When the Winning Move is Not to Play: Games of Deterrence in Cyber Security

Chad Heitzenrater[1,2], Greg Taylor[3], and Andrew Simpson[2]

[1] U.S. Air Force Research Laboratory Information Directorate
525 Brooks Road, Rome NY 13441
[2] Department of Computer Science, University of Oxford
Wolfson Building, Parks Road, Oxford OX1 3QD, UK
[3] Oxford Internet Institute, University of Oxford
1 St. Giles, Oxford OX1 3JS, UK

Abstract. We often hear of measures that promote traditional security concepts such as 'defence in depth' or 'compartmentalisation'. One aspect that has been largely ignored in computer security is that of 'deterrence'. This may be due to difficulties in applying common notions of strategic deterrence, such as attribution — resulting in previous work focusing on the role that deterrence plays in large-scale cyberwar or other esoteric possibilities. In this paper, we focus on the operational and tactical roles of deterrence in providing everyday security for individuals. As such, the challenge changes: from one of attribution to one of understanding the role of attacker beliefs and the constraints on attackers and defenders. To this end, we demonstrate the role deterrence can play as part of the security of individuals against the low-focus, low-skill attacks that pervade the Internet. Using commonly encountered problems of spam email and the security of wireless networks as examples, we demonstrate how different notions of deterrence can complement well-developed models of defence, as well as provide insights into how individuals can overcome conflicting security advice. We use dynamic games of incomplete information, in the form of screening and signalling games, as models of users employing deterrence. We find multiple equilibria that demonstrate aspects of deterrence within specific bounds of utility, and show that there are scenarios where the employment of deterrence changes the game such that the attacker is led to conclude that the best move is not to play.

1 Introduction

When seeking advice on computer security, any combination of the terms 'computer' and 'security' will produce myriad results from academics, businesses looking to sell products, governments at local and national levels, 'hackers' (of the black- and white-hatted varieties), and bloggers; these results are often then moderated by input from friends, family and colleagues. From this conflicting guidance emerge the choices and decisions made by individuals. This can be

Approved for Public Release; Distribution Unlimited : 88ABW-2015-1336 20150323

further complicated by a lack of knowledge or evidence of utility, as some topics are still a matter of active discussion among even the most knowledgeable of practitioners.

Recent years have seen the emergence of security economics, which seeks to augment such discussions with the insight that these failures are often not the result of engineering challenges, but of economic challenges: misaligned incentives, information asymmetries, and externalities [15]. Given this landscape, what can be asked (and expected) of those who lack a technical background, technical staff, and a security budget? This is the question posed by many small businesses and home users, who often must make security decisions based upon their limited resources (with respect to time and money) and their ability to search related terms, digesting the information that appears in (at best) the first few hits. The answer is important, as it is precisely these decisions that affect us all: we all deal with the results of these failures [15].

In examining the source of much of our modern concept of cyber security — the doctrine of the military, an entity whose primary role is security — we see that the concepts that lead to security are well-defined, but multi-faceted. With respect to current research and practice, many concepts have been widely adopted as paradigms for cyber security [20]: "defence in depth", "compartmentalisation", etc. However, one aspect of security that has been largely ignored (outside of military doctrine) is the "first line of defence": deterrence [2, 14]. In examining the role deterrence might play for individuals, we move towards a principled discussion of deterrence through the lens of information security economics. We conclude that, for a set of adversaries that can be defined in the economic context of utility, deterrence as an aspect of a comprehensive security stance is rational, contributory, and quantifiable against specific actor groups.

Section 2 introduces the various concepts at play: the notion of deterrence, the problems of signalling and screening in economics, and conceptual scenarios that are employed to provide context. Section 3 presents two concepts of deterrence as information asymmetries, formed as games of imperfect information. Section 4 provides a discussion of related work, placing this contribution within the broader context of deterrence and security economics. Finally, Section 5 summarises the contribution of this paper.

2 Background

2.1 Concepts of Deterrence

The concept of deterrence has a long history in our collective consciousness, primarily confined to our notions of national security. Throughout the Cold War, our collective security relied on a deterrence strategy of Mutually Assured Destruction (MAD) [14], with much being written on the topic of strategic deterrence: even our definition of the word is linked to this notion, including qualifiers such as "especially deterring a nuclear attack by the capacity or threat of retaliating" [7]. This emphasis on the threat of retaliation would seem to be an

unnecessary deviation in concept from the simple act of *deterring*, where *to deter* is "to discourage or restrain from acting or proceeding" or "to prevent; check; arrest" [6]. We will refer to 'deterrence' in the context of deterring attacks, using the more general notion without emphasis on retaliation (requiring attribution) that is embodied in the former definition. One may argue that this is where the concept of deterrence in cyberspace has been stymied, as attribution is a known hard problem [14]. In decoupling attribution from deterrence, we examine the latter in a sense not possible when the concepts are intertwined.

In conjunction with this line of thought is the movement from concepts of strategic deterrence towards deterrence that results from more commonplace interactions: the deterrence that leads to the everyday security of individuals. In this spirit, Morral and Jackson [16] consider deterrence at the strategic, operational and tactical levels. In [16], strategic deterrence is defined by reducing the net expected utility of a particular means (e.g. attack) for a group to achieve their objective against another group. This is differentiated from operational deterrence by an emphasis on specific operations (or classes of operations), ideally leading to abandonment for that particular operation. Tactical deterrence then refers to the alteration of net utility after the attack is initiated. These definitions map nicely to current concepts of cyber security: strategically deterring attacks against internet users via the installation of various technical protections and procedures; operationally deterring against malware via the use of antivirus; and tactically thwarting the exfiltration of information from a machine via the use of a firewall. This also highlights the obvious links between deterrence and broader security, in that being secure can be a deterrent itself. Morral and Jackson [16] offer interesting insights regarding the nature and role of deterrence in these contexts, with relevance to information technologies and cyber deterrence. One point involves the role of complexity; all else being equal (regarding the utility of the target to the attacker, or other factors such as accessibility), a more complex attack is less appealing to an attacker. The resulting increase in complexity gives rise to an increase in observable signatures, resources expended, etc. — all of which lead to a less attractive target. This is tempered with the caveat that the deterrence cannot be trivial to overcome, no matter the likelihood of engagement by the attacker.

Bunn [4] considers the distinction between deterrence and dissuasion: dissuasion is related to the aim to convince a potential adversary from engaging at all. Using the above example of deterrence measures, dissuasion would be akin to laws against malware-writing and campaigns to warn potential attackers of computer misuse. Relevant to this discussion is the distinction between something that is more closely related to the psychological with respect to dissuasion, against measures that may have a more distinct technical aspect of deterrence. Bunn additionally contributes the notion that one deters *someone* from doing *something*, implying that actors and actions are of importance when considering deterrence. This leads one to conclude that deterrence will manifest itself differently given different scenarios; this is a central tenet of our contribution.

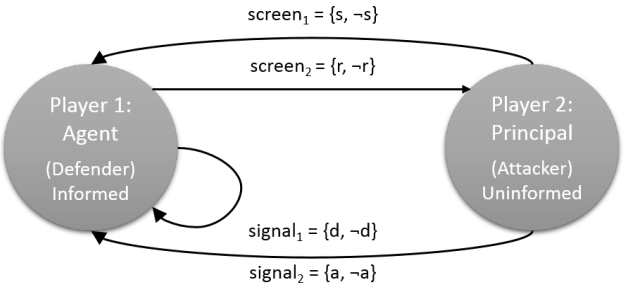

Fig. 1. The Agent–Principal model of deterrence enacted through screening and signalling. In screening, the principal moves first via a screening action $\{s, \neg s\}$ in an attempt to classify the agent's type (e.g. a viable or non-viable target). The agent may choose to respond $\{r, \neg r\}$, potentially betraying their type. In signalling, the agent moves first, broadly signalling $\{d, \neg d\}$, which may or may not be indicative of their type. The principal, observing this signal, chooses to react $\{a, \neg a\}$. The arrows differentiate a directed action by the defender to the attacker (in screening), and the broader action visible to all parties, including other defenders and non-players (in signalling).

2.2 Information Asymmetries in Security

One increasingly popular view of security is that of an information asymmetry between two entities: a user, who has the ability to take some action to secure themself against attack, and therefore has information regarding the state of security; and an attacker, who seeks to identify targets for exploitation, but lacks information regarding the security of any given user. Information asymmetries arise when two entities engage in a transaction with each having access to different levels of information; they are a common source of *market failures* [21], which arise when inefficiencies in markets lead to suboptimal operation, such as one side gaining a distinct advantage in the marketplace. In our construct, the market for security is represented by this interaction between the attacker and the user; this differs from other characterisations that focus on the information asymmetry between users and security products, e.g. [3].

As with other forms of security, we can formally describe deterrence in terms of information asymmetry. We define this market as having an agent and a principal where the user (as the agent) has more information regarding their security level than the attacker (the principal). In this case, the information that is asymmetric is the type of the user, who might (through various actions undertaken prior to this point) be of type 'secure' (t_s), or type 'unsecure' (t_u).

Short of resolution through regulation (a factor for computer security, but something that thus far failed to resolve this market), there are two primary means of dealing with information asymmetries [21]: screening and signalling. Figure 1 depicts these concepts as sets of moves between agents and principals. We consider each in turn.

Screening involves the principal moving first to resolve the asymmetry via an action that (potentially) prompts a response from the agent. The goal of the principal is to separate the agents into groups depending on their type (in this case, secure and unsecure users). Examples of such actions include pings to determine the reachability of devices and services on the network, or operating system fingerprinting using tools such as NMAP.[1] Note that this is not an 'attack' as such, and is perhaps best considered reconnaissance — movement by the principal to gather more information (e.g. reachability of IPs, or patch level of operating systems). However, the results of the screening could certainly be employed in, and contribute to, an attack. In the following, we will use email phishing scams to discuss deterrence in this light, as an example of tactical deterrence of an 'attack' in progress.

Signalling involves the agent moving first via an observable action, prompting the principal to make a decision as to their type and react accordingly. The agent may be honest or dishonest regarding their type, forcing the principal to react based on belief. A good example from cyber security is the bug bounty offered by software providers to indicate the security of their systems. Here, poor software would not be able to offer such a bounty lest the software provider go bankrupt. Therefore the existence of such a scheme both signals to consumers that the software is of high quality, and increases that quality through the awards that are made — which, in turn, prompts further bugs to be found. We will look at the role of 'weak' security mechanisms, such as SSID hiding and MAC filtering, as signals of security that have an operational deterrence effect.

2.3 Adversary Scenarios

Having established the notion of deterrence as an information asymmetry, we now construct two adversarial scenarios corresponding to our concepts of operational and tactical deterrence. As a starting point, we consider attackers as falling on a spectrum, as postulated by Schneier [18]. Schneier characterises attackers along two axes: focus (interest in a specific victim) and skill (technical ability, such as use of existing scripts/tools vice development). Schneier maintains that the majority of attacks are "low-skill and low-focus — people using common hacking tools against thousands of networks world-wide" [18]. It is precisely these kinds of attacks on which we will focus our attention.

In the first scenario we consider a phishing scam, where the attacker (as the principal) moves first. The user, as the agent, drives the beliefs of the attacker through their response (or lack thereof). This is depicted in the upper part of Figure 1. Using the construct of [12], we frame the scenario as an attacker sending spam emails to unwitting users in order to examine tactical deterrence. In [12], Herley conjectures that attackers who profit from attacks that depend on economies of scale (such as the infamous Nigerian scams) face the same economic and technological challenges as other disciplines. In constructing the

[1] The "Network Mapper". See `http://nmap.org/` for a discussion on using NMAP for operating system fingerprinting.

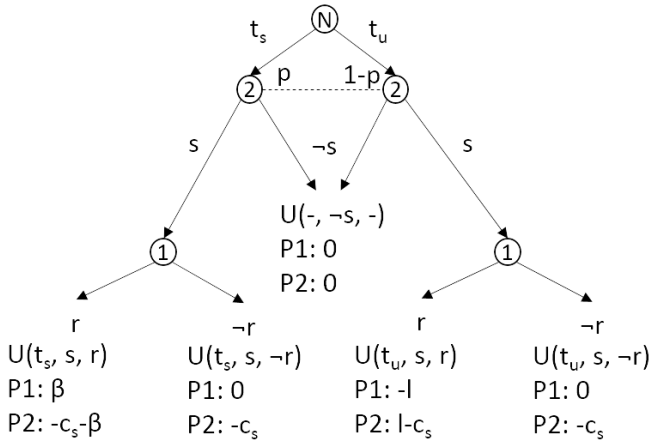

Fig. 2. Extensive form of the deterrence screening game.

scam, attackers must overcome statistical problems such as thresholding and binary classification when selecting victims, and therefore must weigh various aspects in order to make the attack profitable. Herley shows how success for an attacker depends on the density of viable users, d, as a fraction of viable victims M within a population N, $d = M/N$. With each attack costing the attacker C and yielding a net profit of G, it is obvious that, as the density d is small, it is important for the attacker that C is kept low and that G is maximised. To this extent, the attacker must use some criterion to select those to attack, which Herley terms 'viability'. Therefore, in order to identify d, the attacker utilises a 'viability score' x to separate users into a class $\{viable, non\text{-}viable\}$. Herley provides two insights regarding the role of beliefs in such attacks that has implications to deterrence. First, binary classification of users is as much a problem for attack as for defence. Thus, as the attacker's ability to separate viable from non-viable targets decreases, the effect on the true positive rate t_p versus the effect on false positive rate f_p can lead to dramatic shifts in the action of the attacker. Second, optimism on the part of the attacker does not pay, as over-estimation can quickly lead to unprofitability due to the non-zero cost of carrying out the attack. Thus, it is to the attacker's benefit to reduce costs and to be conservative in the choice of thresholding x, which drives both t_p and f_p.

The second scenario covers operational deterrence, and uses the example of an attacker attempting to undermine wireless connections. This network could be the responsibility of a small business proprietor utilising wireless connectivity for their business network, or a home user in a densely occupied space such as an apartment building in a large city. The key to this scenario is that the proprietor or user, acting as the agent, has a wireless network which they seek to secure from eavesdropping and unauthorised use by an attacker, acting as the principal. The security level of the user ('secure' or 'insecure') will serve

to distinguish types of users, corresponding to the user having taken steps to protect against attacks against information disclosure or unauthorised use (e.g. having enabled WPA2 security). In this case, the attacker is assumed to be capable of employing 'standard' measures against the network — attempt to connect to the network, sniff and read message traffic (with proximity to the network), and potentially manipulate and retransmit any packets transmitted in the clear. The attacker is not assumed to be capable of breaking the WPA2 key, although the goal of the user in this context will be to deter the attacker from attempting such an attack in the first place (perhaps due to the user not using a sufficiently secure password, or wanting to minimise the log of failed attempts). As such, the user will seek to employ methods that are widely cited as recommended practices despite being 'weak' security — SSID suppression and MAC filtering — as 'signals' of security to dissuade attacks. We will demonstrate how modelling this scenario as a signalling game indicates that such methods have utility in this context. This is depicted in the lower half of Figure 1.

3 Deterrence as an Information Asymmetry

3.1 Deterrence as Screening: An Example of Tactical Deterrence

We first look at the concept of tactical deterrence (deterrence of an 'attack' that is underway) though the lens of a screening game, using Herley's construct of the Nigerian scammer [12]. The game as conceived is depicted in Figure 2, and unfolds as follows.

1. Player 'Nature' moves, allocating the distribution of the types of users t_s and t_u. We assume a distribution of Player 1 types $(p, 1-p)$ but that neither player observes the realisation of this random variable, as this is reliant on the nature of the scam.
2. Player 2 (the attacker/spammer) makes the first move, not knowing the type of Player 1 (a given victim/user). Player 2 chooses to initiate a screening action s at a cost c_s (the spamming email), or chooses not to engage ($\neg s$) and thus incurs no cost. This is done according to a belief p that Player 2 holds regarding the type of Player 1. It is assumed in this case that c_s is relatively small, but this is not necessarily the case in other scenarios.
3. Player 1's recognition of the scam then dictates their type, as either type secure (t_s) or of type unsecure (t_u). As a result, Player 1 may choose to respond or not to respond to the screening action and this choice may or may not be indicative of their type. Choosing not to respond has no loss or gain — a payoff of 0. Note that, following this exchange, Player 1's type is inferred by both players and the game unfolds similar to that of a game of complete information.
4. For simplicity in this game, the payoff for Player 2 is modelled as capturing l (Player 1's 'loss') upon successfully generating a response from an unsecure user, while Player 1 incurs the same loss $-l$ if unsecure and responding to the scam. Alternatively, a secure Player 1 exacts a benefit β from the scammer

	s	$\neg s$
r	$(p(\beta+l)-l, l-p(\beta+l)-c_s)$	$(0,0)$
$\neg r$	$(0,-c_s)$	$(0,0)$

Table 1. Ex ante expected payoffs for the deterrence screening game of Figure 2.

(the consumption of attacker resources that are not employed elsewhere; a 'social benefit'), while Player 2 loses that benefit along with the plus cost of screen $-c_s - \beta$ if prompting a response that does not result in a payoff (due to missing out on the potential profit from another user).

We make the simplifying assumption that, in getting a user type t_u to respond, the ruse is played out and the attacker captures l. As such, we are not considering instances in which an unsecure user engages but the transaction is thwarted (they instead appear as t_s users, with $\beta < l$).

The payoffs for a distribution of players p are provided in Table 1. We see that the strategy for Player 2 hinges on the value of p: if Player 2 believes Player 1 to be of type t_u ($p = 0$), it is beneficial for Player 2 to attempt the game as long as $l \gg c_s$ (as presumably would be the case for a spam email). In fact, as $p \to 0$ if the spammer is able to push the marginal cost of the attack $c_s = 0$, the strategy to attack is weakly dominant for the attacker. At the other end, as $p \to 1$ the payoffs for secure players are either positive or 0, while Player 2 has only losses (assuming $\beta \geq 0$). In this case, a strictly dominant strategy emerges in which the attacker avoids loss by not incurring any cost; Player 2 chooses not to engage in the game by choosing not to screen in the first place, forming a pure Nash equilibrium at $(\neg s, -), p = 1$.

Between these extremes ($0 < p < 1$), we find the attacker decision driven by both p and potential lost benefit β. In order for the scam to be viable, the attacker must believe that both the attack cost and potential for failed followthrough are sufficiently low to justify the effort of identifying unsecure users ($c_s \leq l - p(\beta + l)$ for $p < 1$). As it is to the attacker's benefit for this distribution to be in their favour ($p < \frac{1}{2}$), as education with respect to such scams grows (e.g. p increases) attackers must also carefully consider c_s. However, even with more unsecure than secure players, as $\beta \to l$ the ability for the scam to be profitable is quickly constrained by the potential payout and the attacker's cost ($c_s \leq l - 2pl$ for $p < \frac{1}{2}$). The attacker relies on Player 1 to find $\beta < \frac{l}{p} - l$ so as not to invoke a response from a secure user (e.g. one who does not complete the transaction), resulting in the consumption of resources for no gain. As well as introducing the potential deterrent of secure users purposefully engaging in the scam in order to consume resources, this threat of engaging with a non-viable target speaks to the heart of Herley's finding: it is to the attacker's best interest to utilise devices in order to identify the most gullible. As per [12], optimism on the part of the attacker is not a viable long-term strategy.

At this point, our findings are mere restatements of the results of [12]. We see evidence to support the conclusion that "at very low densities certain attacks pose no economic threat to anyone, even though there may be many viable

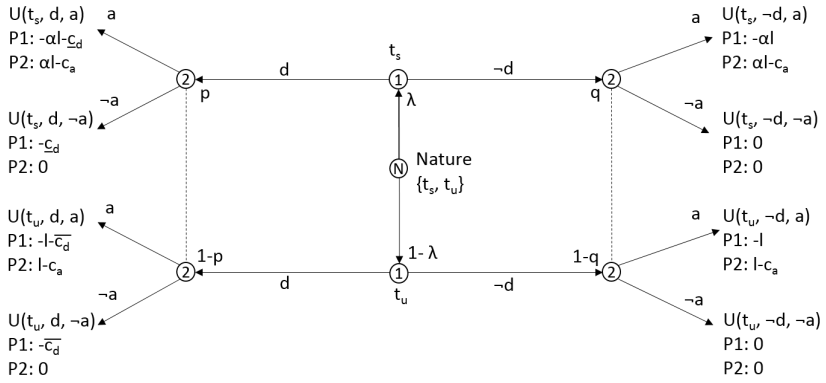

Fig. 3. Extensive form of the deterrence signalling game.

targets" [12]. As shown, the belief of the attacker is critical; as viable target density decreases the attacker's belief that a potential target is secure rationally rises, leading to an attacker trade-space that must consider attacker costs and user benefit — with an increase in either quickly pushing the equilibrium towards deterrence $(\neg s, -)$. We next look at a more complex game in which Player 1 moves first to signal their type and thus deter the attack at the onset. This will serve to account for the actions Player 1 might take in a more interactive defensive posture.

3.2 Deterrence as Signalling: An Example of Operational Deterrence

We now examine operational deterrence (in which a class of operations is deterred, but not the attackers themselves) within the context of a signalling game, as depicted in Figure 3. This construct is based upon the concept of actions signalling a particular security level (secure or unsecure) for the purpose of deterring an attack. Using our conceptual scenario of a wireless network, we examine the employment of 'weak' security constructs (such as SSID hiding) as a means to signal that a user is secure. This game proceeds as follows.

1. Player 'Nature' moves, allocating the distribution of the types of users. As the real distribution of secure versus unsecure users is scenario-specific, we represent this as a probability λ of being secure (t_s), and a corresponding probability of $(1 - \lambda)$ of being unsecure (t_u).
2. Player 1 (the agent) then chooses to send (d) or not to send $(\neg d)$ a 'message' — that is, chooses to deter (e.g. hiding the SSID) or not — with the former action implying a cost that differs between types of user. Thus, the action costs secure users a low amount $\underline{c_d}$, while unsecure users will incur a higher cost of $\overline{c_d}$. In this model, messages have no meaningful effect on security; the question to be addressed is whether they can nevertheless deter attacks.

3. Player 2 (the principal) observes the message (deterrent) and subsequently chooses to attack or not attack, a or $\neg a$. Attacking incurs a cost of attack, c_a. Attacking a user of type t_u will be assumed to succeed, resulting in a gain of l (Player 1's loss); whereas attacking a user of type t_s will be assumed to succeed only with some small probability α, resulting in a gain of αl. At any point the attacker chooses not to attack ($\neg a$), the resulting cost is 0.

We assume a difference in the cost to the secure user t_s and to the unsecure user t_u to send this signal, in which the latter is significantly higher ($t_u \gg t_s$). The cost for an unsecure user to act secure (without actually being secure) warrants such a distinction, with experience costs being the primary differentiator. For instance, setting up wireless security on a modern home router can often be accomplished through a single action, as manufacturers seek to improve user experience. We can identify 'secure' users as those with experience enough to either use such mechanisms or by virtue of their own knowledge of how to do this themselves, and 'unsecure' users who may simply plug in the device and take no action — and who, presumably, are so due to a lack of understanding that would impose significant awareness costs if they were to only hide their SSID, but not implement any further security.

One important aspect of this type of game is Player 2's inability to discern the type of the user. As such, the best that Player 2 can do is to form a set of beliefs as to which type of agent (t_s or t_u) they are playing. This is represented by the value p, corresponding to the belief probability that a message d (that is, deterrent action) corresponds to a user of type t_s, and the corresponding belief probability $(1-p)$ it indicates a player of type t_u. The belief probability q (and $(1-q)$) serves the same function for $\neg d$.

We now analyse this game for equilibria, which for such games is defined by Perfect Bayesian Nash Equilibria (PBNE). Here, multiple conditions have to be met for a strategy profile to be in equilibrium: players must have a belief (probability distribution) as to the node reached any time they make a decision; players must act optimally given their beliefs and the continuation of the game; and beliefs are determined by Bayes' rule on the equilibrium path (as well as those off the path where possible). There are three types of equilibria that can come into play in such games:

- *Separating equilibria*, where a message (the deterrence action) perfectly separates the types of users.
- *Hybrid equilibria*, where a user type may receive the same expected utility from multiple actions and therefore randomise their response.
- *Pooling equilibria*, whereby one or both types find it profitable to take the same action (deter or not deter).

We start by examining for a separating equilibrium, noting that two types of such equilibria are possible: secure users deter, while unsecure users do not; and unsecure users deter, while secure users do not. Looking first at the latter, we note that this corresponds to beliefs of $p = 0$ and $q = 1$. We examine the utilities to Player 2 and see that, given these beliefs, we examine the strategy

for Player 2 and find a likely course of action to be a in the case of seeing d, in that

$$E[U_{Player2}(d, a)] \geq E[U_{Player2}(d, \neg a)] \Rightarrow l - c_a \geq 0$$

where $l > c_a$.

Likewise, Player 2 may attack upon seeing $\neg d$ according to the value of $l > \frac{c_a}{\alpha}$:

$$E[U_{Player2}(\neg d, a)] \geq E[U_{Player2}(\neg d, \neg a)] \Rightarrow \alpha l - c_a \geq 0$$

However, in this instance we see that there exists a profitable deviation by Player 1. Given a by Player 2, while a type t_s player has no motivation to deviate (since $-\alpha l > -\alpha l - \underline{c_d}$), a player of type t_u finds it beneficial to switch and play $\neg d$ as $-l > -l - \overline{c_d}$. As such, a separating equilibrium cannot exist in this case since a profitable deviation exists. In general, we can see from the game that, due to the symmetry of the payoff to Player 2 in the case of t_u, Player 1 of this type will always find it profitable to deviate and play $\neg d$ when $p = 0$ and $q = 1$ due to the cost of deterrence.

Looking now at the case where secure users deter and unsecure users do not, we employ beliefs $p = 1$ and $q = 0$. In this instance it is beneficial for Player 2 to refrain from attack upon seeing the signal d, as when $p = 1$,

$$E[U_{Player2}(d, \neg a)] \geq E[U_{Player2}(d, a)] \Rightarrow 0 \geq \alpha l - c_a$$

where $c_a \geq \alpha l$.

Likewise, consistent with $q = 0$, Player 2 finds the best move to be a upon failing to see a deterrent, as long as the gain from attack (e.g. Player 1's loss) is more than the cost of attack, $l > c_a$:

$$E[U_{Player2}(d, a)] \geq E[U_{Player2}(\neg d, \neg a)] \Rightarrow l - c_a \geq 0$$

Examining for deviation, we consider types t_s and see that a deviation to $\neg d$ may be desirable, since $-\alpha l > -\alpha l - \overline{c_d}$. While in this case Player 1 would no longer incur the additional cost of deterring $\underline{c_d}$, consistent with belief $q = 0$, Player 2 should now respond with a since $l - c_a > 0$. As such, deviation is only profitable for Player 1 if $-\alpha l > -\underline{c_d}$; that is, the potential loss (with small probability α) is greater than the cost to deter.

Looking now at type t_u players, we see that in any event a switch from $\neg d$ to d is going to incur an additional cost $\overline{c_d}$. As such, we can conclude that such an equilibrium exists under the condition $\frac{c_d}{\alpha} < l < \overline{c_d}$. Put another way, this equilibrium exists as long as it is inexpensive for secure users to implement a deterrence mechanism (specifically, less than αl), and the cost to unsecure users is greater than their loss l (given attacker beliefs $p = 1$ and $q = 0$). The meaning of this result is somewhat nuanced and requires further exposition; as such, the implication will be further discussed in Section 5. For now, we note that an equilibrium exists under these beliefs and conditions.

Considering hybrid equilibria, we note that the existence of such equilibria would require that actions exist between d and $\neg d$ such that the payoff is the same for one of the user types t_s or t_u. We can see from the game's construct

that no such equilibrium exists. This is due to the cost of deterring which, despite presumably being small (at least for the case of secure users), changes the payoff function for Player 1. It is important to note that if the cost of deterrence to Player 1 or Player 2 reduces towards 0, this game becomes somewhat symmetric in its payoffs and multiple hybrid equilibria become possible. In such an instance the best course of action for the attacker is to randomise their attacks. Such a game would more closely follow the notion of a 'Cheap Talk' game [8], and arguably may have correspondence to current reality. However, we point out that the asymmetry induced serves to strengthen the case for deterrence measures having utility in a comprehensive defensive posture — but only when they impose an attacker cost that is non-negligible. This is consistent with the conceptualisation of deterrence presented by Morral and Jackson in [16].

We now examine the possibility of pooling equilibria, and first consider the case of an equilibrium at d under the assumption that both player types benefit from deterring. Consistent with the belief upon seeing d that $p = \lambda$:

$$E[(U(d,a)] = \lambda(\alpha l - c_a) + (1-\lambda)(l - c_a) = l + \lambda\alpha l - \lambda l - c_a$$

while

$$E[U(d,\neg a)] = 0$$

Therefore we can see this will hold in instances where $\lambda > \frac{c_a - l}{\alpha l - l}$, rendering this possibility plausible with Player 2 playing $\neg a$. However, as we now look at potential deviation, we see that Player 1 has a potential profit in both instances: type t_s players can find a profitable deviation with $0 > -\underline{c_d}$, as can type t_u players with $0 > -\overline{c_d}$. Put another way, Player 1 can get the same amount of payoff (security) without incurring the cost of deterring (consistent with the idea that deterrents have no security value themselves).

Considering now Player 2's move given these potential deviations, we compare the cost of $\neg a$ and a under the belief $q = \lambda$ and find that Player 2 attacks only as:

$$E[(U(d,a)] \geq E[U(d,\neg a)] \Rightarrow l - \lambda\alpha l - \lambda l - c_a \geq 0$$

Therefore, Player 2 would only change from $\neg a$ to a in the event that $\lambda > \frac{c_a - l}{\alpha l - l}$, which is inconsistent with the belief stated previously. Given this, Player 1 has found a profitable deviation and so we can conclude that a pooling equilibrium does not exist at this point.

Next, we examine the possibility of pooling equilibria existing at $\neg d$ (both players finding it beneficial not to deter), noting that the attacker's a posteriori belief in this case must now be $q = 1 - \lambda$. Upon seeing a play of $\neg d$, it is always to the benefit of Player 1 to play a (as $\alpha l - c_a > 0$ and $l - c_a > 0$), with the consideration that

$$E[U(\neg d,a) \geq E[U(\neg d,\neg a) \Rightarrow 2\lambda c_a - c_a + l(\alpha - \lambda\alpha - \lambda) \geq 0$$

We see that this indeed holds in the event that $\lambda < \frac{1}{2}$ (or $\frac{\alpha}{\alpha+1} > \lambda$), with Player 1 payoffs of $-\alpha l$ for t_s and $-l$ for type t_u. Put another way, this is true only when the distribution of unsecure users is dominant, or the probability of

success against a secure user is much greater than the instances of secure users. Examining now for deviation, we see in both cases that the payoff for Player 1 is reduced in each case (secure and unsecure users), as each faces the same potential loss and additionally incurs the cost of deterring. Therefore, a pooling equilibrium potentially exists whereby Player 1 chooses not to deter and Player 2 chooses to attack, with the beliefs $p = \lambda$, $q = 1 - \lambda$, and $\lambda < \frac{1}{2}$.

Finally, using the same approach, it is straightforward to show that another potential pooling equilibrium exists, with both types of Player 1 choosing not to deter and Player 2 choosing not to attack, as

$$E[U(\neg d, \neg a) \geq E[U(\neg d, a) \Rightarrow 0 > 2\lambda c_a - c_a + l(\alpha - \lambda\alpha - \lambda)$$

when $\lambda > \frac{1}{2}$, all else being the same.

Discussion on the realism of these beliefs is saved for Section 3.3; for now, we summarise that we have identified the following potential equilibria:

- A separating equilibrium when $\frac{c_d}{\alpha} < l < \overline{c_d}$, with

$$(P1_s(d), P1_u(\neg d), P2_d(\neg a), P2_{\neg d}(a), p = 1, q = 0).$$

- A pooling equilibrium when $\lambda < \frac{1}{2}$, with

$$(P1_s(\neg d), P1_u(\neg d), P2_d(a), P2_{\neg d}(a), p = \lambda, q = (1 - \lambda)).$$

- A pooling equilibrium when $\lambda > \frac{1}{2}$, with

$$(P1_s(\neg d), P1_u(\neg d), P2_d(\neg a), P2_{\neg d}(\neg a), p = \lambda, q = (1 - \lambda)).$$

3.3 Discussion

Starting with the screening game, the salient question that emerges is: how do we represent shifting attacker beliefs? In [12], Herley touches on this through the notion that the attacker would employ a series of one or more observables for which they can base a value for x in an attempt to classify the victim. We can think of x as now encompassing the necessary information for the choice of belief of the attacker. In this particular scenario since there is only one move by each player this is fully based upon the response of Player 1 to Player 2's screening message s, such that the choice of Player 1 to respond (r) or not to respond ($\neg r$) corresponds to a belief $p = 1$ or $p = 0$, respectively. However, in other scenarios we can conceive of how this might be a combination of positive observables o_+ and negative observables o_-, such that these observations raise or lower the overall value of x and affect the attacker's assessment of viability. In this construct, we can now think of o_- observables as taking on the role of deterrents. Since the value Player 2 assigns to x is directly tied to the true and false positive rates of their classifier, this affects the risk to the attacker, who as noted cannot afford optimism. Minimising the value Player 2 assigns to x will result in two inter-related effects that will contribute to unprofitability: as a given

assessment x is decreased (via such negative observables), the associated user is more likely to be placed into the category of 'not viable' and thus not subject to attack; and as the perceived set of viable users becomes smaller, attackers are faced with having to find ways to increase true positive and reduce false positive rates, or be faced with decreased attacker profits in the ways described in [12]. This rests not on the user type actually being secure or unsecure (i.e. the 'truth' of Player 1's response), but rather on the belief of the attacker. The response (or lack thereof) represents a single measurement upon which the attacker must infer viability.

We could conceive of a more general game, in which multiple measures beyond a single exchange result in complex screening scenarios (e.g. multiple emails) using the notion of positive and negative observables. Such a construct could be useful in characterising activities such as 'reconnaissance' leading to an attack, port probing (reporting open ports or services running on those ports), information contained within a DNS response that may lead the attacker to believe the system is up to date or of a specific type, or system fingerprinting (reporting specific patch levels, installed applications, etc.).

The separating equilibrium in the signalling version of our deterrence game tells us exactly what we might expect: there is a benefit for players to deter, as it conveys belief that the user's type is t_s. Note that for a user of type t_u playing d is off the equilibrium path, and so no information can be ascertained. In fact, due to this equilibrium, such a move is likely to swing the belief of the attacker towards inferring that the user is of type t_s and refrain from attack, thereby providing a type t_u player the best outcome. This equilibrium required beliefs that seeing a deterrent indicated security, and likewise not seeing such deterrents indicated a lack of security; we claim that this is a reasonable assumption, given the abundance of websites advocating such measures. Users who have taken the time to acquire such devices and follow recommendations on their set-up have likely completed true security tasks as well, such as setting up WPA2 encryption. Additionally, this result requires the constraint that $c_a \geq \alpha l$, such that the expected result of attacking a secure player is less than the cost to attack. This is in line with accepted notions of security.

This result shows that the deterrent must also meet the requirement that $\frac{c_d}{\alpha} < l < \overline{c_d}$, so that the cost of deterring for an unsecure user is higher than the expected loss. This may or may not hold, depending on the conceptualisation employed in the game analysis: in our scenario of a wireless user, someone with a lack of equipment, or improper or unusable equipment, might have a hardware investment to overcome. A lack of technical expertise might result in a user finding that developing an understanding of what an SSID is, or how to find a MAC address and set up filtering, to simply be too burdensome — more so than having to cancel a credit card and deal with the removal of a few charges. This strays into aspects such as time valuation and technical expertise, which is clearly going to vary based on the specifics of the scenario. However, for two users with similar routers — one of whom has set up security, and the other who has simply plugged in out-of-the-box — this becomes more reliant

on the user's perceptions and how they value their time. We note that, as the deterrence costs converge $\overline{c_d} \rightarrow \underline{c_d}$, the asymmetry in payoffs between deterring and not deterring disappears, and Player 1 becomes agnostic (as discussed in Section 3.2). This leads to various hybrid equilibria in which secure players are attacked. As $\overline{c_d} \rightarrow \underline{c_d} \rightarrow 0$, this will only hold if the value of the loss decreases as well, and thus nothing of value is being protected. Thus, one result that can be interpreted from this inquiry is that as such 'security' mechanisms become more user-friendly, they may also lose value in their utility to signal security if they don't result in a sufficient cost to the attacker; this is again consistent with accepted concepts of deterrence.

Turning to the pooling equilibrium, we see that the nature of the equilibrium depends on the distribution of secure users λ. Hard metrics of this type are often scarce and difficult to estimate reliably. Fortunately, some empirical research for the wireless network security scenario exists, placing the occurrence of secure routers at 61% in 2006 [13]. While such analyses are fraught with difficulty and only temporally relevant, this result allows us to assert that instances of secure router set-up are (at least somewhat) more common than not. We can now place a value on our a posteriori beliefs (e.g. $\lambda = 0.61$), and find that our first pooling equilibrium is unlikely to hold as it was dependent on $\lambda < \frac{1}{2}$. However, this distribution is consistent with our second equilibrium, in which neither Player 1 type is deterred but Player 2 chose not to attack. This reflects a belief held by Player 2 that secure players are more prevalent (backed by empirical evidence), and that the likelihood of successful attack is small.

All of these outcomes naturally rely on the attacker incurring a sufficient cost $c_a \geq \alpha l$, as with a small c_a the attacker becomes indifferent to various plays (since they incur little or no cost). As $c_a \rightarrow 0$, we again expect a number of hybrid equilibria situations, leading to probabilistic attack strategies. This results in interesting ramifications, especially as network-sniffing software reduces this to a point-and-click exercise.

Combining these results, we can see that changing the outcome of the game involves changing one or more of the salient parameters. Focusing first on costs, we see that in the screening game the key inequality is between the attacker cost (c_s) and the potential payout (l) or benefit (β). In the signalling case, while a sufficient attacker cost ($c_a \geq \alpha l$) must still exist, the key cost relationship shifts to the defender cost ($\underline{c_d}$ or $\overline{c_d}$) and payout (l), driving a similar inequality that is also conditioned on the attacker's success probability (α). In both cases, this finding reinforces our current notions of security — and forms the basis for much of the effort to combat such crimes. In the case of spamming, efforts in the form of botnet take-down, capture, fines and jail time dominate; probabilistic costs which the attacker must consider within c_s, and when considered explicitly are a confirmation of the role law enforcement in a specific country/region has in deterrence. In the case of signalling, the focus within wireless security has been towards improving usability, and thereby lowering user costs. These respective costs represent government and industry actions in response to these issues.

Ultimately, in both of these games it is the perpetuation of the information asymmetry that is of benefit to the user. This of course stands to reason: the less the attacker can determine of the user's security, the greater the benefit to the user's security. What additionally becomes clear through this analysis is that the effect of such mechanisms can be either direct, by signalling the type or viability of a victim, or indirect, leaving the attacker without actionable information. It is here that the user (defender) appears to have the most direct impact on the resulting security, regardless of prior investment or external constructs. Most directly, in the case of screening the action (or inaction) of the defender provides the conditions to drive a binary attacker belief ($p = 1$ or $p = 0$), and, coupled with the threat of a failed engagement, forces equilibrium. It is this adherence to recommended 'good practice' that sets attacker beliefs, and one could conceive different scenarios in which continued iterations require the defender to continually follow such advice (as characterised by the 'the user is the weakest link' ethos). This reinforces the findings of [12] that it is the small, gullible minority who respond to spam that enables the perpetuation of such scams by allowing attackers to believe it is profitable, given its low cost of entry.

In the case of signalling, while by the construct of the game the signal itself (i.e. SSID hiding) fails to have any security impact, the equilibrium found indicates that the value it provides is in affecting attacker belief. This may help explain the continued endorsement of the practice despite widespread understanding that it does little to affect wireless security, and would appear to provide the justification of heeding such advice. Again, this appears to perpetuate the continued adherence to security guidance even if it has dubious contribution to the actual security stance — as long as the good advice is also followed, and the rest 'looks like' security and comes at a sufficiently low cost.

Naturally, these results only hold in specific circumstances. In these games, Player 1 has knowledge of their type, which may not be the case in many circumstances (or is arguably more likely only in that a 'secure' type would identify as such, with all others falling into the 'unsecure' category). Additionally, these results are in the presence of attacks at scale, as wholly different constructs (with different utilities) are required for examination of directed, focused attacks. Given these conditions, from these results we come to the conclusion foreshadowed by the title of the paper: in both cases of games constructed here, there exists a deterrence outcome in which the winning move is not to play.

4 Related Work

The work described in this paper is intertwined with the wider literature on deterrence, cyber security, and adversarial behaviour, although to the authors' knowledge it is the first to tackle the concept of deterrence from the operational and tactical level in cyberspace.

The role of game theory as the construct for examining deterrence is well studied. Relevant to this work is that of Robbins *et al.* [17], in which they present an extension of the 1960's US–USSR game-theoretic model for strate-

gic nuclear deterrence. Their concept of decision criteria being in the "mind's eye" of the adversary and leading to probability assessments has synergy with the signalling game as defined in this paper. Other attempts at defining deterrence mathematically have also employed game-theoretic constructs to measure reductions in intent to attack [19], although it is not clear how this is to be employed when the potential target set is not specifically known. Generally, the interplay between adversary belief manipulation and cost–benefit realisation are the common themes in definitions of deterrence [4, 16].

Attempts to define cyber deterrence typically stem from these traditional military concepts of strategic deterrence. Much of this literature is focused around cyber attack and notions of 'cyberwar' likened to approaches deterrence in the nuclear era; there is no lack of examples of such treatments [9]. Regarding the role of deterrence as a part of the larger concept of cyber defence, Jabbour and Ratazzi [14] discuss deterrence as a combination of denial, increased costs and decreased benefits, noting that the first of these aspects (denial) relies on a defence-dominated environment — which cyberspace is not. This links the second and third aspects (increased costs and decreased benefits) to the notions of assurance and avoidance, but the authors do not specify how this might be exacted or quantified. While this characterisation soundly dismisses the notion that deterrence can be thought of exclusively in traditional terms of 'Mutually Assured Destruction (MAD)' or retaliatory action, it doesn't reach the level of describing how this could be measurably performed — noting only that it will vary with the domain and application.

In the field of security economics, research involving deterrence has thus far focused primarily on the role it plays to dissuade large-scale malicious behaviour. Previous treatments have included deterrence of employee behaviour with regards to information security (to include employee knowledge of the security mechanisms in play) [11], as well as the application of various theories of deterrence with respect to combatting specific cyber crimes, such as online fraud [1]. These contributions represent a growing trend towards examining deterrence in various perspectives outside of war, but retain the emphasis on larger-scale engagements (e.g. many potential bad actors) and are generally abstracted beyond specific interactions between actors.

Grossklags *et al.* [10] investigate the application of static game-theoretic models to cyber security problems. This scope permits the authors to investigate security concerns ranging from redundant network defence, software vulnerability introduction, and insider threats. The primary focus is in the analysis of the trade-off between 'security' and 'insurance' measures in these instances, and on decisions regarding approach rather than allocation. As such, their results lead to conclusions regarding the role of centralised policy planning and organisational structure in defensive posturing. Differences in approach and emphasis aside, our work follows the same vein of utilising such models to provide insights to enhance development, planning and decision-making processes.

Finally, the contribution of Cremonini and Nizovstev [5] examines the role of security investment on attacker decisions. While never using the term 'deter-

rence' to describe this concept, the authors examine the duality of the security contribution ("ability to withstand attacks of a given intensity") and the behavioural contribution ("change in attacker's perception of the target in question") present in any given security investment; as with our work, they rely on the presence of alternative targets. Cremonini and Nizovstev argue that this second component is often ignored, and develop a model in four scenarios to capture this effect — the fourth of which is an incomplete, asymmetric information game of similar construct to our operational deterrence model. With their focus on the investment aspects, Cremonini and Nizovstev come to the conclusion that the magnitude of the behavioural contribution can greatly exceed that of the direct investment. Additionally, they find that in the incomplete, asymmetric information case attacker treatment of each target is the same, and thus this behavioural component is removed. They argue that this lack of "transparency" in favour of "opacity" is a benefit for the less secure, at the detriment of the more secure users who as a result are disincentivized to invest in security. It is this phenomenon to which they attribute the failure of econometrics such as Annual Loss Expectancy (ALE) to properly capture the security investment, as it fails to account for such indirect contribution and may lead to underinvestment or misallocation of resources.

The model of [5] shares many common themes and concepts with our construct, with both drawing conclusions along complementary lines. In addition to considering the role of screening within potential behavioural contributions, our model most identifies a concrete example of such a mechanism. This addresses a concern of Cremonini and Nizovstev [5] as to what can "serve as a credible signal of strong inner security and not undermine that security at the same time". In addition, our construct further extends the discussion of 'transparency' and 'opacity' to more fully characterise instances (in the form of game equilibria) where the role of belief can be observed. In the instance that the signal is not seen as an indicator of security, the two resulting pooling equilibria are then driven by the attacker beliefs (and therefore can be considered to be related to the prior probabilities). The first equilibrium is analogous to the findings of Cremonini and Nizovstev, where 'opacity' leads to each defender being attacked equally when $\lambda < \frac{1}{2}$. We also find that in the case that the distribution shifts toward secure users ($\lambda > \frac{1}{2}$) another equilibrium is possible, whereby the situation flips such that everyone benefits as the attacker chooses not to move. This result is not considered by Cremonini and Nizovstev, although it serves to support their conclusions regarding the role of the behavioural component within security. In addition, we also find that when the cost of signalling by a less secure player is sufficiently expensive (coupled with attacker beliefs regarding the role of such signals) a 'transparent' environment with a separating equilibrium emerges, which clearly benefits investment in security. Through a more descriptive treatment of signalling by user types within this environment, we complement [5] with a description that relates cost to loss l and loss probability α. This allows the actions of both low and high security users to be more granular with respect to the desired outcome. As such, our construct suggests

that the behavioural component to security and its benefit is indeed still present in these cases, and is reliant on attacker beliefs. These findings further bolster the arguments made in the conclusion of [5] regarding the rejection of 'security through obscurity' and the role of layered defence.

5 Conclusions

We have demonstrated the explanatory power gained by treating the concept of deterrence as an information asymmetry, which is then modelled as a set of games: a screening game, where the attacker moves first and attempts to identify targets for attack, and a signalling game in which the user undertakes measures to attempt to deter potential attackers. In both cases, we showed how the propagation of the asymmetry through the action (or inaction) of the user provided security benefits that can be measured in terms of utility.

We do not attempt to make an argument for deterrence to replace security (to, for example, forgo WPA2 encryption and merely hide one's SSID). In fact, the results show that such constructs have no value in the absence of secure users. Notably, this construct has relevance only to low focus, low skill attacks. As such, they operate as part of a filter for the 'background noise' of internet attacks, but as noted don't hold for directed attacks. The model as presented is highly simplified in its consideration of the cost to the attacker and the user. In addition to the various parameters of the model that may vary from case to case, there are assumptions (such as the equality in the loss of Player 1 and the gain of Player 2) that would be far more complex in reality.

We plan to further investigate the effects of more detailed modelling; however, it is the authors' belief that the value of such concepts lie not in more complex models, but in their explanatory power to describe alternative and complementary concepts of security. As such, such models are expected to have an impact on the formation of requirements and the approaches to security engineering that result from such insights. Movement from the existing paradigms require that we think differently about security throughout the engineering life cycle, and expand our ability to conceptualise and quantify our approaches.

Acknowledgements. We would like to thank Paul Ratazzi and Kamal Jabbour for sharing their previous work, Kasper Rasmussen for the discussion that led to this line of investigation, and Luke Garratt for his insights. We are also grateful to the anonymous reviewers for their helpful and constructive comments.

References

1. Alanezi, F., Brooks, L.: Combatting online fraud in Saudi Arabia using General Deterrence Theory (GDT). In: Proceedings of the 20th Americas Conference on Information Systems (AMCIS 2014) (2014)
2. Alberts, D.S.: Defensive Information Warfare. National Defense University Press, Washington, D.C. (1996)

3. Anderson, R.: The economics of information security. Science 314(5799), 610–613 (October 2006)
4. Bunn, M.E.: Can deterrence be tailored? Tech. Rep. 225, Institute for National Strategic Studies National Defense University, http://www.ndu.edu/inss/nduhp (January 2007)
5. Cremonini, M., Nizovtsev, D.: Understanding and influencing attackers' decisions: Implications for security investment strategies. In: Proceedings of the 5th Annual Workshop on the Economics of Information Security (WEIS 2006) (2006)
6. Dictionary.com: Deter — define deter at dictionary.com (January 2014), http://dictionary.reference.com/browse/deter
7. Dictionary.com: Deterrence — define deterrence at dictionary.com (January 2014), http://dictionary.reference.com/browse/deterrence?s=t
8. Gibbons, R.: Game Theory for Applied Economists. Princeton University Press, Princeton, NJ (1992)
9. Gray, C.S.: Deterrence and the nature of strategy. Small Wars & Insurgencies 11(2), 17–26 (2000)
10. Grossklags, J., Christin, N., Chuang, J.: Secure or insure?: A game-theoretic analysis of information security games. In: Proceedings of the 17th International Conference on World Wide Web (WWW '08). pp. 209–218. ACM (2008), http://doi.acm.org/10.1145/1367497.1367526
11. Herath, T., Rao, H.R.: Protection motivation and deterrence: A framework for security policy compliance in organisations. European Journal of Information Systems 18(2), 106–125 (2009)
12. Herley, C.: Why do Nigerian scammers say they are from Nigeria? In: Proceedings of the 11th Annual Workshop on the Economics of Information Security (WEIS 2012) (2012), http://research.microsoft.com/apps/pubs/default.aspx?id=167719
13. Hottell, M., Carter, D., Deniszczuk, M.: Predictors of home-based wireless security. In: Proceedings of the 5th Annual Workshop on the Economics of Information Security (WEIS 2006) (2006)
14. Jabbour, K.T., Ratazzi, E.P.: Deterrence in cyberspace. In: Lowther, A. (ed.) Thinking About Deterrence: Enduring Questions in a Time of Rising Powers, Rogue Regimes, and Terrorism, pp. 37–47. Air University Press (2013)
15. Moore, T., Anderson, R.: Economics and internet security: A survey of recent analytical, empirical and behavioral research. Tech. Rep. TR-03-11, Computer Science Group, Harvard University (2011)
16. Morral, A.R., Jackson, B.A.: Understanding the role of deterrence in counterterrorism security. Tech. Rep. OP-281-RC, RAND Corporation, Santa Monica, CA (2009)
17. Robbins, E.H., Hustus, H., Blackwell, J.A.: Mathematical foundaitons of strategic deterrence. In: Lowther, A. (ed.) Thinking About Deterrence: Enduring Questions in a Time of Rising Powers, Rogue Regimes, and Terrorism, pp. 137–165. Air University Press (2013)
18. Schneier, B.: Schneier on security: Lessons from the Sony hack. www.schneier.com/blog/archives/2014/12/lessons_from_th_4.html (December 2014)
19. Taquechel, E.F., Lewis, T.G.: How to quantify deterrence and reduce critical infrastructure risk. Homeland Security Affairs 8, Article 12 (2012)
20. Tirenin, W., Faatz, D.: A concept for strategic cyber defense. In: IEEE Military Communications Conference 1999 (MILCOM 1999). vol. 1, pp. 458–463 (1999)
21. Varian, H.R.: Intermediate Microeconomics: A Modern Approach. W. W. Norton & Company, 7th edition edn. (2005)