When Hackers Err – The Impacts of False Positives on Information Security Games

Abstract

False positive rates and their impacts have been a focal point for information security research. However, most of this research investigates false positives exclusively from the system defender's perspective, while in reality an attacker also faces the classification decision in identifying feasible targets and the consequences of false positive rates. In this paper, we present the first comprehensive analytical model that incorporates the false positives both from the perspective of the attacker as well as of the system defender. Our results show that such false positives from the attacker's perspective have a significant impact on the attacker's decision making for an attack, as well as the optimal protection strategy for the defender. Our results help to shed new light on a wide range of diverse information security phenomena such as spam emails, the Nigerian scams, and the design of Honeypot as a security mechanism. In addition, we show how an attacker's mis-estimation of certain parameter would affect the defender's strategy, and how the heterogeneity of the systems impacts the defender's strategy to manipulate the attacker's possible mis-estimation.

Key words: information security game, false positive, classification, spam, Nigerian scam, Honeypot, mis-estimation

1. Introduction

In extant information security studies, a significant research focal point has been the study of the impacts of false positive rates in detecting an attacker from a normal user on the performance of the security mechanism in place (e.g., Chivers et al. 2013, Rezk et al. 2011, Cavusoglu et al. 2013, Persico and Todd 2005). When designing a security mechanism for an information system, it is fundamental that the security mechanism contains a component that is able to discriminate a potential malicious attacker from a normal user, and as a result interact with the two types of users differently. Email spam filter is a simple example of such a classification component in an email system. A spam filter scans all incoming emails, and based on certain criteria it will classify some emails as spam emails and others as normal emails. As a result, the two different types of incoming emails are treated differently: the former will be flagged, quarantined, or even directly deleted, while the latter will be delivered directly to the recipient's inbox. Another example of such security classification components is an Intrusion Detection Systems (IDS), which scans all incoming user traffic requesting access to the information system, and classifies some user traffic as malicious and the remainder as normal. As a result, traffic deemed malicious is terminated and usually further measures are taken to handle users generating that traffic, and those users whose traffic is deemed normal are granted access to the information system.

No classification is perfect, and two types of classification errors arise (Trees et al. 2013): the false positives and the false negatives. In the IDS example, a false positive refers to the scenario where a normal user is incorrectly classified as an attacker, whereas a false negative refers to the scenario where an attacker is incorrectly classified as a normal user. By the very nature of such a classification scheme, the two types of errors are inter-related in such a way that they must be traded off against each other (Trees et al. 2013). One critical design issue for the classification component is to find the optimal trade-offs between these two types of errors. In different fields, the related costs for false positives and false negatives vary significantly and thereby, the trade-off decision also varies significantly among various fields. For example, as described by Herley (2012), in the medical field when deciding on whether to have a surgery, the potential cost structure may result in the situation where a false positive (i.e. conducting an unnecessary surgery) is more tolerable than a false negative (i.e., not conducting a necessary surgery); on the other hand, in criminal justice, the potential social costs may be such that a false negative (i.e., letting a guilty person go free) is more acceptable than a false positive (i.e., punishing an innocent person). In the information security domain, the impacts of false positives in the classification have been a focal point of research. Extensive studies have been conducted to investigate how false positives affect information security investment, security systems design, and security systems performance (e.g. Cavusoglu et al 2005, 2009, Ulvila and Gaffney 2004, Guan et al 2017, Salman et al 2014, Duan et al 2012).

In the decision analysis literature, the impact of false positives on decision making is also a focal point of research. Cavusoglu and Raghunathan (2004) and van Binsbergen and Marx (2007) had argued about the relative merits of decision analysis and game theory in attacker detection, and incorporated false positives as important factors in the decision making. Sophisticated decision models that specifically incorporate the element of false positives have been proposed to address the optimal decision making in a wide variety of scenarios such as national security (Blum and Pate-Cornell 2016), cargo security (Bakir 2008), bio-surveillance

(Willis and Moore 2014), breast cancer diagnosis (Alagoz et al 2013), and counter-terrorism (Ni et al 2013).

Almost all studies in the information security and decision analysis domain that investigate the impact of false positives do so from the perspective of the primary system defender classification (i.e., what happens if the system defender mistakenly classifies a normal user as an attacker, or vice versa?). But in reality, a potential attacker also faces this classification issue and the trade-off problem that it entails. From the perspective of the attacker, not all potential targets are equal: some are feasible targets (e.g., targets with weak security mechanisms) that may yield a high payoff to the attacker if successfully attacked while the others are infeasible targets (e.g., targets with strong security mechanisms) that may result in a low payoff or even a severe penalty to the attacker if attacked.

Security professionals have long advocated the importance of good security decision making and they promote the optimal adoption of security technology and implementation of security policies. However, not all entities follow these (usually excellent) pieces of advice (e.g., Bulgurcu et al. 2010, Liang et al. 2013, Siponen et al. 2014, Stahl and Pease 2011). Even within one single organization, where there is usually a centralized security team that implements security policies and adopts security technologies across all information systems organization-wide, the actual usage of the security technology and adherence to security policies eventually falls on the shoulders of the people who are operating the individual systems. Unfortunately, not all of those involved in operations adhere to the established policies or use the security tools in the proper way (Solms and Solms 2004, Son 2011, Hannah and Roberson 2015), thus making such information systems more vulnerable than others within the same organization. As a result, when an attacker attempts to attack information systems targets inside an organization, he may be motivated to endeavor to identify those weaker targets before engaging in the actual attack behavior. Compounded with the fact that implementing an attack usually incurs a non-zero cost, it seems plausible that an attacker would classify potential targets before carrying out the attacks, presumably focusing on those targets classified as weak. One implication of this strategy is that the attacker must also face the consequences of target classification errors, including both false positives and false negatives. Furthermore, such attacker behavior will also influence the system defender's optimal strategy in protecting the information assets facing possible attacks.

While the majority of the security literature studies the security decision making from the perspective of the defender, there are a number of studies that investigate the phenomena from the attacker's perspective. For example, Brandt and Sandler (2010) studied how terrorists adjust their attack target choices based on target hardening. Hausken (2016) conducted a cost–benefit analysis of terrorist attacks. Hausken (2017) studied the attacking and information sharing strategy of multiple attackers when choosing among multiple targets. Herley (2013) studied the behaviors of multiple attackers seeking victims among a finite population.

However, to the best of our knowledge, few studies have been conducted to understand the impacts of false positives in classifying potential targets from an attacker's standpoint (with the notable exception of Herley 2012) and furthermore, no current study (including Herley 2012) has been done to integrate this type of attacker's behaviors into the optimal strategy design from the standpoint of the system defender.

In this paper, we fill the void of this aspect of information security decision analysis and provide the first analytical model that incorporates the classification components for both the attacker and the system defender. In our setting, the attacker-defender interaction is modeled as a Stackelberg game. The attacker has the option to use a classification component to identify weak targets and then implement the attack. The system defender optimally configures the classification components in her organization's security systems to identify the attacker, taking into full consideration the potential attacker's possible attacking strategy based on their classification, to maximize the overall expected payoffs to the organization.

The remainder of the paper is organized as follows: in Section 2, we provide a general introduction delineating how to model errors in classification. In Section 3, we illustrate our model of the game between the attacker and the system defender, followed by Section 4, where we describe and analyze the decision process of both the attacker and the system defender, based on our model. In Section 5, we extend our model to investigate the situation where the attacker does not have complete information about the target systems. In Section 6, we extend our model to analyze the situation where there is further heterogeneity among the users and the systems. We summarize our findings and conclude in Section 7.

2. Modeling the errors in classification

The classification components described in Section 1 are binary classifications for decision making: the attacker tries to discriminate feasible targets from infeasible targets, and the system defender tries to discriminate attacker from normal users. The trade-offs between the two types of errors can be modeled by a Receiver Operator Characteristic (ROC) curve (Trees et al. 2013). The ROC curve captures the inter-relationship between a classifier's true positive rate and its false positive rate, and provides the foundations for the trade-offs between the two types of errors.

An ROC curve depicts the classifier's true positive rate as an increasing concave function of its false positive rate. Any classifier can be configured only according to the points located on its ROC curve. Intuitively, we want the true positive rate to be as high as possible and the false positive rate to be as low as possible. However, in reality, the ROC curve dictates that as one increases the true positive rate of the classifier, its false positive rate also increases, and one can only configure the classifier such that its true positive and false positive rates fall on the ROC curve. That is the core reason for the trade-offs between the two types of errors. It is reasonable to assume that any good classifier would perform better than random guessing. In other words, it must have its true positive rate higher than its false positive rate. This is also the assumption for all the classification mechanisms used in our model.

There is a large body of literature on how to construct an ROC curve for a classifier either experimentally or empirically (e.g., Durst et al. 1999, Lippmann et al. 2000, McHugh 2000). In this paper, we adopt an explicit functional form for the relationship between the true positive rate (tp) and false positive rate (fp) as: $tp = fp^w$, where $0 < w < 1$. As derived by Cavusoglu et al. (2005), this power function is obtained if the risk score calculated by the classifier follows an exponential distribution. This functional form is theoretically rigorous, and widely accepted by security decision analysis researchers in their models for security detection (e.g., Cavusoglu et al. 2013, Zhuang and Bier 2007, Gao and Zhong 2015).

3. Modeling the game between system defender and users

We now model the interaction between the system defender and users of information systems as a complete information[1] sequential game. In our setting, we define the components of the game, as well as their parameters, as follows:

Users: There are a total of N users trying to access information systems in an organization. All users are homogeneous in all aspects except one: λ (λ<1) proportion of the users are attackers (denoted as A), while the others are normal users (denoted as L). We define attackers as the users whose successful access to the system will cause damage *d* to the organization; we define normal users as the users whose successful access to the system will generate positive utility *u* to the organization[2]. We assume that any user will try to access one and only one information system. While any of the users' type cannot be directly observed, the value of λ is assumed to be public knowledge.

Systems: The organization has a total of M systems accessible to all users. All systems are assumed to be homogeneous in all aspects except one: θ (θ <1) proportion of the systems are weak (denoted as S_w), while the others are strong (denoted as S_s)[3]. We define weak systems as those systems whose security mechanism has not been functioning due to any reason(s); we define strong systems as those systems whose security mechanism is fully functioning, as configured by the system defender. In reality, weak systems may be caused by their operators not following security protocols in operating the system, not adhering to security policies when

[1] In Section 5 we will relax this assumption of complete information and study the impact of attackers misjudging the model parameters.

[2] In Section 6 we will extend our model and study the situation where there are multiple priority levels of users as they generate different levels of utility.

[3] Again in Section 6 we will study the situation where there are multiple priority levels of the systems as they generate different levels of utility.

using the system, or not properly maintaining the security mechanism of the system. In any case, we assume the outcome of such negligence as the security mechanism of these systems not functioning, which results in any user attempting to access such systems being able to gain access directly. It is also assumed that while it is not directly observable which systems are weak, the value of θ is public knowledge. This type of assumption is quite common in the analytical modeling literature about information security. For example, in Cavusoglu and Raghunathan (2004) and Cavusoglu et al (2008), it is assumed that all parameter values in their analytical model for information security are public knowledge, including the firm's investment in information security and the target information system's vulnerability level. In Section 5, we extend our model by relaxing this assumption of θ being known by the attacker, and study how the attacker's misjudgment of θ would affect his decision making, the defender's strategy, and the organization's overall expected payoffs.

Security Mechanism: In our model, a system contains a security component that can monitor the incoming user traffic, and try to differentiate and thereby detect attackers from normal users. The performance of this detection component is, as described in Section 2, captured by its ROC curve. We denote the component's detection (true positive) rate as P_d, which is configurable by the system defender, and its associated false positive rate as P_f. We assume the relationship captured by the detection component's ROC is $P_d = P_f^s$, with s (0<s<1) as the parameter defining the shape of the ROC.

System Defender's Decision Making: There is a system defender who will decide the optimal configurations of the security detection component of the systems to maximize the overall expected payoffs for the organization, fully aware that in θ portion of the M systems the detection component will not function. When the system defender, through the functioning detection component of the systems, identifies a user as a normal user, she grants the user system access (and subsequently reaps the utility *u* if the user is indeed a normal user, or sustains the damage *d* if the user is actually an attacker); when she identifies a user as an attacker, she denies the user system access (thus avoiding the potential damage of *d* if the user is indeed an attacker, but also missing out the potential utility of *u* if the user is actually a normal user).

Attacker's Decision Making: The attacker's decision is whether to randomly select a target system to try to gain access (i.e., randomly attack), or to detect a weak system and then try to gain access to that detected system (i.e., target attack). If the attacker chooses to use a detection mechanism to identify weak systems, the performance of such detection mechanism is captured by its own ROC curve. We assume such detection mechanism has detection rate of Q_d, and false positive rate of Q_f, where $Q_d = Q_f^r$, with r (0<r<1) as the parameter defining the ROC for the attacker's detection mechanism.

Normal User's Decision Making: We assume that a normal user always randomly selects a target system and tries to gain access.

Cost / Payoff Parameters: To execute the attack, it costs the attacker c_a; if the attack is successful, the attacker gains payoff of *v*. To implement the detection mechanism to identify weak systems, it costs the attacker c_t. To operate the security detection component on the systems, it costs the system defender z. If the attacker is correctly identified by the system defender, the attacker suffers a loss of h, and if the normal user is mistakenly identified as an attacker by system defender, the system defender suffers a loss of goodwill of l.

The timeline of the agent interactions is illustrated in the following Figure 1.

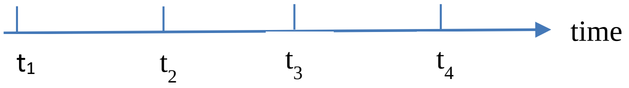

Figure 1. The timeline of the sequence of the agent interactions

At time t_1, nature determines which user is a normal user or an attacker. At time t_2, the actual two-period Stackelberg game starts with the system defender configuring the security detection mechanisms on the organization's information systems. At time t_3, the attacker responds with determining and implementing their attacking strategy. Finally at time t_4, all payoff/loss are realized.

We summarize the model notation in the following table:

Table 1. Summary of notation used in our model.

Notation	Definition
λ	Portion of the user population that are attackers
d	The damage to the organization after a successful attack
u	The utility to the organization generated by a normal user after accessing the system
θ	Portion of the organization's systems that are weak, i.e., their security detection mechanisms do not function
P_d	Detection (true-positive) rate of the security mechanism in detecting attacker
P_f	False-positive rate of the security mechanism in detecting attacker
s	The parameter defining the ROC curve for the security detection mechanism
Q_d	Detection (true-positive) rate of the detecting mechanism attacker uses in detecting weak systems
Q_f	False-positive rate of the detecting mechanism attacker uses in detecting weak systems
r	The parameter defining the ROC curve for the attacker's detecting mechanism
c_a	Attacker's cost to execute an attack
c_t	Attacker's cost to detect weak systems
v	Attacker's payoff after a successful attack
z	System defender's cost of operating the security mechanism on the systems
h	Loss to the attacker if he's been correctly detected by system
l	Loss of goodwill to the organization if a normal user is mis-identified as an attacker

4. The decision processes of the agents

4.1. The attacker's decision process

The attacker will decide between two options: randomly attack versus target attack.

If the attacker decides to randomly attack, there is a probability θ that he encounters a weak system, which the attacker will gain direct access to and carry out the attacks. As a result, the attacker will gain a payoff of v-c_a. But, there is also a probability of (1- θ) he will encounter a strong system, and face the security detection mechanism. There is a probability of P_d that the attacker is correctly detected by the security detection mechanism and thereby suffers a loss of –c_a-h; and a probability of (1-P_d) that he is mistakenly identified as a normal user and granted access to the system and thus obtains a payoff of v-c_a.

The attacker can also decide to target attack. In this case, his probability of encountering a weak system will be increased to P_1. Based on Bayes' Theorem, P_1 = (target is weak | target detected as weak) = P(target detected as weak | target is weak) * P(target is weak) / P(target detect as weak) = Q_d*θ / [Q_d*θ+Q_f*(1-θ)]. It should be apparent that the probability the attacker will encounter a strong system in this case is P_2=1-P_1.

Thus the possible realization for attacker's decision making setup can be illustrated by the following Figure 2: square A represents the attacker, square S_w represents a weak system and square S_s represents a strong system; the lines connecting the squares represent either the attacker's decision or the probability of the decision reaching the next node.

Therefore, the attacker's overall expected payoff if he chooses to randomly attack is: $E_r^a = (v-c_a)\theta + (-c_a - h)(1-\theta)P_d + (v-c_a)(1-\theta)(1-P_d) = (v-c_a) - (v+h)(1-\theta)P_d$. The attacker's overall expected payoff if he chooses to target attack is:

$E_d^a = (v - c_a - c_t) - (v+h)(1-P_1)P_d$, where $P_1 = Q_d\theta \big/ Q_d\theta + Q_f(1-\theta)$, as defined earlier.

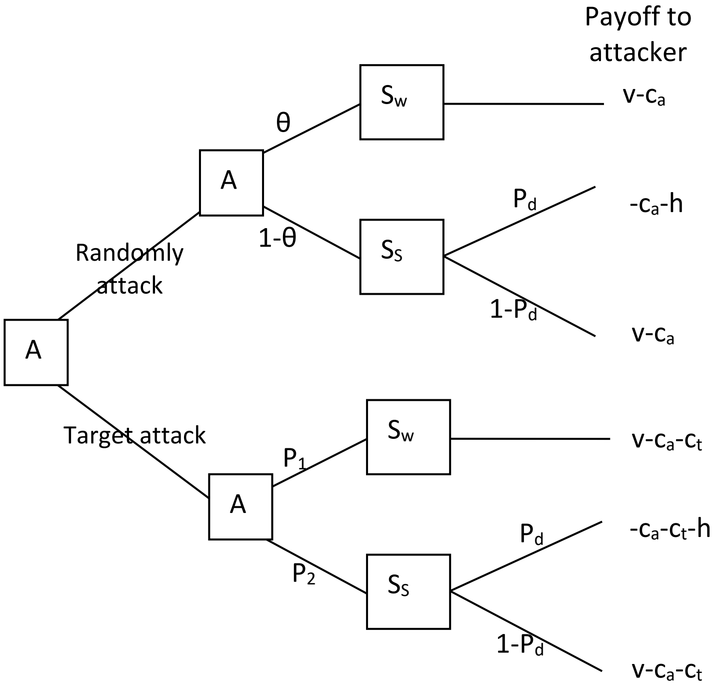

Figure 2. The possible realization for attacker's decision making

<u>The attacker's decision criteria:</u> We assume that when indifferent, the attacker will choose to randomly attack. Doing an algebraic comparison of the expected overall payoffs from the two strategies, we obtain the following result.

Lemma 1. The attacker will target attack, instead of randomly attack, if and only if $P_d(P_1 - \theta) > c_t \big/ v+h$. (1)

This result implies that the attacker will be motivated to target attack if any combinations of the following conditions hold:

1) The cost of detecting a weak system is low (i.e., the value of c_t is low);
2) The attacker's payoff after a successful attack is high (i.e., the value of v is high);

3) The attacker's loss after getting detected is high (i.e., the value of h is high);

4) The probability that the attacker will be detected is high (i.e., the value of P_d is high); or

5) The performance of the mechanisms the attacker uses in detecting weak system is good (i.e. the value of (P_1-θ) is high).

This result provides a theoretical explanation for some real-life attacker behaviors. For example, regular email spammers seldom detect recipients' susceptibility to spam messages before sending out spams. Our results show that some rationale behind their behavior in this case include: 1) the cost of detecting those vulnerable recipients is relatively high; and 2) since the spammer's main goal is to deliver the spam messages to the recipients' email inbox, the payoff consequences whether the goal is reached or not are relatively minor (i.e., a spam message successfully reaching a recipient's inbox in no way implies guaranteed payoffs to the spammer, and a spam message detected and quarantined by the recipient's email spam filter does not cause much, if at all, loss to the spammer at all). In other words, conditions 1 – 3 are not satisfied in this case. Therefore, email spammers seldom discriminate potential recipients before sending out spam emails.

On the other hand, this result also provides an explanation for the phenomenon that spam messages are used by the attacker as a useful tool to detect targets that are vulnerable to the real attacks the attacker is conspiring. For years, people have been continuously receiving spam messages involving scams such as recovering lost funds in Nigeria, transferring inheritance for a West African prince, or wiring money for a friend stuck in London. Usually the wording of these scam messages is so over-the-top, ridiculous, and outlandish that in reality quite possibly only those recipients who are the most naïve and gullible would fall for these type of scams. People would wonder why the scammers are creating spam messages in this seemingly unreasonable way.

Our result provides some insights into the rationale behind this behavior. For these types of scammers, a successful implementation of the full scam would yield significant payoff, and if they get caught, the punishment will also be harsh. Furthermore, the implementation of the whole scam is such a complicated process that at any point of the process the scam may be derailed and the scammers may get caught. Based on our results, this implies that conditions 2 – 4 are all satisfied. Therefore, the scammers are highly motivated to detect the feasible targets before implementing the scam.

BBC (2016) reported such a case, in which a scammer mastermind known as "Mike" was "alleged to head a network of 40 individuals behind global scams worth more than $60 million" (satisfying condition 2); the criminals also "allegedly ran a money laundering network" that spanned across three continents (satisfying condition 3); and they were arrested, facing charges "including hacking, conspiracy and obtaining money under false pretences" (satisfying condition 4). Our results indicate that these types of scammers would be highly motivated to detect the feasible targets before implementing the scam.

Preferably, the detection mechanism the scammers use should be of low cost (condition 1) and high discriminating power in differentiating feasible targets from infeasible ones (condition 5). The "Nigerian fund transfer" types of spam messages as described earlier serve as an excellent mechanism for the scammers to detect the feasible targets. First of all, sending out spam messages cost almost nothing (satisfying condition 1). Furthermore, by wording the spam message in such an outrageous way, the scammers are ensuring that the only targets who will respond positively to these messages are the ones who are not sophisticated enough to see through this thin facade of conspiracy. In other words, those are exactly the type of targets the scammer wants to attack. So these types of spam messages have high discriminating power in identifying those feasible targets (satisfying condition 5).

Therefore, these seemingly preposterously-worded emails are actually a very rational decision made by the attacker designed to serve as a low-cost, high-accuracy detection mechanism in identifying viable targets for the attacks. Our result complements the findings of Herley (2012) who showed that these spam messages can be used to filter out the infeasible targets.

Impact of the portion of weak systems: Now we rewrite the result in Equation (1) as

$$P_d > {}^{c_t}\!\big/_{(v+h)(P_1-\theta)} \qquad (2),$$

and denote its RHS as Ω. Taking the first order derivative of Ω w.r.t. θ, we obtain

$\frac{d\Omega}{d\theta} = \frac{c_t}{(h+v)(1-\theta)^2\theta^2} * \frac{Q_d\theta^2 - Q_f(1-\theta^2)}{Q_d - Q_f}$. Algebra shows that if $\theta \geq \frac{\sqrt{Q_d Q_f} - Q_f}{Q_d - Q_f}$, $\frac{d\Omega}{d\theta} \geq 0$, and if $\theta < \frac{\sqrt{Q_d Q_f} - Q_f}{Q_d - Q_f}$, $\frac{d\Omega}{d\theta} < 0$. We denote $\theta^{'} = \frac{\sqrt{Q_d Q_f} - Q_f}{Q_d - Q_f}$. This leads to the following result.

Proposition 1. There exists a threshold value for the proportion of weak systems in the organization (denoted as θ′) that if surpassed, the value of Ω will increase (i.e., the attacker will be more likely to randomly attack) with the increase of θ; and that if not surpassed, the value of Ω will decrease (i.e., the attacker will be more likely to target attack) with the increase of θ.

The implication of this result provides some theoretical basis for an attacker's behavior with regard to the number of weak systems in the organization. When the portion of weak systems in the organization is higher than the threshold value $\theta^{'}$ (e.g., because the information security education of the employees is insufficient), our result shows that the value of Ω will increase with the increase of the portion of the weak systems. That means it becomes more likely (1) will not hold, and thus the attacker will decide to randomly select a system to attack. The rationale behind this behavior is that there are a lot of weak systems in the organization, and the attacker doesn't need to take additional efforts (beyond randomly selecting a system) to detect weak systems to be able to successfully gain access to the weak systems. On the other hand, when the portion of the weak systems in the organization falls below the threshold value $\theta^{'}$ (e.g., because the information security culture is well established in the organization), even as the portion of the weak systems increases, the value of Ω still drops, making the likelihood higher for (1) to hold. That is, the attacker has to make more efforts to try to identify weak systems to attack, simply because there are not many weak systems in the organization.

Impacts of the false positives by attacker: Next we rewrite the result in Equation (1), as $Q_f < [\frac{(v+h)\theta P_d(1-\theta)-\theta c_t}{(v+h)\theta P_d(1-\theta)+(1-\theta)c_t}]^{\frac{1}{1-r}}$, and denote the RHS as Γ. It is noticed that $\frac{d\Gamma}{dr}$<0, which means as the value r increases, it becomes less likely Equation 1 would hold (i.e., less likely the attacker will target attack, instead of randomly attacking). Remember that for the attacker's detection mechanism for weak systems, we have $Q_d = Q_f^r$, which implies that as the value r increases, the values of Q_d and Q_f converge more, which in turn means that the detection mechanism's discriminating power is not very high. Combining these two results, we gain some insight on the role the attacker's false positive rates play in the attacker's strategic decision of attacking. Intuition suggests that if the attacker's detection mechanism is of high quality, the attacker should be more likely to decide to target attack. Our results not only confirm this intuition, but also illustrate the underlying mechanism through which the attacker makes such a decision: the attacker wants to detect the weak systems with a certain detection rate, and with a lower r value, he can assure a lower false positive rate at the same detection rate level, which makes it more likely Equation 1 would hold, which in turn makes it better off for the attacker to target attack. Our results suggest that for the attacker, the ideal mechanism for detecting weak systems should provide not only a high detection rate (i.e., high value for Q_d), but also high discriminating power (i.e., low value for r). Our earlier illustration of attacker using Nigerian scam as a detection mechanism to identify feasible targets for attacking is a good example of the attacker's strategic behavior in handling the false positives in identifying weak targets.

One important implication of the above analyses of the attacker's decision-making process and behavior is that to protect the organization's information systems, the system defender's decision making and strategic planning must take into account the attacker's potential behavior of detecting weak systems before attacking. This is what we will analyze next.

4.2. The system defender's decision process.

We now investigate how the system defender can configure its security detection mechanism to optimize the organization's overall expected payoffs.

From the perspective of the system defender, nature determines whether the user is an attacker or a normal user. Then the user tries to gain access to the system. The normal user always randomly selects a system to access. The attacker will target attack if Equation (1) or (2) holds, and randomly attack otherwise.

When a normal user accesses the system, there is a probability of θ that he will encounter a weak system and directly gain access, thus generating a payoff of u to the organization. There is also a probability of (1-θ) that he will encounter a strong system, whose detection mechanism will try to identify the user's type. There is a probability of P_f that the detection mechanism will mistakenly identify the user as an attacker, deny access to the user and thus result in the organization suffering a loss of (z+l). There is also a probability of (1-P_f) that the detection mechanism correctly identifies the normal user, grants access, and results in the organization obtaining a payoff of (u-z).

Similar analyses can be conducted on cases where an attacker randomly attacks, as well as where an attacker target attacks. In the case of attacker target attacking, the probability of encountering a weak system is increased from θ to P_1, as described earlier.

To summarize, the possible realization of the system defender's decision making can be illustrated by Figure 3: square User represent the system's user, who may be a normal user (represented by square L) or an attacker (represented by square A); the lines connecting the squares represent either attacker's decision or the probability of reaching the next node.

When the attacker decides to randomly attack (i.e., Equation (1) does not hold), the organization's overall expected payoff is:

$$\begin{aligned}\pi^r &= (1-\lambda)\theta u + (1-\lambda)(1-\theta)(1-P_f)(u-z) - (1-\lambda)(1-\theta)P_f(l+z) - \lambda\theta d \\ &\quad - \lambda(1-\theta)P_f^s z - \lambda(1-\theta)(1-P_f^s)(d+z)\end{aligned} \quad . \qquad (3)$$

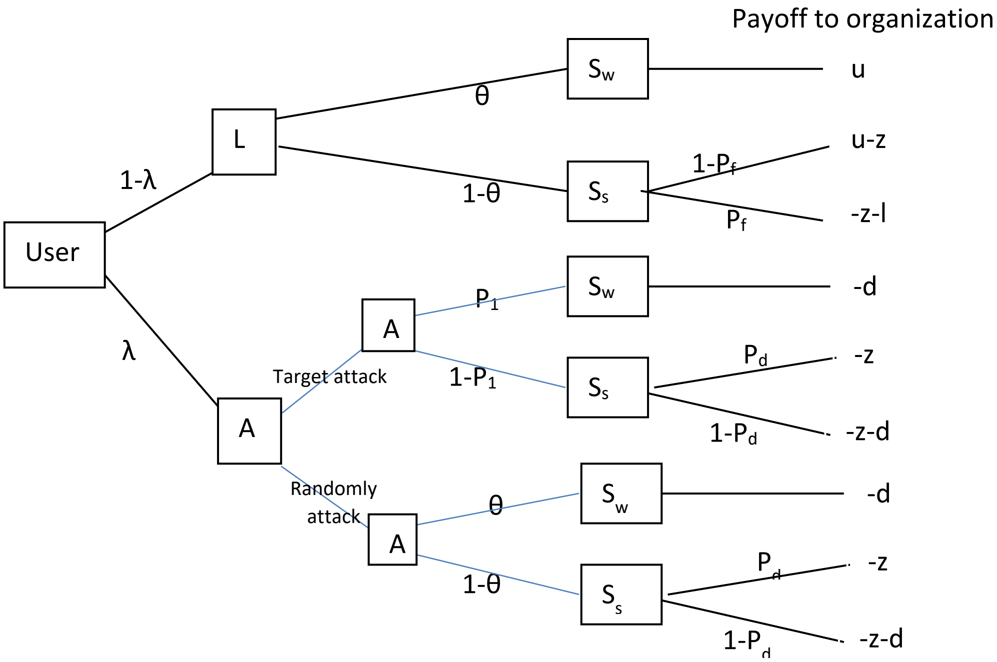

Figure 3. The possible realization for system defender's decision making

When the attacker decides to target attack (i.e., (1) holds), the organization's overall expected payoff is:

$$\begin{aligned}\pi^{d} &= (1-\lambda)\theta u + (1-\lambda)(1-\theta)(1-P_f)(u-z) - (1-\lambda)(1-\theta)P_f(l+z) - \lambda\left({}^{Q_f^r\theta}\!/_{Q_f^r\theta+Q_f(1-\theta)}\right)d \\ &\quad - \lambda\left(1-{}^{Q_f^r\theta}\!/_{Q_f^r\theta+Q_f(1-\theta)}\right)P_f^s z - \lambda\left(1-{}^{Q_f^r\theta}\!/_{Q_f^r\theta+Q_f(1-\theta)}\right)(1-P_f^s)(d+z)\end{aligned} \quad . \qquad (4)$$

Payoff functions (3) and (4) are the objective functions the system defender wants to maximize.

Therefore, the system protector's decision making would involve solving the following two constrained optimization problems:

1. $\boxed{\begin{array}{ll} \underset{P_d}{Max} & \pi^{d} \\ s.t & P_d(P_1-\theta) > c_t/(v+h) \end{array}}$; and

2. $\boxed{\begin{array}{ll} \underset{P_d}{Max} & \pi^{r} \\ s.t & P_d(P_1-\theta) \le c_t/(v+h) \end{array}}$

Since we assume that the attacker will decide to randomly attack when indifferent between the two strategies, the constraint in problem 1 is a strict inequality.

Solving the optimization problems yields the following solutions:

1. For problem 1, considering that in this optimization problem the constraint is not binding, we obtain the global maximum as $P_d^* = \left[\frac{ds\lambda}{(l+u)(1-\lambda)} * \frac{Q_f}{Q_d\theta+Q_f(1-\theta)}\right]^{\frac{s}{1-s}}$;
2. For problem 2, we obtain the following two solutions:
    a. If the constraint is not binding, we have the global maximum $P_d^{**} = \left[\frac{ds\lambda}{(l+u)(1-\lambda)}\right]^{\frac{s}{1-s}}$; and
    b. If the constraint is binding, we have $P_d^{***} = \frac{c_t}{(v+h)(P_1-\theta)} = \Omega$

We thus obtain the system defender's optimal configuration of the security detection mechanism on the systems under various conditions.

One important observation directly obtained from this result is that $P_d^* = P_d^{**}\left(\frac{Q_f}{\theta Q_d + (1-\theta)Q_f}\right)^{\frac{s}{1-s}}$. Since we assume that the detection systems used by attacker are better than random guessing, i.e., $Q_d \geq Q_f$, we have $\left(\frac{Q_f}{\theta Q_d + (1-\theta)Q_f}\right)^{\frac{s}{1-s}} \leq 1$. We denote the LHS of this inequality as Ψ. Therefore, we have $P_d^* \leq P_d^{**}$.

As a result, the defender's decision making process would be: 1) based on the parameter values, she calculates the values of P_d^*, P_d^{**}, and Ω; 2) based on the ordinal relationships among the values of P_d^*, P_d^{**}, and Ω, she determines the optimal configuration of the detection mechanism.

More specifically, the defender's decision making and the attacker's corresponding strategy can be summarized as the following Proposition.

Proposition 2: The equilibria of the defender's optimal configurations for the security mechanism's detection rate and the attacker's corresponding attacking strategy are given as the following table 2:

Table 2. The system defender's optimal configuration and attacker's respective strategy under different scenarios.

Scenario	Defender's optimal configuration	Attacker's attacking strategy
1. $P_d^* \le P_d^{**} \le \Omega$	P_d^{**}	Randomly attack
2. $P_d^* \le \Omega \le P_d^{**}$	Ω	Randomly attack
3. $\Omega < P_d^* \le P_d^{**}$	P_d^*	Target attack

This surprising result seems to be counter-intuitive: we obtain the optimal detection rate P_d^* if the attacker tries to target attack, while we obtain the optimal detection rate P_d^{**} when the attacker tries to randomly attack. Why is P_d^* lower than P_d^{**}?

The intuition behind this result can be explained as follows. When all parameter values are public knowledge, as in our case, the system defender can determine, a priori, whether the threshold has been surpassed for the attacker to detect weak systems before attacking. If that threshold has already been surpassed with the configuration of P_d^*, increasing the detection rate on the strong systems would incur more detection costs, would further encourage attacker to target attack, and would not alleviate any loss from the attacked weak systems. Thus overall the system defender may be worse off if they increase the detection rate on the security mechanisms. On the other hand, if the threshold won't be surpassed with the configuration of P_d^{**}, the attacker will randomly attack, and in this case configuring the system's detection rate at a high level of P_d^{**} would catch the attacker more effectively.

Therefore, for any given set of parameter values, the value of Ω can fall into one of the three regions as illustrated in the following figure:

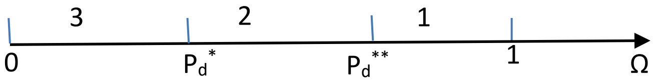

Figure 4. The different value regions for Ω

The different regions correspond to the scenario 1 – 3 as described in Proposition 2.

The "Nigerian Scammer" types of attacks described in Section 4.1 are good examples for scenario 3. As illustrated by the real world case of BBC (2016), the attacker is highly motivated to target attack. In response, the system defender would set the optimal detection rate of the security mechanism as P_d^* instead of P_d^{**}. In the above example that means the defender's best strategy would be to implement the defense mechanism that is just sufficient in identifying the potential scammers (e.g., basic information security training of the danger of online scams, and principles of handling suspicious email messages, etc.), instead of a highly accurate detection system (e.g., additional hardware/software that can automatically identify the Nigerian scammers).

A good example for scenario 1 would be the design and deployment of Honeypot technology. A Honeypot can be understood as a decoy information system that appears to be an easy target that "tends to provide the attacker with all the necessary resources needed for a successful attack" (Kuwatly, et al. 2004). System defender often builds a Honeypot network that attracts attacker's attention and lures them into interacting with the systems. The system defender can then study "the methods and tools used by the intruders" to better deter them. In this case, Honeypot systems are usually designed to be easily accessible to the attacker as weak systems, thus the attacker usually won't make extra efforts in detecting true weak systems. As a result, the system defender should configure the detection rate to be sufficiently high (P_d^{**} instead of P_d^{*}) to ensure that they are interacting with real attacker. This is exactly what has been advocated by both Honeypot researchers and industry practitioners. For example, Spitzner (2003) has pointed out that Honeypots must be designed to attract real hackers (i.e., implementing detection mechanism with a high true positive detection rate). Similarly, Antonopoulos (2010) also pointed out that in industry practices, Honeypots are often used in conjunction with other technological tools to enhance hacker detection (i.e., to achieve high true positive detection rate).

Another important insight resulted from the Proposition 2 is that algebra directly shows that $\frac{d\Psi}{dQ_f} > 0$. Therefore, we obtain the following result:

Corollary 1. The value of P_d^{*} is increasing in the value of Q_f.

Under scenario 3, the defender will configure the detection rate at P_d^{*} and the attacker will target attack. This corollary shows that if the attacker's detection mechanism is not very good (i.e., higher value for Q_f), the defender's optimal detection rate configuration should be higher. Intuitively, the higher the value is for attacker's false positive rate in his detection of weak systems, the more likely he will classify a strong system as a weak one and choose to attack. In that case, it is beneficial for the defender to configure a higher detection rate for detecting the attacker to capitalize on the attacker's mistakes.

Next, we investigate the properties of the optimal configurations of the detection systems, as well as those of the overall expected payoffs for the defender.

4.3. Sensitivity Analysis

Differentiating the system defender's optimal detection rate for attacker under scenario 3 with respect to the attacker's detection rate for weak systems, we obtain that $\frac{dP_d^*}{dQ_d} > 0$. Differentiating the system defender's optimal detection rate for attacker under scenario 3 with respect to the portion of the weak systems in the organization, we obtain that $\frac{dP_d^*}{d\theta} < 0$.

Substituting the optimal detection rate of the defender into the overall expected payoff functions for both the randomly-attack scenario and the detect-and-attack scenario, and differentiating the overall expected payoffs with respect to λ, we get that $\frac{d\pi}{d\lambda} < 0$ for both π^r and π^d. Differentiating the defender's optimal expected payoff when attacker target attacks (i.e., π^d) with respect to Q_d, we get that $\frac{d\pi^d}{dQ_d} < 0$.

We summarize the above results in the following Proposition 3:

Proposition 3:

a. When the attacker's accuracy in detecting weak systems increases, the optimal configuration of the security detection mechanism by the defender should also increase, and the overall expected payoffs to the organization decrease;
b. The optimal payoff to the organization is decreasing in the proportion of attackers among the users; and
c. The defender's optimal detection rate decreases with an increase in the proportion of weak systems in the organization.

Proposition 3 directly illustrates the impact of an attacker's strategic behavior of detecting weak systems. Part a and b indicate that if the attacker is able to increase their accuracy in detecting weak systems, the organization is worse off, and the system defender must also increase her accuracy in detecting the attacker to protect the organization. The more attackers there are in the user population, the worse off the organization is. These results conform to our intuition about the impacts of the number of attackers in the user population and their capability in identifying weak systems on the organization's overall payoffs and the system defender's optimal configurations for the security detection mechanism.

One implication from Proposition 3 that is not immediately obvious is the relationship between the optimal detection rate and the proportion of weak systems in the organization (i.e., part c of the Proposition). As the proportion of weak systems in the organization increases, it becomes more likely for the attacker to encounter weak systems. Our result shows that one approach for the system defender for handling this is to adjust the detection rate downward for the security detection mechanism to deter the attacker from targeting the (now more) weak systems. Again, the ability of attacker detecting the weak systems has significant impacts on the system defender's optimal strategy. Sometimes, the system defender would rather lower the security detection rate to deter attacker from target attacking and thereby steer the attacker in the direction of random attacks.

5. The Impact of Incomplete Information

In this section, we relax the assumption that the attacker is able to obtain the accurate value of θ, and consider the scenario under which the attacker will overestimate or underestimate the value of θ. We investigate how such mis-estimation would impact the attacker's behavior, the defender's strategy, and the organization's overall payoffs.

5.1 The Impact of Mis-estimation on the Attacker's Decision Making

For the attacker, his decision-making is based on comparing his expected payoff from randomly attacking (E_r^a) and detecting the weak systems before attacking (E_d^a). We first notice that E_r^a is linearly increasing in θ, and E_d^a is concave and increasing in θ (i.e., $\frac{dE_d^a}{d\theta} > 0, \frac{d^2E_d^a}{d\theta^2} < 0$). Algebra shows that $E_r^a > E_d^a$ if and only if $\theta > \overline{\theta}$, or $\theta < \underline{\theta}$, where $\underline{\theta} = \frac{1}{2}[1 - \frac{c_t}{P_d(h+v)} - \Delta]$, and

$$\overline{\theta} = \frac{1}{2}[1 - \frac{c_t}{P_d(h+v)} + \Delta] \text{, where } \Delta = \frac{\sqrt{(Q_d - Q_f)[Q_d(c_t - P_d(h+v))^2 - Q_f(c_t + P_d(h+v))^2}}{(Q_d - Q_f)P_d(h+v)}.$$

Therefore, for any given detection rate P_d configured by the defender, there is a corresponding pair of $\underline{\theta}$ and $\bar{\theta}$ values such that the value of θ can fall into one of the three regions (A, B, or C) as illustrated in the following Figure 5.

When the value of θ falls in region A or C (i.e., the proportion of weak systems in the target organization is either very high or very low), the attacker should choose to randomly attack. When the value of θ falls in region B (i.e., the proportion of weak systems in the target organization is moderate), he should choose to target attack.

But if the attacker mis-estimate the value of θ as $\hat{\theta}$, he may change his attacking strategy. If the value of $\hat{\theta}$ deviates from the true value of θ, but both values remain in the same region as depicted in Figure 5, the mis-estimation does not impact the attacker's decision-making or resultant payoffs. However, if this deviation is sufficient to make the two values belong to two difference regions, the mis-estimation will change attacker's decision and thus impact the attacker's payoffs. This result can be illustrated by the following Lemma 2.

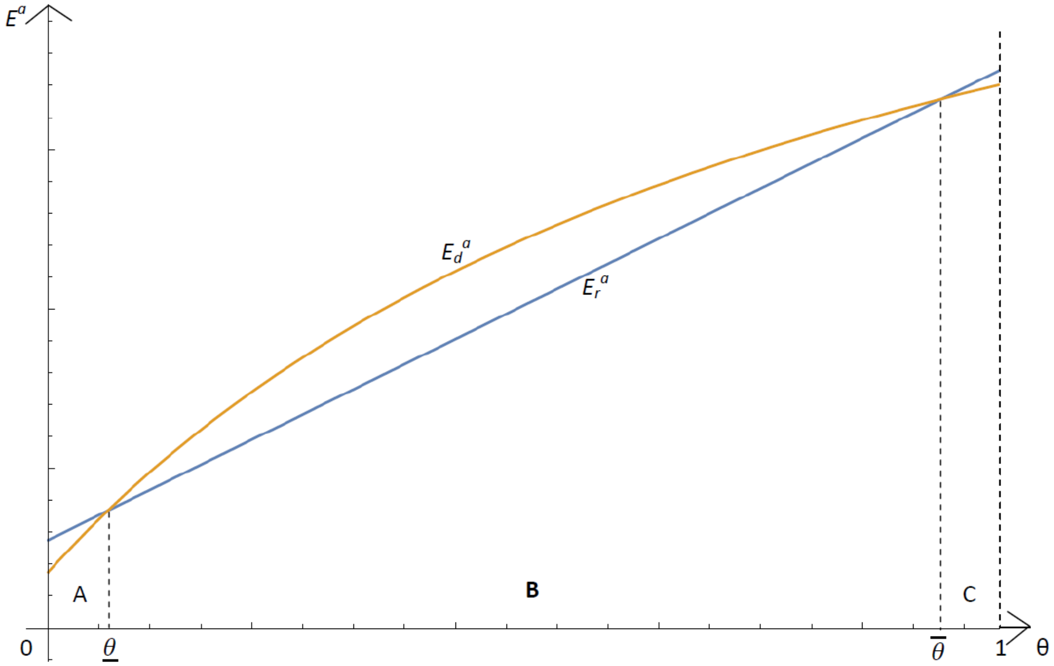

Figure 5. The different value regions for the value of θ

Lemma 2: If the value of $\hat{\theta}$ deviates from the true value of θ, the resultant attacker's decision and payoff impacts (shown in parenthesis) can be summarized as the following table 3:

Table 3. The attacker's decision (attacker payoff impacts) under different scenarios of estimation of θ

		The region the value of $\hat{\theta}$ falls into:		
		A	B	C
The region the true value of θ falls into:	A	Randomly attack (same payoffs)	Target attack (Lower payoffs)	Randomly attack (same payoffs)
	B	Randomly attack (lower payoffs)	Target attack (same payoffs)	Randomly attack (lower payoffs)
	C	Randomly attack (same payoffs)	Target attack (Lower payoffs)	Randomly attack (same payoffs)

From Lemma 2, it is noticed that mis-estimating the value of θ never makes the attacker better off. Next, we investigate the impact of such mis-estimation on the defender's strategy and organization's overall payoffs.

5.2 The Impact of Mis-estimation on the Defender Strategy and Organization Payoffs

Once the defender decides on the optimal configuration of the security mechanism's detection rate based on Proposition 2, the attacker will respond accordingly. For example, under scenario 1 and 2 as specified in Proposition 2, the defender configures the optimal detection rate as $P_d = P_d^{**}$, or $P_d = \Omega$, and the optimal strategy for the attacker will be to randomly attack, resulting in the organization's overall expected payoff being π^r. However, if the attacker misjudges the value of θ to such a degree that he decides to target attack instead of randomly attack, then the organization's overall expected payoff becomes π^d. Similarly, under scenario 3, the defender configures the optimal detection rate as $P_d = P_d^{*}$, and the optimal strategy for the attacker will be to target attack, resulting in the organization's overall expected payoff being π^d. However, if the attacker misjudges the value of θ to such a degree that he decides to randomly attack in this scenario, then the organization's overall expected payoff becomes π^r.

Comparing π^r and π^d, we obtain that for any given value of the defender's optimal detection rate (denoted as $P_d^{'}$), $\pi^r > \pi^d$ if and only if $P_d^{'} > z/d$. We denote the RHS of this inequality as Σ.

If $\Sigma < \Omega$, then there are three regions of the possible value of P_d, as illustrated in the following Figure 6a.

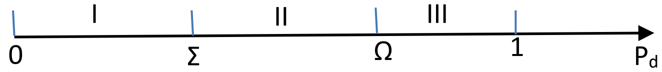

Figure 6a. The different value regions for the defender's optimal configuration of the detection rate $P_d^{'}$ when $\Sigma < \Omega$

When $P_d^{'}$ falls in region I, since $P_d^{'} < \Omega$, based on result in Equation (2), we have $E_r^a > E_d^a$. That is, the true value of θ is in Region A or C as depicted in Figure 5, and thus the attacker should randomly attack. On the other hand, when $P_d^{'}$ falls in region I, since $P_d^{'} < \Sigma$, we have $\pi^r <$

π^d. Thus the defender has incentive to encourage the attacker to target attack. The attacker would randomly attack if he accurately estimates the value of θ, and would target attack if he mis-estimates the value of θ sufficiently (i.e., the value of $\hat{\theta}$ falls in Region B while the value of θ falls in Region A or C, as depicted in Figure 5). Therefore, the defender has incentive to make the attacker mis-estimate the value of θ.

Similarly, when $P_d{}'$ falls in region II, since $P_d{}' < \Omega$, we have $E_r^a > E_d^a$. That is, the true value of θ is in Region A or C, and thus the attacker should randomly attack. On the other hand, when $P_d{}'$ falls in region II, since $P_d{}' > \Sigma$, we have $\pi^r > \pi^d$. Thus the defender has incentive to encourage the attacker to randomly attack. The attacker would randomly attack if he accurately estimates the value of θ, and would target attack if he mis-estimates the value of θ sufficiently. Therefore, the defender has incentive to make the attacker accurately estimate the value of θ.

If $P_d{}'$ falls in region III, the true value of θ is in Region B, and thus the attacker should target attack. We also have $\pi^r > \pi^d$, and thus the defender has incentive to encourage the attacker to randomly attack. The attacker would target attack if he accurately estimates the value of θ, and would randomly attack if he mis-estimates the value of θ sufficiently. Therefore, the defender has incentive to make the attacker mis-estimate the value of θ.

Similarly, if $\Sigma > \Omega$, then there are also three regions of the possible value of P_d, as illustrated in the following Figure 6b.

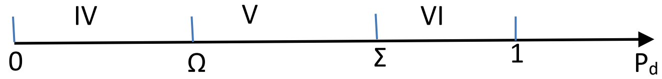

Figure 6b. The different value regions for the defender's optimal configuration of the detection rate P_d when $\Sigma > \Omega$

If $P_d{}'$ falls in region IV, the situation is the same as in region I.

If $P_d{}'$ falls in region V, the true value of θ is in Region B and thus the attacker should target attack. We also have $\pi^r < \pi^d$, and thus the defender has incentive to encourage the attacker to target attack. The attacker would target attack if he accurately estimates the value of θ, and would randomly attack if he mis-estimates the value of θ sufficiently. Therefore, the defender has incentive to make the attacker accurately estimate the value of θ.

If $P_d{}'$ falls in region VI, the situation is the same as in region III.

5.3 The Defender's Strategy in Influencing the Attacker's Estimation of θ

Obviously, if the attacker mis-estimates the value of θ, but remains within the same region as the true θ value, as illustrated by Figure 5, the mis-estimation has no impact on the defender's strategy. Only when the attacker mis-estimates the value of θ in such a way that the value of $\hat{\theta}$ moves across different regions from the true θ value will such mis-estimation have an impact on the organization's overall expected payoff. In Section 5.2 we show that such impact can be either positive or negative in terms of the organization's overall payoffs.

The above analyses in Section 5.1 and 5.2 can therefore be summarized as Proposition 4.

From the defender's perspective, based on all the parameter values, she first calculates the potential configuration values $P_d{}^*$, Ω, and $P_d{}^{**}$, and makes the optimal choice P_d based on Proposition 2. Then she calculates the threshold values Σ, $\underline{\theta}$ and $\bar{\theta}$, and chooses the correct strategy based on Proposition 4 to influence the attacker's estimation of the value of θ.

One significant observation from this result is that the attacker's mis-estimation of θ does not automatically lead to benefits for the organization. There are three scenarios, as identified in Table 4, under which the attacker's mis-estimation will result in a lower overall expected payoff to the organization, and as a consequence, the defender should endeavor such that the attacker would be able to accurately estimate the value of θ.

Proposition 4: The defender's strategies in influencing attacker's estimation of the value of θ can be summarized in the following table:

Table 4. The defender's strategies in influencing attacker's estimation of the value of θ under various scenarios

		The Region the True Value of θ Falls Into:		
		A	B	C
The Region the Optimal Detection Rate (P_d) Falls Into:	I or IV	Encourage attacker to over-estimate	N/A	Encourage attacker to under-estimate
	II	Encourage attacker to estimate accurately	N/A	Encourage attacker to estimate accurately
	III or VI	N/A	Encourage attacker to mis-estimate	N/A
	V	N/A	Encourage attacker to estimate accurately	N/A

If the defender's optimal detection rate falls in region I, the real value of θ would be either very high or very low, and the attacker will randomly attack. However, in this case, it will result in a higher overall expected payoff to the organization if the attacker target attack instead of randomly attacking. Therefore, the defender should endeavor to encourage the attacker to over-estimate θ if its real value is very low, and to encourage the attacker to under-estimate if its real value is very high, in order to nudge the attacker to target attack.

If the optimal detection rate falls in region II, the real value of θ would also be either very high or very low, and the attacker will randomly attack. In this case, the organization yields higher expected payoff if the attacker randomly attacks. So it is to the best interest of the organization that the defender endeavors such that the attacker can estimate the value of θ more accurately to ensure that the value of $\hat{\theta}$ remains in the same region of the true value of θ, and attacker will indeed randomly attack[4].

If the optimal detection rate falls in region III, the real value of θ would be moderate, and the attacker will target attack. In this case, the organization yields higher expected payoff if the attacker randomly attack. Therefore, the defender should endeavor to encourage the attacker to either over-estimate the value of θ if its true value is close to $\bar{\theta}$, or under-estimate the value of θ if its true value is close to $\underline{\theta}$. If the defender can make the value of $\hat{\theta}$ move to region A and C, it will result in a higher expected payoff to the organization.

If the optimal detection rate falls in region IV, the defender's best strategy would be similar to that when the detection rate falls in region I.

If the optimal detection rate falls in region V, the real value of θ would be moderate, and the attacker will target attack, and it will yields higher payoff to the organization. Thus the defender should endeavor such that the attacker can accurately estimate the value of θ and thus will indeed target attack.

Finally, if the optimal detection rate falls in region VI, the defender's best strategy would be similar to that when the detection rate falls in region III.

Next, we extend our model to investigate the situation where there are further heterogeneity among the normal users and the organization's systems.

6. The Impact of Heterogeneous Users and Systems

[4] Obvious, in this scenario, there is another possibility. If the true value of θ is in region A (C), defender can theoretically endeavor to make the attacker grossly over-estimate (under-estimate) the value of θ such that the value of $\hat{\theta}$ falls into region C(A). That way, it is also ensured that the attacker will indeed randomly attack.

In this section, we extend our model to analyze the situation in which there are multiple levels of normal users or systems as they may generate different levels of utility to the organization. We study how this heterogeneity impacts the defender's strategy and organization's payoffs.

6.1. The Impact of Heterogeneous Users

We now consider the situation in which different normal users generate different levels of payoffs to the organization by using the information system. We denote the overall expected payoff from a single user as E(u). Since the normal users will always randomly select a target system to access, from the defender perspective, her decision making will still be realized based on Figure 4, with the sole exception that the payoff to the organization will be E(u) instead of u, if the normal user is granted access to the system.

As a result, the analyses and analytical results from Section 4 and 5 would remain intact (except for the notation where u would be replaced by E(u).

6.2. The Impact of Heterogeneous Systems

In our original model, all the information systems in the organization are homogeneous in every aspect except that θ proportion of the systems is considered weak due to their security mechanism not functioning. But in reality, an organization's information systems may also be differentiated by their values to the organization: some systems may generate more utility to the organization than other systems when accessed by a normal user; or some may result in more severe damages if attacked than other systems. For example, an organization may have information systems to keep records of business transactions, and also systems to process business transactions. While both systems are important to the organization, it is reasonable to think that the transaction processing systems would generate more utility to the organization; it is also reasonable to think that compromised transaction processing systems would cause more severe loss than compromised record keeping systems. Such heterogeneity would affect the defender's optimal strategy to protect the systems.

In this paper, we consider two scenarios in which this heterogeneity would impact the security level of the systems. In scenario 1, we can assume that the security levels of the information systems are completely dependent on their values. In other words, the difference between strong and weak security levels among systems is the direct outcome of the difference between high and low expected payoffs (or losses) of those systems. That is, the organization allocates resources differently according to the value of the systems to the organization: systems with higher value (i.e., systems with higher payoffs/losses to the organization) will become strong systems, and systems with lower value (i.e., systems with lower payoffs/losses to the organization) will become weak systems. A huge body of literature exists that investigate this scenario (e.g., Bier at al. 2007 and Levitin et al, 2012). In scenario 2, we can assume that that the security levels of the information systems are completely independent on their values. In other words, the difference between high and low values of the systems would be another independent dimension, in addition to the difference between strong and weak security levels, in differentiating the systems. That is, among all the organization's information systems, some have strong security while others have weak security; and some have low values to the organization while others have high values. In addition, in this scenario, the security level of the systems is assumed to be independent of the value of the systems to the organization. Next we'll analyze both scenarios to study the impacts of system heterogeneity on the security game and its outcome.[5]

6.2.1. Scenario 1

In this scenario, the heterogeneity of the system values directly results in the different security level of the systems. Thus we re-define the parameter θ in our original model as follows.

We now assume among all the information systems in the organization, there is a portion of θ that would generate low utility (or cause less severe loss) to the organization if accessed by normal user (or attacker) than the other (1- θ) portion[6]. As a result of low utility, this θ portion of the information systems become S_w type and the other (1- θ) portion become S_s type. As in our original model, we assume that other than this dimension, all other aspects of the information system are identical among all systems. The attacker still decides on whether to randomly attack or target attack the weak system (i.e., the low utility system), and the defender still optimizes the detection rate to maximize the organization's overall expected payoffs.

For the attacker, since his payoff parameters, as well as the probability parameters, do not change, the realization of his decision making would remain the same as was illustrated in Figure 2. All our analyses and results presented in Section 4.1 remain unchanged. Similarly, for the defender, the realization of her decision making will be the same as illustrated in Figure 3. All our analyses in Section 4.2 to 5 remain unchanged.

[5] An interesting future research avenue would be to consider the scenario where the security levels of the information systems are partially dependent on their values.

[6] Here we make the implicit assumption that the systems that generate higher payoffs if accessed by normal user would also cause worse damages if attacked, though no further assumptions are made about the nature of this correlation.

One significant insight can be obtained from these results by re-interpreting Proposition 4 within this new scenario. Proposition 4 illustrates the defender's strategies in influencing the attacker's estimation of the value of θ in order to optimize the organization's overall expected payoffs. It shows that if the defender's optimal detection rate is configured to be within region I, IV, or V, she has incentive to encourage the attacker to target attack (by either disseminating true information under region V or spreading false information under region I or IV). Based on our scenario description, this essentially implies that under certain conditions (e.g., the optimal detection rate falls into region I, IV, or V), it is optimal for the defender to deflect the attacker to the low value systems. This insight directly illustrates the relationships between system security level as a direct result from system value, and the defender's optimal strategy.

6.2.2. Scenario 2

In this scenario, the system value is an independent dimension, in addition to the security level, of the organization's information systems. With all parameters of our original model intact, now we add more parameters that capture the heterogeneity of the system values, and its impacts on the organization.

We now assume among all the information systems in the organization, there is a portion of γ that would generate higher utility (cause more severe loss) to the organization if accessed by normal user (attacker) than the other (1-γ) portion. We denote this γ portion of the information systems as HI type, and the other (1-γ) portion of the information systems as LO type. We also assume $\gamma \perp \theta$. That is, the value of the systems is independent of the security level of the system. Furthermore, we assume a HI system accessed by normal user will generate a utility denoted as u_h to the organization, and LO system accessed by normal user will generate a utility denoted as u_l, with $u_h > u_l$. Therefore, when a normal user is granted access to a system, the organization obtains an expected payoff of $\gamma u_h+(1-\gamma)u_l$. We denote this value as u^*. Similarly, it is assumed that an attacker accessing a HI system would cause damage d_h to the organization and an attacker accessing a LO system would cause damage d_l, with $d_h > d_l$. Thus the organization's overall expected damage when an attacker is granted access to a system is γd_h + $(1-\gamma)d_l$. We denote this value as d^*. Finally, we assume that other than these two dimensions, all other aspects of the information system are identical among all systems.

For the attacker, again the realization of his decision making as well as the analyses and results presented in Section 4.1 remain unchanged. For the defender, the realization of her decision making will be slightly different from what is illustrated in Figure 3: the payoffs to the organization change. When the user/attacker accesses a system, he will be facing either a HI system or a LO system with probability of either γ or 1-γ, respectively, which results in an expected payoff/damage of either u* or d*, respectively, to the organization. The new realization of the defender's decision making is illustrated in the following Figure 7.

From Figure 7, we solve for the defender's optimal configuration of the detection system, and obtain $P_d^* = \left[\frac{d^*s\lambda}{(l+u^*)(1-\lambda)} * \frac{Q_f}{Q_d\theta+Q_f(1-\theta)}\right]^{\frac{s}{1-s}}$, $P_d^{**} = \left[\frac{d^*s\lambda}{(l+u^*)(1-\lambda)}\right]^{\frac{s}{1-s}}$, and $P_d^{***} = \frac{c_t}{(v+h)(P_1-\theta)} = \Omega$. All analyses and results in Section 4.2 and 4.3 remain qualitatively similar.

To illustrate the impact of system heterogeneity on defender's optimal strategy, we obtain the following Lemma 3:

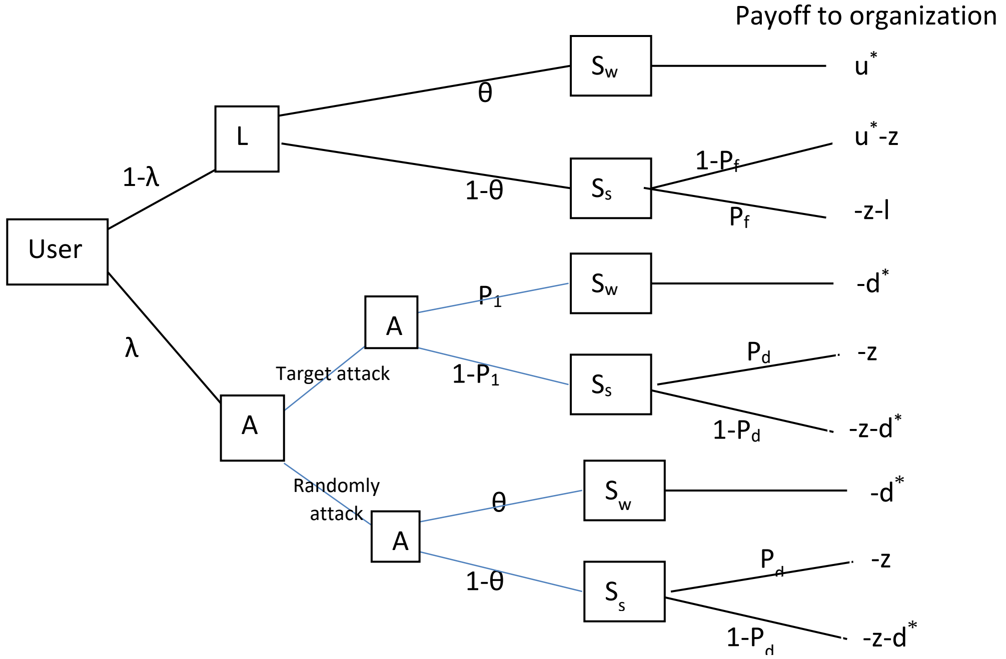

Figure 7. The possible realization for system defender's decision making when the systems are heterogenous

Lemma 3: As the value of γ increases, the values of P_d^* and P_d^{**} will also increase

if and only if $\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$ (6);

and will decrease otherwise.

Proof: we differentiate P_d^* and P_d^{**} w.r.t. γ, and solve for the inequality of $\frac{d\mathrm{P}_\mathrm{d}^*}{d\gamma} > 0$, and $\frac{d\mathrm{P}_\mathrm{d}^*}{d\gamma} < 0$, as well as $\frac{d\mathrm{P}_\mathrm{d}^{**}}{d\gamma} > 0$, and $\frac{d\mathrm{P}_\mathrm{d}^{**}}{d\gamma} < 0$, and obtain the above result.

▪

Inequality (6) ($\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$) illustrates the characteristics of the system heterogeneity. The LHS denotes the ratio of the potential damages between the two different types of systems and the RHS denotes the ratio of the potential benefits (plus the opportunity cost) between them. Lemma 3 indicates that when the heterogeneity of the two types of systems is captured more by their potential loss if attacked, the optimal detection rate for the defender would increase as the portion of the HI system increases; when the heterogeneity of the two types of the systems is captured more by their potential payoffs to the organization, the optimal detection rate for the defender would decrease as the portion of the HI system increases. The practical implication of this result is straightforward: if damage from a security breach is the more salient focus for the organization, with more HI systems in the organization, the defender should tighten the detection of the attacker so as to protect the systems from getting accessed by attacker; if payoff from normal access is the more salient focus, with more HI systems in the organization, the defender should loosen the detection of the attacker so as to ensure access to the systems from the normal users.

Next, we investigate how this additional dimension of system heterogeneity impacts defender's strategy in deflecting attacker to the weak systems as illustrated in Proposition 4.

The defender's optimal configuration of the detection rate P_d is a function of γ. We can denote this function as $P_d(\gamma)$. We can further solve $P_d(\gamma)=\Omega$ for γ, and denote the result as γ'. Lemma 3 shows when $\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$, $P_d(\gamma)$ increases in γ. Lemma 1 shows that the attacker will target attack instead of randomly attack if and only if $P_d(\gamma) > \Omega$. Therefore, when $\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$, there are two cases regarding how the value of γ impacts the defender's strategy to deflect the attacker.

Case 1: if $\gamma > \gamma'$, the attacker will target attack (i.e., the value of θ falls in region B). Following our analyses in Section 5, for the defender to want the attacker to target attack in this case, the value of P_d must fall into region V. As a result, when $\gamma > \gamma'$, the defender would want to deflect the attacker to the weak system if and only if the value of P_d falls into region V.

From Proposition 4, we know that under these circumstances the defender should encourage attacker to accurately estimate the value of θ.

Case 2: if $\gamma < \gamma'$, the attacker will randomly attack (i.e., the value of θ falls in region A or C). This implies that in order for the defender to want the attacker to target attack in this case, the value of P_d must fall into region I or IV. As a result, when $\gamma < \gamma'$, the defender would want to deflect the attacker to the weak system if and only if the value of P_d falls into region I or IV. She should endeavor to encourage attacker to over-estimate the value of θ if it is in region A, and under-estimate if it is in region C.

When $\frac{d_h}{d_l} < \frac{l+u_h}{l+u_l}$, $P_d(\gamma)$ decreases in γ. If $\gamma < \gamma'$, we encounter Case 1 described above, and if $\gamma > \gamma'$, we encounter Case 2 described above.

The above analyses can be summarized as the following Proposition.

Proposition 5: when the systems are heterogeneous in their values to the organization, the conditions under which defender would want to deflect the attacker to the weak systems under various combinations of the systems values, as well as the defender's strategy to influence the attack's estimation (in parenthesis), can be summarized in the following table:

Table 5. The conditions under which defender would want to deflect attacker to weak systems (and the defender's strategy to influence the attacker's estimation of the value θ)

	If $\frac{d_h}{d_l} < \frac{l+u_h}{l+u_l}$	If $\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$
If $\gamma > \gamma'$	Value of P_d falls into region I or IV (encourage attacker to mis-estimate)	Value of P_d falls into region V (encourage attacker to accurately estimate)
If $\gamma < \gamma'$	Value of P_d falls into region V (encourage attacker to accurately estimate)	Value of P_d falls into region I or IV (encourage attacker to mis-estimate)

Proposition 5 shows the impact of system value heterogeneity on the defender's strategy of deflecting attacker to the weak systems. When the heterogeneity of the two types of the systems is captured more by their potential loss if attacked ($\frac{d_h}{d_l} > \frac{l+u_h}{l+u_l}$), if the HI systems are of large quantity ($\gamma > \gamma'$), the defender's optimal detection rate will be relatively high (falls into region V instead of region I or IV), and she would endeavor to help attacker accurately estimate the value of θ in order to encourage him target attack the weak systems; if the LO systems are of larger quantity ($\gamma < \gamma'$), the defender's optimal detection rate will be relatively low (falls into region I or IV instead of region V), and she would endeavor to help attacker mis-estimate the value of θ in order to nudge him into target attacking the weak systems. On the other hand, when the heterogeneity of the two type of the systems is captured more by their potential payoffs to the organization ($\frac{d_h}{d_l} < \frac{l+u_h}{l+u_l}$), the defender's strategy to deflect attacker to the weak systems will reverse with respect to the relative quantity of HI or LO systems.

7. Conclusions

In this paper, we investigate the phenomenon of false positives from the attacker's perspective in a typical attacker-defender information security game. False positives in detection have long been identified as a critical factor in security scenarios and have been extensively studied. However, in almost all of the extant literature, false positives have been studied exclusively from the system defender's perspective, i.e., the existing literature focuses primarily on the research question of what happens if the system defender mis-identifies potential attacker.

To the best of our knowledge, we present the first comprehensive analytical model that incorporates the false positives both from the perspective of attacker, as well as that of system defender. We attempt to answer the research question of what are the consequences of attacker mis-identifying potential targets for attacking. We model the attacker-defender interactions, taking into consideration the attacker's possible option of detecting weak systems before implementing attacks. We analyze the impacts of attacker's false positives on the decision making of both the attacker and the system defender. Our results show that the false positives from the attacker's perspective have a significant impact on the attacker's decision making process for an attack, as well as the optimal protection strategy of the defender. We present analytical results for the defender's optimal strategy when facing strategic attacker who may detect weak systems before attacking, and provide some insights on how a system defender should be more stringent or less rigid in detecting attacker under different scenarios. Our results also help to shed new light on some common and interesting phenomena in the information security landscape such as the prevalence of spam emails, the outrageous wordings of Nigerian scams, and the design of Honeypot as a security mechanism.

An interesting finding of our analysis is that when attacker mis-estimates the proportion of the weak systems in the target organization, they are never better off; however, from the defender's perspective, the attacker's mis-estimation does not always benefit the defender. We show several scenarios under which the defender should endeavor to drive the attacker toward accurate estimation.

Our work also sheds light on the impact of system value heterogeneity on the defender's strategy of deflecting potential attacker to the weak systems. Our findings suggest that the defender has a fairly complex combination of optimal strategies contingent on a critical ratio of parameters that capture damage sustained when an attacker attacks the systems, utility gained when a normal user accesses the systems and the loss of goodwill on account of normal user being mis-identified as an attacker. These strategies again suggest, under certain conditions, an endeavor to influence the attacker in his estimation of the proportion of weak systems.

There are several interesting directions in which this study can be extended. First of all, we present a parsimonious analytical model to investigate the attacker and defender behaviors. An important extension would be an empirical study that examines the pragmatic validity of our model and analytical results. Secondly, in our model, the detection rate of attacker's classification mechanism is an exogenous parameter, and the attacker's decision making is focused exclusively on choosing between randomly attacking and detecting before attacking. It would be interesting to extend our model to endogenize the detection rate of attacker's classification mechanism as attacker's decision variable, and study how it would impact the defender's optimal strategy. Also, we model the heterogeneity of the target information systems in a binary discrete setting: weak vs. strong systems. It would be a worthy effort to study a continuous setting in which the security strength of the systems may follow certain distributions. Finally, while we show that mis-estimating the proportion of weak systems never makes the attacker better off, it would be interesting to find if this mis-estimation is symmetric or asymmetric in terms of the potential loss in payout to the attacker when the attacker overestimates versus underestimates.

References:

Alagoz O, Chhatwal J, Burnside ES (2013) Optimal policies for reducing unnecessary follow-up mammography exams in breast cancer diagnosis. *Decision Analysis* 10(3):200-224.

Antonopoulos A (2010) Honeypots for hacker detection. *Network World*, July 6, 2010, http://www.networkworld.com/article/2213251/network-security/honeypots-for-hacker-detection.html, accessed on December 21, 2016.

Bakır NO (2008) A decision tree model for evaluating countermeasures to secure cargo at United States southwestern ports of entry. *Decision Analysis* 5(4):230-248.

BBC (2016) Online fraud: Top Nigerian scammer arrested. August 1, 2016, http://www.bbc.com/news/world-africa-36939751, accessed on December 21, 2016.

Bier V, Oliveros S, Samuelson L (2007) Choosing what to protect: strategic defensive allocation against an unknown attacker. *Journal Of Public Economic Theory* 9(4):563-587.

Blum DM, Paté-Cornell E (2016) Probabilistic warnings in national security crises: Pearl Harbor revisited. *Decision Analysis* 13(1):1-25.

Brandt P, Sandler T (2010) What do transnational terrorists target? Has it changed? Are we safer? *The Journal of Conflict Resolution* 54(2):214-236.

Bulgurcu B, Cavusoglu H, Benbasat I (2010) Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly* 34(3): 523-548.

Cavusoglu H, Kwark Y, Mai B, Raghunathan S (2013) Passenger profiling and screening for aviation security in the presence of strategic attackers. *Decision Analysis* 10(1):63-81.

Cavusoglu H, Mishra B, Raghunathan S (2005) The value of intrusion detection systems in information technology security architecture. *Information Systems Research* 16(1):28 – 46.

Cavusoglu H, Raghunathan S (2004) Configuration of detection software: A comparison of decision and game theory approaches. *Decision Analysis* 1(3):131-148.

Cavusoglu H, Raghunathan S, Cavusoglu H (2009) Configuration of and interaction between information security technologies: the case of firewalls and intrusion detection systems. *Information Systems Research*, 20(2):198 – 217.

Cavusoglu H, Raghunathan S, Yue W (2008) Decision-theoretic and game-theoretic approaches to it security investment. *Journal of Management Information Systems* 25(2):281-304.

Chivers H, Clark J, Nobles P, Shaikh S, Chen H (2013) Knowing who to watch: Identifying attackers whose actions are hidden within false alarms and background noise. *Information Systems Frontier* 15:17 – 34.

Duan Z, Chen P, Dong Y, Stephenson M, Barker J (2012) Detecting spam zombies by monitoring outgoing messages. *IEEE Transactions on Dependable and Secure Computing* 9(2):198 – 210.

Gao X, Zhong W (2015) Information security investment for competitive firms with hacker behavior and security requirements. *Annals of Operation Research* 235:277-300.

Guan P, He M, Zhuang J, Hora SC (2017) Modeling a multitarget attacker–defender game with budget constraints. *Decision Analysis* 14(2):87-107.

Hannah D, Robertson K (2015) Why and how do employees break and bend confidential information protection rules? *Journal of Management Studies* 52(3).

Hausken K (2016) A cost-benefit analysis of terrorist attacks. *Defense and Peace Economics DOI: 10.1080/10242694.2016.1158440*.

Hausken K (2017) Security investment, hacking, and information sharing between firms and between hackers. *Games* 8(2):23.

Herley C (2012) Why do Nigerian scammers say they are from Nigeria? *Proceedings of the Workshop on the Economics of Information Security*, Berlin, Germany.

Herley C (2013) Small world: Collisions among attackers in a finite population. *Proceedings o f the 12th Workshop on the Economics of Information Security*, Washington D.C.

Kuwatly I, Sraj M, and Al Masri Z (2004)A dynamic Honeypot design for intrusion detection. *Proceedings of IEEE/ACS International Conference on Pervasive Services*.

Liang H, Xue Y, Wu L (2013) Ensuring employees' IT compliance: Carrot or stick? *Information Systems Research* 24(2):279 – 294.

Levitin G, Hausken K, Taboada H, Coit D (2012) Data survivability vs. security in information systems. *Reliability Engineering & System Safety* 100:19-27.

Ni KS, Faissol D, Edmunds T, Wheeler R (2013) Exploitation of ambiguous cues to infer terrorist activity. *Decision Analysis* 10(1):42-62.

Persico N, Todd P (2005) Passenger profiling, imperfect screening, and airport security. *The American Economic Review* 95(2):127-131.

Rezk A, Ali H, El-Mikkawy M, Barakat S (2011) Minimize the false positive rate in a database intrusion detection system. *International Journal of Computer Science & Information Technology* 3(5).

Salman M, Budiardjo B, Ramli K (2014) Measuring accuracy in identifying and detecting unauthorized access point using proactive intrusion detection approach in wireless networks. *International Journal of Computer Science and Network Security* 14(6).

Siponen M, Mahmood M, Pahnila S (2014) Employees' adherence to information security policies: An exploratory field study. *Information & Management* 51:217 – 224.

Solms B, Solms R (2004) The 10 deadly sins of information security management. *Computers & Security* 23:371-376.

Son J (2011) Out of fear or desire? Toward a better understanding of employees' motivation to follow IS security policies. *Information & Management* 48:296 – 302.

Spitzner L (2003) Honeypots: Tracking Hackers, Boston, MA: Pearson Education.

Stahl S, Pease K (2011) Seven requirements for successfully implementing information security policies and standards. *Citadel Information Group White Paper*.

Trees H, Bell K, Tian Z (2013) Detection, Estimation and Modulation Theory—Part I, Hoboken: New Jersey, John Wiley & Sons.

Ulvila JW, Gaffney Jr. JE (2004) A decision analysis method for evaluating computer intrusion detection systems. *Decision Analysis* 1(1):35-50.

van Binsbergen JH, Marx LM (2007) Exploring relations between decision analysis and game theory. *Decision Analysis* 4(1):32-40.

Willis HH, Moore M (2014) Improving the value of analysis for biosurveillance. *Decision Analysis* 11(1):63-81.

Zhuang J, Bier V (2007) Balancing terrorism and natural disasters - Defensive strategy with endogenous attacker effort. *Operations Research* 55(5):976 – 991.