efforts to find alternatives should continue. One that is frequently suggested is using security experts to "Red Team" the airport and plan or simulate attacks against it. While this would undoubtedly provide useful information, it is very costly and would require the approval of the airport authorities. Truly live red team operations are generally not conducted due to the risks they create for security personnel.

3. It would be useful to correlate the number of suspected attackers stopped at checkpoints with the number of suspected attackers stopped by other security methods over time. If the number of people detained at checkpoints increases after ARMOR was deployed and the number of people

[6] Although this may seem myopic, institution-level security measures are designed to protect a single target; if ARMOR causes attackers to be deterred and attack elsewhere, the security measure has successfully defended LAX. If a measure was designed to cause attackers to *never* attack (or fail at any attack, anywhere), our definition of deterrence would have to be significantly modified.

Evaluation Summary

Test	Type	Accuracy	Cost	Prevented	Deterred	Plan Cost	Resources	Cost	Damage	Qualitative
Game Theory	Mathematic	High	Low	✓	✓				✓	
Attacker Resources/Payoff	Mathematic	High	Low	✓		✓	✓		✓	
Defense Dollars / Damage	Mathematic	High	Low	✓				✓	✓	
Simulated Attacks	Simulation	High	Low	✓					✓	
Human Studies	Human	Med	Med	✓					✓	
Foiled Attacks	Natural	Low	Low	✓						
Red Team	Situated	Low	High	✓					✓	
Expert Evaluation	Qualitative	Low	Med	✓	✓	✓	✓		✓	✓
Deterrence Measurement	Math. / Qual.	Low	Low		✓	✓				✓
Cost Study	Math. / Qual.	Med	Low	✓				✓		✓

Table 1: This table summarizes our proposed evaluation methods by suggesting where each falls along the three general dimensions and which of the seven security-specific metrics are measured.

detained by other methods stayed the same (or fell), it is likely that ARMOR is more successful than the previous checkpoint strategy. Currently, such arrest statistics are considered sensitive and are not available to researchers.

4. Another approach for testing the assumptions of our game-theoretic model and the quality of the payoffs elicited from the security experts is to build more detailed computer simulations of airport operations and potential attack scenarios. These simulations themselves also make assumptions, but it would potentially improve reliability to model and understand the domain using two very different modeling frameworks at different levels of abstraction.

5. A weakness of the current evaluation is the lack of an effective measure of deterrence. This is an inherently difficult aspect to capture, as the important variables cannot be observed in practice. One possibility is to explore deterrence more carefully in the game-theoretic model. For example, attackers could be given the option of attacking other targets in addition to LAX. Combined with sensitivity analysis and behavioral experiments, this could be a way to better understand the effects of deterrence.

7 Related Work

Security is a complex research area, spanning many disciplines, and policy evaluation is a persistent challenge. Many security applications are evaluated primarily on the basis of theoretical results; situated evaluations and even laboratory experiments with human subjects are relatively rare. In addition, existing general methodologies for risk and security evaluation often rely heavily on expert opinions and qualitative evaluations.

Lazarick [1999] is a representative example which relies heavily on expert opinions. In the study, seven tools/approaches used to evaluate airport security were compared as part of a competitive bidding process. At the end of the multi-year security initiative, the highest ranked evaluation methodology relied on averaging qualitative expert evaluations.

A second example of a high-level methodology for per-facility and regional risk assessment, such as described by Baker [2005]. The methodology relies heavily on expert opinions and evaluations from local technical staff/experts, similar to Lazarick [1999]. The three key questions in the methodology are: (1) Based on the vulnerabilities identified, what is the likelihood that the system will fail? (2) What are the consequences of such failure (e.g., cost or lives)? (3) Are these consequences acceptable? Such an approach enumerates all vulnerabilities and threats in an attempt to determine what should (or must) be improved. There is no quantitative framework for evaluating risk.

Many in the risk analysis community have recently argued for game theory as a paradigm for security evaluation, with the major advantage that it explicitly models the adaptive behavior of an intelligent adversary. Cox [2008] provides a detailed discussion of the common "Risk = Threat × Vulnerability × Consequence" model, including analysis of an example use of the model. There are several arguments raised as weaknesses of the approach, including (1) the values are fundamentally subjective (2) rankings of risk are often used, but are insufficient (3) there are mathematical difficulties with the equation, including dependencies between the multiplied terms, and (4) the model does not account for adaptive, intelligent attackers. One of the main recommendations of the paper is to adopt more intelligent models of attacker behavior, instead of more simple, static, risk estimates.

Bier et al. [2009] provide a high-level discussion of game-theoretic analysis in security applications and their limitations. The main argument is that the *adaptive* nature of the terrorist threat leads to many problems with static models — such models may overstate the protective value of a policy by not anticipating an attacker's options to circumvent the policy. They explicitly propose using quantitative risk analysis to provide probability/consequence numbers for game-theoretic analysis.

Beir [2007] performs a theoretical analysis of the implications of a Bayesian Stackelberg security game very similar to the one solved by ARMOR, although most of the analysis assumes that the defender does *not* know the attacker's payoffs. The primary goal is to examine intuitive implications of the model, such as the need to leave targets uncovered in some cases so as not to drive attackers towards more valuable targets. There are no "real world" evaluation of the model. Other work [Bier *et al.*, 2008] considers high-level budget allocation (e.g., to large metropolitan areas). While the study uses real data, its focus is not model evaluation but the implications resulting from the model.

Game theory does have much to offer in our view, but should not be considered a panacea for security evaluation. One difficulty is that human behavior often does not correspond exactly to game-theoretic predictions in controlled studies. Weibull [2004] describes many of the complex issues associated with testing game-theoretic predictions in a laboratory setting, including a discussion of the ongoing argument regarding whether people typically play the Nash equilibrium or not (a point discussed at length in the literature, such as in Erev et al. [2002]). This is one reason we believe behavioral studies with humans are an important element for security system evaluation.

Many of the issues we describe in acquiring useful real-world data for evaluation purposes are mirrored in other types of domains. Blundell and Costa-Dias [2009] describe approaches for experimental design and analysis of policy proposals in microeconomics, where data is limited in many of the same ways: it is often not possible to run controlled experiments and many desired data cannot be observed. They describe several classes of statistical methods for these cases, some of which may be valuable in the security setting (though data sensitivity and sparse observations pose significant additional challenges). In truth, it is often hard to evaluate complex deployed systems in general — in our field a test of the prototype often suffices (c.f., Scerri et al. [2008]).

Jacobson et al. [2005] describe a deployed model for screening airline passenger baggage. The model includes detailed information regarding estimated costs of many aspects of the screening process, including variables for probability of attack and cost of a failed detection, but these are noted to be difficult to estimate and left to other security experts to determine. One particularly interesting aspect of the approach is that they perform sensitivity analysis on the model in order to assess the effect of different values on the overall decisions. Unfortunately, the authors have little to say about actually setting the input values to their model; in fact, there is no empirical data validating their screening approach.

Kearns and Ortiz [2003] introduce algorithms for a class of "interdependent" security games, where the security investment of one player has a positive externality and increases the security of other players. They run the algorithms on data from the airline domain but do not directly evaluate their approach, instead looking at properties of the equilibrium solution and considering the broad insight that this solution yields regarding the benefits of subsidizing security in such games.

Lastly, the field of *fraud detection* [Kou *et al.*, 2004], encompassing credit card fraud, computer intrusion, and telecommunications fraud, is also related. Similar to the physical security problem, data is difficult to access, researchers often do not share techniques, and deterrence is difficult (or impossible) to measure. Significant differences include:

1. Humans can often classify (in retrospect) false positives and false negatives, allowing researchers to accurately evaluate strategies.

2. Companies have significant amounts of data regarding known attacks, even if they do not typically share the data outside the company. Some datasets do exist for common comparisons (c.f., the 1998 DARPA Intrusion Detection Evaluation data[7]).

3. The frequency of such attacks is much higher than physical terrorist attacks, providing significant training/evaluation data.

4. Defenders can evaluate multiple strategies (e.g., classifiers) on real-time data, whereas physical security may employ only, and evaluate, one strategy at a time.

8 Conclusions

While none of the evaluation tests presented in Section 5 can calculate a measure's utility with absolute accuracy, understanding what each test *can* provide will help evaluators better understand what tests *should* be run on deployed systems. The goal of such tests will always be to provide better understanding to the "customer," be it researchers, users, or policy makers. By running multiple types of tests, utility (the primary quantity) can be approximated with increasing reliability.

At a higher level, thorough cost-benefit analyses can provide information to policy makers at the inter-domain level. For instance, consider the following example from Tengs and Graham [1996]:

> To regulate the flammability of children's clothing we spend $1.5 million per year of life saved, while some 30% of those children live in homes without smoke alarms, an investment that costs about $200,000 per year of life saved.

While such a comparative cost-benefit analysis is beyond the scope of the current study, these statistics show how such an analysis can be used to compare how effective measures are across very different domains, and could be used to compare different proposed security measures.

In the future we plan to use this framework to help decide which evaluation tests are most important to determine ARMOR's utility, as suggested in Section 6. Additionally, we intend to continue collaborating with security experts to determine if our framework is sufficiently general to cover all existing types of security tests, as well test how the framework can guide evaluation in additional complex domains.

Acknowledgements

The authors would like to thank James Pita and Vicki Bier for helpful suggestions, Prakhar Garag and Rahul Jain for assistance in running experiments, Sergeant Jim Holcomb for supplying ARMOR-related arrest data, and First Sargent Cruz for discussing pre-ARMOR scheduling procedures. This research was supported by the United States Department of Homeland Security through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) under grant number 2007-ST-061-000001. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the United States Department of Homeland Security.

References

[Baker, 2005] G. H. Baker. A vulnerability assessment methodology for critical infrastructure sites. In *DHS Symposium: R and D Partnerships in Homeland Security*, 2005.

[Bier *et al.*, 2008] V. M. Bier, N. Haphuriwat, J. Menoyo, R. Zimmerman, and A. M. Culpen. Optimal resource allocation for defense of targets based on differing measures of attractiveness. *Risk Analysis*, 28(3):763–770, 2008.

[Bier *et al.*, 2009] V. M. Bier, Jr. L. A. Cox, and M. N. Azaiez. Why both game theory and reliability theory are important in defending infrastructure against intelligent attacks. In *Game Theoretic Risk Analysis and Security Theats*, volume 128. Springer US, 2009.

[Bier, 2007] V. M. Bier. Choosing what to protect. *Risk Analysis*, 27(3):607–620, 2007.

[Blundell and Costa-Dias, 2009] R. Blundell and M. Costa-Dias. Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 2009.

[Conitzer and Sandholm, 2006] V. Conitzer and T. Sandholm. Computing the optimal strategy to commit to. In *Proc. of EC*, 2006.

[7] See `http://www.ll.mit.edu/mission/communications/ist/index.html` for data and program details.

[Cruz, 2009] First Sargent Cruz. Personal communication, August 20 2009.

[Erev *et al.*, 2002] I. Erev, A. E. Roth, R. L. Slonim, and G. Barron. Predictive value and usefulness of game theoretic models. *International Journal of Forecasting*, 18(3):359–368, 2002.

[Jacobson *et al.*, 2005] S. H. Jacobson, T. Karnai, and J. E. Kobza. Assessing the impact of deterrence on aviation checked baggage screening strategies. *International J. of Risk Assessment and Management*, 5(1):1–15, 2005.

[Kearns and Ortiz, 2003] M. Kearns and L. E. Ortiz. Algorithms for interdependent security games. In *Neural Information Processing Systems (NIPS)*, 2003.

[Kiekintveld *et al.*, 2009] Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordóñez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. In *AAMAS*, 2009.

[Kou *et al.*, 2004] Y. Kou, C. Lu, S. Sinvongwattana, and Y.P. Huang. Survey of fraud detection techniques. In *Proc. of IEEE Networking*, 2004.

[L. A. Cox, 2008] Jr. L. A. Cox. Some limitations of “risk = threat x vulnerability x consequence” for risk analysis of terrorist attacks. *Risk Analysis*, 28(6):1749–1761, 2008.

[Lazarick, 1999] R. Lazarick. Airport vulnerability assessment – a methodology evaluation. In *Proc. of 33rd IEEE International Carnahan Conference on Security Technology*, 1999.

[Paruchuri *et al.*, 2008] Praveen Paruchuri, Jonathan P. Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Playing games with security: An efficient exact algorithm for Bayesian Stackelberg games. In *AAMAS-08*, 2008.

[Pita *et al.*, 2008] J. Pita, M. Jain, C. Western, C. Portway, M. Tambe, F. Ordonez, S. Kraus, and P. Paruchuri. Deployed ARMOR protection: The application of a game theoretic model for security at the Los Angeles International Airport. In *Proc. of AAMAS*, 2008.

[Pita *et al.*, 2009] J. Pita, M. Jain, M. Tambe, F. Ordonez, S. Kraus, and R. Magori-Cohen. Effective solutions for real-world stackelberg games: When agents must deal with human uncertainties. In *Proc. of AAMAS*, 2009.

[Scerri *et al.*, 2008] P. Scerri, T. Von Goten, J. Fudge, S. Owens, and K. Sycara. Transitioning multiagent technology to UAV applications. In *Proc. of AAMAS Industry Track*, 2008.

[Stevens *et al.*, 2009] D. Stevens, T. Hamilton, M. Schaffer, D. Dunham-Scott, J. J. Medby, E. W. Chan, J. Gibson, M. Eisman, R. Mesic, C. T. Kelly, J. Kim, T. LaTourrette, and K. J. Riley. Implementing security improvement options at Los Angeles international airport, 2009. `www.rand.org/pubs/documented_briefings/2006/RAND_DB499-1.pdf`.

[Tengs and Graham, 1996] T. O. Tengs and J. D. Graham. Risks, costs, and lives saved: Getting better results from regulation. In *The opportunity costs of haphazard social investments in lifesaving*. American Enterprise Institute, Washington, 1996.

[Wagenaar, 1972] W. A. Wagenaar. Generation of random sequences by human subjects: A critical survey of literature. *Psychological Bulletin*, 77(2):65–72, 1972.

[Weibull, 2004] J. Weibull. Testing game theory. In Steffen Huck, editor, *Advances in Understanding Strategic Behavior: Game Theory, Experiments and Bounded Rationality.*, pages 85–104. Palgrave MacMillan, 2004.