A Game-Theoretic Investigation of the Effect of Human Interactive Proofs on Spam E-mail

Dimitrios K. Vassilakis, Ion Androutsopoulos and Evangelos F. Magirou
Department of Informatics
Athens University of Economics and Business
Patission 76, GR-104 34 Athens, Greece

ABSTRACT

We show how a game-theoretic model of spam e-mailing, which we had introduced in previous work, can be extended to include the possibility of employing Human Interactive Proofs (HIPs) in conjunction with filters that classify incoming messages as legitimate or spam. Using our extended model, we show that making HIPs widely available to e-mail users will reduce the volume of spam on the Internet and increase the benefit that legitimate users obtain from e-mail.

1. INTRODUCTION

Human Interactive Proofs (HIPs) are puzzles intended to be solved easily by humans, but not computers; they are typically used to prevent computer agents from accessing particular on-line services [2]. For example, e-mail providers often require users who apply on-line for accounts or who have exceeded a maximum number of outgoing messages to retype random sequences of characters that appear in distorted images, in an attempt to deter spammers from acquiring automatically and abusing large numbers of accounts [21]. Similar HIPs have been employed in on-line polls, chatrooms etc.

In anti-spam e-mail filtering, HIPs can be used on their own or to complement other counter-measures, such as filters based on machine learning algorithms [20, 8, 15, 9, 13, 14]. For instance, to avoid the risk of missing misclassified legitimate (ham) messages, users of learning-based filters might choose to return to the senders any messages their filters classify as spam, each along with a request to solve a HIP.[1] The sender of each returned message would be instructed to repost the original message, this time with the solution of the HIP in, for example, the first line of the body. The message would pass unfiltered the second time it would be received, provided that the answer to the HIP is correct; the sender's address might also be added to the recipient's white-list, to ensure that future messages from the same

---

[1] Alternatively, the sender's *computer* may be requested to perform a mildly costly computation, which becomes prohibitive for spammers, who have to respond to large volumes of such requests [10]. This, however, provides an incentive for spammers to compromise and exploit larger numbers of innocent users' computers, known as zombies. Instead, here we focus on challenges that require responses by the senders themselves.

sender would always be treated as legitimate. The rationale behind using HIPs in spam filtering is that spammers post automatically very large numbers (often millions) of messages. Even if only a small percentage of spam recipients request HIPs, the total number of HIP requests received by the spammers will still be very large for them to handle. Hence, spammers will not be able to reply but to a negligible percentage of HIPs.

Although challenges based on visually distorted sequences of characters may sometimes be easier for computers to solve than one might expect [4, 5], and hence spammers may occasionally be able to automate solving this type of HIPs, legitimate e-mail users can easily adopt challenges that are much more difficult for computers. For example, personalized filters could allow each user to specify a different natural language question to be used in his/her HIPs (e.g., "Add the capital of Greece in the first line of the message's body, replacing the first letter of the capital with a question mark."). The filters could also prompt their users to change periodically their questions (e.g., once a month), along with advice on how to make their questions more difficult for computers, for example by avoiding straight factoid questions. Responding automatically (and correctly) to the resulting variety of questions is beyond the capabilities of modern question-answering technology [22].

Even HIPs that can be solved easily by humans, however, can be annoying to, and, when amassed, time-consuming for legitimate senders, and there is the danger that some legitimate users (e.g., potential first-time clients) may not reply to HIPs; this may eventually lead the community of legitimate users to abandon HIPs. More generally, it is unclear what effect HIPs would have if they were adopted widely in spam filtering. For instance, one may hypothesize that a wide adoption of the combination of HIPs and learning-based filters that we highlighted above would initially reduce the number of spam messages that are read by their recipients. This, however, might then lead spammers to send larger volumes of spam (e.g., sending more messages, possibly mutated, to the same recipients; or sending messages to more recipients, who may use different filters). Learning-based filters will always misclassify a (possibly very small, but never zero) percentage of spam messages, and these messages would not trigger HIPs in the scenario we highlighted above; hence, by increasing the volume of outgoing spam (and, consequently, the number of misclassified spam messages that do not trigger HIPs), spammers might be able to ensure that the number of spam messages that are read remains the same as before HIPs were introduced. If the re-

quired increase in spam volume is not prohibitively costly for spammers, the only visible effects of HIPs, then, would be to increase the volume of spam on the Internet and annoy legitimate users by requiring them to reply to challenges.

In this paper, we investigate the effects of HIPs on the problem of spam e-mail using an extension of the game-theoretic model that we introduced in previous work [1]. In the extended model of this paper, the incoming messages of each user are first classified as legitimate or spam by a learning-based filter, and the user can further select a strategy that specifies when HIPs should be requested, depending on the decisions of the filter. As discussed above, we assume that HIPs are natural language questions that change periodically and are different per recipient, so that spammers cannot solve them but in negligible percentages. Provided that reasonable conditions are met, we show that HIPs will in fact both increase the overall benefit that legitimate users obtain from using e-mail and decrease the volume of spam on the Internet. Although alternative game-theoretic models of spam filtering [19] and fraud detection [3] have appeared, as well as investigations of adversarial classification [6], to the best of our knowledge this is the first game-theoretic analysis of HIPs.

Section 2 below presents the game theoretic model, focusing mostly on its differences from the model of our earlier work; Section 3 shows how the Nash equilibria of the game can be computed; Section 4 then uses the equilibria to study the expected effect of HIPs on the overall benefit of legitimate users and the volume of spam on the Internet; Section 5 concludes and proposes directions for further research.

2. THE EXTENDED MODEL

Based on our previous work [1], we model the interaction between spammers and legitimate e-mail users as a one-shot game between the two communities, called players I and II, respectively.[2] Figure 1 shows the extensive form of the game, which is repeated whenever a member of the community of legitimate users requests to obtain his/her next incoming message. Spammers, who play first, may interfere by inserting a spam message (action S) in the user's incoming e-mail stream, causing him/her to obtain a spam message.[3] Alternatively, they may decide not to interfere (action L), in which case the user will obtain his/her next legitimate message. The frequency with which spammers adopt action S over repetitions of the game determines the average ratio of spam to legitimate messages in the users' incoming streams.

Although in reality spam senders do not have the ability to decide whether or not they will insert a spam message in a user's incoming stream on a message per message basis, the overall effect of making this assumption is that spam senders control the average ratio of spam to legitimate messages the legitimate users receive, which we believe is a reasonable assumption. In fact it can be shown that if we construct an

alternative model where the spammers can insert a volume N_S of spam messages, given a volume N_L of legitimate mail, then the same equations are valid and equilibria in the new game correspond to Nash equilibria derived in Section 3 for the one-spam-message formulation of Figure 1. This holds, provided that the final costs and benefits for both players in the new game are strictly proportional to the volumes of spam, legitimate, misread, read, etc. messages involved. Namely, the adverse effects of spam are strictly measured by the spam traffic received. Any overall deterioration of service due to spam related congestion is ignored.

As in our previous work, we assume that all user mailboxes are fitted with spam filters that classify incoming messages as spam ("S") or legitimate ("L"). On average, the filters misclassify spam messages as legitimate ($S \to$"L") with probability ε, and legitimate messages as spam ($L \to$"S") with probability η. Hence, the overall effect of spam filters can be modelled using the chance nodes, labelled F, of Figure 1. The users are not aware of the true classes of the messages unless they read them. Consequently, if they see that a message has been classified as legitimate ("L"), they do not know which of the two nodes of set $II.1$ in Figure 1 the game is at, and similarly for $II.2$. The two sets, $II.1$ and $II.2$, are called *information sets*.

If the incoming message has been classified as spam ("S"), users can trust the filter's decision and delete the message without reading it (action D), or they can ignore the filter and read the message (action R). As an extension to the model of our previous work, they also have a third action available: requesting a HIP (action H). Actions R, D and H are also available when users encounter a message the filter has classified as legitimate ("L").

As already noted, we assume that HIPs are natural language questions that spammers in effect cannot solve. In contrast, we assume for simplicity that legitimate senders always reply (correctly) to HIPs; alternatively, one can incorporate the average cost of missing a legitimate message that incurs when the sender does not reply to a HIP into the cost of requesting a HIP for a legitimate message, to be discussed below; the formulation of the game and our analysis are not affected.[4] Hence, action H always reveals the true class of the message, at the expense of a small cost for the community of legitimate users, if the message is legitimate; the cost comprises the sender's effort to solve the HIP, the cost of the possible delay in reading the message, etc. In contrast, if the message is spam, we assume that requesting a HIP has negligible cost for the community of legitimate users, since the questions of the challenges change rarely (e.g., once every a few months), all other aspects of the process to request HIPs can be automated, and the cost of posting a message back to its sender is very small compared to the users' other cost factors, which we discuss below.

A possible problem is that spammers may counter-attack HIPs by forging spam messages to make them appear as if they were sent by existing legitimate users; then HIP requests for spam messages will be delivered to legitimate users, and the cost for the community of legitimate users will not be zero. Our model could be extended in this direction by introducing a cost $-\alpha'_H$ in Figure 1 when a HIP is requested for

---

[2]Although developed independently, the game of our previous work is remarkably similar to the one Cavusoglu and Raghunathan [3] had used to model fraud detection. Our analysis of the game, however, was very different.

[3]Note that the theory of repeated games cannot be applied, since the two communities, and especially the community of users, change dynamically; hence, the basic condition that users observe and remember the history of past repetitions of the game does not hold.

---

[4]One can also model explicitly the possibility of replying to a HIP or not, by introducing additional nodes and actions in Figure 1. Our key findings remain unchanged, but the analysis of the game becomes more complex.
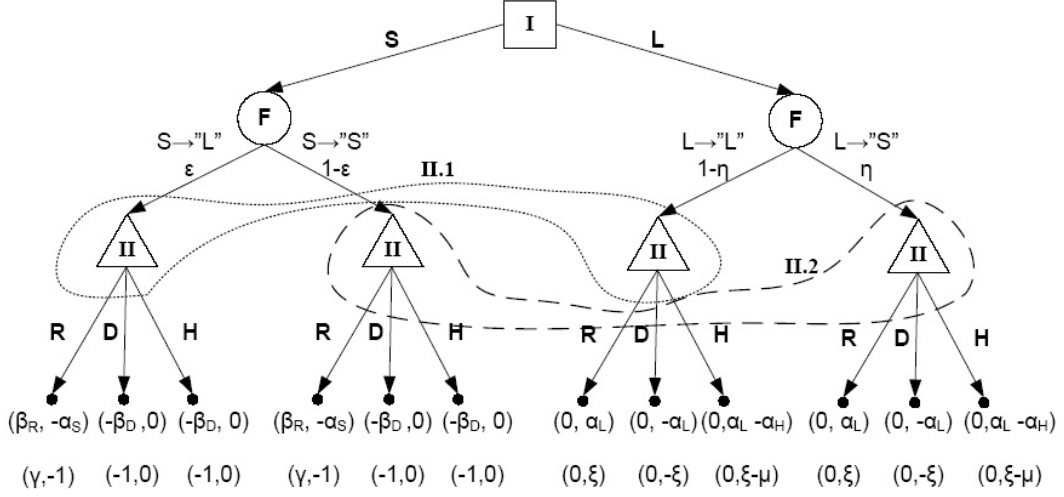
Figure 1: Extensive form of the game, including HIPs.

a spam message, but the key results of our analysis would not change; we return briefly to this possibility in Section 4. For simplicity, our discussion here presupposes that technologies like DKIM or SENDER ID, will have been adopted widely, making forging the senders' addresses very difficult.

Whenever the game is repeated, the actions that players I and II select, along with the filter's decision, lead to a specific leaf node of the tree in Figure 1. Each leaf node corresponds to a possible outcome of the game and contains a pair of values (x, y), where x and y denote the payoffs for players I (spammers) and II (users), respectively. The β_R quantity is the average net benefit the spammers receive from each spam message that is read. The average cost of posting a spam message that is never read is $-\beta_D$. We assume that $\beta_R, \beta_D > 0$. The users' payoff is described by three positive parameters: α_L, α_H, and α_S. When a legitimate message is read without requesting a HIP, the average benefit for the community of users is α_L, whereas an average cost of $-\alpha_L$ is incurred when a legitimate message is missed; on the other hand, reading a spam message costs on average $-\alpha_S$. (A more detailed discussion of the α and β parameters introduced so far, including motivation for setting the cost of missing a legitimate message to $-\alpha_L$, can be found in our previous work.) Requesting a HIP for a legitimate message costs on average $-\alpha_H$ to the users' community, which is deducted from the benefit of reading the message (α_L). Note that requesting a HIP for a legitimate message guarantees that the message will be read, since legitimate senders always respond to HIPs, and reposted messages accompanied by correct HIP answers are always read; in contrast, requesting a HIP for a spam message has zero cost for the users, following our assumption discussed above, and it guarantees that the message will be deleted, since spammers do not reply to HIPs.

To normalize payoffs, let $\xi = \alpha_L/\alpha_S$, $\mu = \alpha_H/\alpha_S$, and $\gamma = \beta_R/\beta_D$, with $\xi > 0$, $\mu > 0$, and $\gamma > 0$. The value of ξ measures how much worse it is for users to miss a legitimate message compared to reading a spam message, and μ compares the average cost of requesting a HIP for a legitimate message to the average cost of reading a spam message; γ is the ratio of the spammers' average benefit from a spam message that is read to the average cost of sending a spam

message that is never read. For simplicity, we pick the units of measurement for the payoffs of players I and II such that $\alpha_S = 1$ and $\beta_D = 1$; then, the payoffs are as in the lowest row of Figure 1. We assume that the payoffs have the properties of utilities.

3. NASH EQUILIBRIA

Each user has to select a strategy that determines what to do with each incoming message, given the decision of the filter; for example, read messages the filter has classified as legitimate and delete messages classified as spam (strategy RD); read messages classified as legitimate and request HIPs for messages classified as spam (RH); read all messages, regardless of the filter's decision (RR). In our case, there are nine such *pure strategies*, namely RR, RD, RH, DR, DD, DH, HR, HD, HH, where the first and second letters determine the user's action when the message has been classified as legitimate or spam, respectively. More generally, a user may adopt a *mixed strategy* σ, whereby the probabilities $\sigma(RR)$, $\sigma(RD)$, $\sigma(RH)$, etc., with $\sum \sigma(\cdot) = 1$, determine how often each one of the nine pure strategies is selected whenever the game is repeated. Similarly, we may assume that the overall community of users (player II) adopts a mixed strategy σ, whose probabilities reflect the frequencies with which the pure strategies are adopted by its members whenever the game is repeated. In the same manner, the community of spammers (player I) adopts an overall strategy π, which specifies the probabilities $\pi(S)$ and $\pi(L) = 1 - \pi(S)$ of selecting actions S and L.

In two-player games, a *Nash equilibrium* is any pair of pure or mixed strategies of the two players, such that no player has an incentive to deviate unilaterally from his/her strategy. When mixed strategies are allowed, every game has at least one Nash equilibrium; and if there is a single Nash equilibrium, it is reasonable to expect that the game will settle at that equilibrium.[5] As with the model of our previous work, we show below that, with the exception of

---

[5]Consult [7] for an informal introduction to game theory. For a more technical introduction, see [18, 16, 17]. Consult Section 3.2 of [17] for an excellent discussion of the possible interpretations of Nash equilibria.

some boundary situations, the spam game of this paper has in general a single Nash equilibrium, despite the additional H action.

Determining the Nash equilibria with mixed strategies in $2 \times M$ <u>non-zero</u> sum games can be achieved using a quasi-diagrammatic procedure, in the spirit of the well known graphical solution for $2 \times M$ <u>zero</u> sum games [18, 11], as shown in our previous work [1]. As M increases, however, that procedure becomes tedious. Instead, it is easier to determine Nash equilibria with *behavioral* strategies, following an approach similar to that of [3]. Unlike mixed strategies, which assign probabilities to the available pure strategies (e.g., $\sigma(RD)$, $\sigma(RH)$, etc. for player II in our game), behavioral strategies assign probabilities to individual actions (e.g., R, D, H) given the information set (e.g., $II.1$, $II.2$) the game is at.

In the framework of behavioral strategies, we assign probabilities p_1, q_1, and $1 - p_1 - q_1$ on actions R, D, and H, respectively, when a user encounters a message the filter has classified as legitimate (information set $II.1$); and p_2, q_2, and $1 - p_2 - q_2$, respectively, on the three actions when the message has been classified as spam (information set $II.2$). Any pair of valid triplets $(p_1, q_1, 1 - p_1 - q_1)$ and $(p_2, q_2, 1 - p_2 - q_2)$ is a behavioral strategy for player II. Similarly, any probability distribution $(t, 1 - t)$ on the spammers' possible actions S and L, with $t \in [0, 1]$, is a behavioral strategy for player I; in the case of player I, behavioral and mixed strategies in effect coincide. The concept of Nash equilibria also applies to behavioral strategies, and will be defined shortly.

Following [12], in a game of perfect recall, i.e., a game where players never forget information they have acquired, which is the case in the game of Figure 1, Nash equilibria for mixed and behavioral strategies are equivalent. Hence, we can derive equilibria in mixed strategies from equilibria in behavioral strategies.

The users' expected payoff at $II.1$ (a "legitimate" message has been received) is:

$$
\begin{aligned}
U_{II.1}(t, p_1, q_1) = & -p_1 P_{1S}(t) + \\
& P_{1L}(t) \left[\xi - \mu + \mu p_1 - (2\xi - \mu) q_1 \right],
\end{aligned}
$$

where $P_{1S}(t)$ denotes the probability that a message classified as legitimate is in fact spam:

$$
P_{1S}(t) = P(S|\text{"}L\text{"}) = \frac{t\varepsilon}{t\varepsilon + (1-t)(1-\eta)},
$$

while $P_{1L}(t)$ denotes the probability that a message classified as legitimate is indeed legitimate:

$$
P_{1L}(t) = 1 - P_{1S}(t).
$$

At $II.2$ (a "spam" message has been received), the users' expected payoff is:

$$
\begin{aligned}
U_{II.2}(t, p_2, q_2) = & -p_2 P_{2S}(t) + \\
& P_{2L}(t) \left[\xi - \mu + \mu p_2 - (2\xi - \mu) q_2 \right],
\end{aligned}
$$

where:

$$
\begin{aligned}
P_{2S}(t) = P(S|\text{"}S\text{"}) & = \frac{t(1-\varepsilon)}{t(1-\varepsilon) + (1-t)\eta} \\
P_{2L}(t) & = 1 - P_{2S}(t).
\end{aligned}
$$

The spammers' expected payoff is:

$$
\begin{aligned}
V(t, p_1, p_2) = & \; t\varepsilon \left[(\gamma + 1)p_1 - 1 \right] + \\
& t(1-\varepsilon) \left[(\gamma + 1)p_2 - 1 \right].
\end{aligned}
$$

A Nash equilibrium in behavioral strategies is a set of values t^* for the spammers and p_1^*, p_2^*, q_1^*, q_2^* for the users that are best responses to each other with respect to the above payoff functions $U_{II.1}$, $U_{II.2}$ and V, in the following sense:

$$
\begin{aligned}
V(t^*, p_1^*, p_2^*) & = \max_t V(t, p_1^*, p_2^*) \\
U_{II.1}(t^*, p_1^*, q_1^*) & = \max_{p_1, q_1} U_{II.1}(t^*, p_1, q_1) \\
U_{II.2}(t^*, p_2^*, q_2^*) & = \max_{p_2, q_2} U_{II.2}(t^*, p_2, q_2).
\end{aligned}
$$

These conditions mean that, if the spammers know the p^*'s and q^*'s to be used by the users, their optimal response will be t^*; similarly the users employ p^*, q^*, provided they know the spam frequency t^* – and in both information sets. The determination of t^*, p^*'s and q^*'s is tedious, since the payoff functions are linear or affine in the decision parameters, but can be done in a straightforward parametric analysis as shown in the Appendix.

Given the operational characteristics (ε, η) of the average filter (see Figure 1), we obtain the values of p_1, p_2, q_1, q_2 and t that correspond to Nash equilibria in behavioral strategies.[6] These values are shown in Table 1. With the exception of the boundary cases $\varepsilon = \frac{1}{1+\gamma}$ and/or $\mu = 2\xi$, where we obtain whole continua of equilibria, there is always a single equilibrium. Note that these results are valid under the reasonable assumption that $\varepsilon + \eta < 1$.

The boundary cases $\mu = 2\xi$ and/or $\varepsilon = \frac{1}{1+\gamma}$ can be ignored, because the probability of μ, ξ, ε, and γ having values that satisfy exactly the boundary equations is practically zero. In all other cases, it can be seen from Table 1 that whether or not users will use HIPs depends on the relation between μ and 2ξ, or equivalently α_H and $2\alpha_L$. When the average cost of requesting a HIP for a legitimate message is more than twice the cost of missing a legitimate message ($\alpha_H > 2\alpha_L$, or equivalently $\mu > 2\xi$), users never use action H ($1 - p_1 - q_1 = 0$ and $1 - p_2 - q_2 = 0$); in that case, the corresponding Nash equilibria in mixed strategies are identical to those of our previous work [1], where HIPs were not available. In contrast, when a_H is less than the critical value $2\alpha_L$, users replace action D with action H in their equilibria. Hereafter, we call "HIP region" and "NO-HIP region" the regions $\mu < 2\xi$ and $\mu > 2\xi$, respectively.

4. THE EFFECTS OF HIPS

We are now ready to investigate the effects of making HIPs available to users, by comparing the equilibria of Table 1 in the HIP and NO-HIP regions. As already noted, if $\mu > 2\xi$ (NO-HIP region), introducing HIPs is pointless, since the users never use them and the equilibria are the same as if HIPs were unavailable. The NO-HIP region can, thus, be thought of as corresponding to the case where HIPs are unavailable or they cost too much for legitimate messages.

When filters operate with a false-negative error rate (ε) that is below $\frac{1}{1+\gamma}$, users always read messages classified as

---

[6]The model can also be used to select the optimal (ε, η) filter configuration, as shown in our previous work.

Table 1: Nash equilibria in behavioral strategies.

Region		$\mu < 2\xi$ (HIP region)	$\mu > 2\xi$ (NO-HIP region)	$\mu = 2\xi$
$\varepsilon < \frac{1}{1+\gamma}$	I:	$t = \frac{\eta\mu}{1-\varepsilon+\eta\mu} \equiv t'_1$	$t = \frac{2\xi\eta}{1-\varepsilon+2\xi\eta} \equiv t_1$	$t = t_1$
	II.1:	$p_1 = 1,\ q_1 = 0$	$p_1 = 1,\ q_1 = 0$	$p_1 = 1,\ q_1 = 0$
	II.2:	$p_2 = \frac{1-\varepsilon-\varepsilon\gamma}{(1-\varepsilon)(1+\gamma)} \equiv \sigma_1,\ q_2 = 0$	$p_2 = \sigma_1,\ q_2 = 1-\sigma_1$	$p_2 = \sigma_1,\ q_2 \in [0, 1-\sigma_1]$
$\varepsilon > \frac{1}{1+\gamma}$	I:	$t = \frac{(1-\eta)\mu}{\varepsilon+(1-\eta)\mu} \equiv t'_2$	$t = \frac{2\xi(1-\eta)}{\varepsilon+2\xi(1-\eta)} \equiv t_2$	$t = t_2$
	II.1:	$p_1 = \frac{1}{\varepsilon(1+\gamma)} \equiv \sigma_2,\ q_1 = 0$	$p_1 = \sigma_2,\ q_1 = 1-\sigma_2$	$p_1 = \sigma_2,\ q_1 \in [0, 1-\sigma_2]$
	II.2:	$p_2 = 0,\ q_2 = 0$	$p_2 = 0,\ q_2 = 1$	$p_2 = 0,\ q_2 \in [0,1]$
$\varepsilon = \frac{1}{1+\gamma}$	I:	$t \in [t'_1, t'_2]$	$t \in [t_1, t_2]$	$t \in [t_1, t_2]$
	II.1:	$p_1 = 1,\ q_1 = 0$	$p_1 = 1,\ q_1 = 0$	$p_1 = 1,\ q_1 = 0$
	II.2:	$p_2 = 0,\ q_2 = 0$	$p_2 = 0,\ q_2 = 1$	$p_2 = 0,\ q_2 \in [0,1]$

"L" ($p_1 = 1$), and they also read a fraction σ_1 of "S" messages ($p_2 = \sigma_1$); for the rest $1 - \sigma_1$ of "S"-labeled messages, users request a HIP in the HIP region, or delete them without reading in the NO-HIP region. When the false-negative error rate is larger ($\varepsilon > \frac{1}{1+\gamma}$), users do not follow blindly the "L" decisions of their filters: for messages classified as "L", they mix actions R and H in the HIP region, or R and D in the NO-HIP region; and they always adopt H or D, in the two regions respectively, when they encounter an "S" message, i.e., that they no longer read any messages classified as spam.

The frequency with which spammers play S over the repetitions of the game determines the average ratio of spam to legitimate messages in the users' incoming streams and on the Internet. When $\varepsilon < \frac{1}{1+\gamma}$, this ratio is t_1 in the NO-HIP region and t'_1 in the HIP region. It can be shown that $t'_1 \le t_1$, with equality when $\mu = 2\xi$. Hence, provided that the cost of requesting a HIP for legitimate messages is kept small enough (HIP region), making HIPs available to legitimate users reduces the average ratio of spam to legitimate messages. We obtain the same effect when $\varepsilon > \frac{1}{1+\gamma}$: spammers now play S with frequency t_2 in the NO-HIP region, and t'_2 in the HIP region; and again it can be shown that $t'_2 \le t_2$. Hence, regardless of the false-negative error rate (ε) of the filters, introducing HIPs always reduces the ratio of spam to legitimate messages, provided that the cost of requesting a HIP for legitimate messages (a_H) is kept within the HIP region. As one would expect, the traffic of spam messages vanishes when a_H becomes zero; this is true both when $\varepsilon < \frac{1}{1+\gamma}$ and when $\varepsilon > \frac{1}{1+\gamma}$.

Let us also investigate the effect of HIPs on the users' expected payoff. The latter is:

$$U = [t\varepsilon + (1-t)(1-\eta)]U_{II.1} + [t(1-\varepsilon) + (1-t)\eta]U_{II.2},$$

where $U_{II.1}$ and $U_{II.2}$ are as in Section 3. ¿From the formula above, we can derive the users' expected payoff in each one of the four equilibria of Table 1, by substituting in $U_{II.1}$ and $U_{II.2}$ the corresponding values of p_1, q_1, p_2, q_2, and t; again, the boundary cases $\mu = 2\xi$ and/or $\varepsilon = \frac{1}{1+\gamma}$ can be ignored, because the probability of encountering them is practically zero. Table 2 shows the users' expected payoff in the four equilibria.

It can be shown that the following inequalities hold:

$$U^1_{\text{HIP}} \ge U^1_{\text{NO-HIP}} \text{ and } U^2_{\text{HIP}} \ge U^2_{\text{NO-HIP}},$$

Table 2: Users' expected benefit in Nash equilibria.

Region	$\mu < 2\xi$	$\mu > 2\xi$
$\varepsilon < \frac{1}{1+\gamma}$	$U^1_{\text{HIP}} = \frac{\xi(1-\varepsilon)-\eta\mu}{\eta\mu+1-\varepsilon}$	$U^1_{\text{NO-HIP}} = \xi\frac{1-2\eta-\varepsilon}{2\xi\eta+1-\varepsilon}$
$\varepsilon > \frac{1}{1+\gamma}$	$U^2_{\text{HIP}} = \frac{\varepsilon(\xi-\mu)}{\varepsilon+(1-\eta)\mu}$	$U^2_{\text{NO-HIP}} = \xi\frac{\varepsilon}{2\xi\eta-2\xi-\varepsilon}$

with equalities holding when $\mu = 2\xi$. Hence, making HIPs available to users always increases their expected payoff (intuitively, the average benefit that they obtain from using e-mail), regardless of the false-negative error rate (ε) of the filters, provided again that the cost of requesting a HIP for a legitimate message is kept in the HIP region ($a_H < 2\alpha_L$, or equivalently $\mu < 2\xi$).

Notice (see Table 1) that the percentage of spam messages that are read is unaffected by the availability of HIPs, unlike what one might expect: when $\varepsilon < \frac{1}{1+\gamma}$, users (player II) always read misclassified spam messages (since $p_1 = 1$, they always read "L" messages) and they read σ_1 of the correctly classified spam messages (since $p_2 = \sigma_1$ in both the HIP and NO-HIP region); and when $\varepsilon > \frac{1}{1+\gamma}$, they always read σ_2 of the misclassified spam messages (since for "L" messages, $p_1 = \sigma_2$ in both the HIP and NO-HIP region) and they never read any of the correctly classified spam messages (since $p_2 = 0$). Hence, the increase that the availability of HIPs brings to the users' expected payoff cannot be attributed to a decrease in the percentage of spam messages the users read; it is purely due to the fact that the spammers post fewer messages ($t'_1 \le t_1$ and $t'_2 \le t_2$, as discussed above) and fewer legitimate messages are missed.

The decrease in spam traffic as a result of the introduction of HIPs is puzzling at first. A plausible explanation of this effect is as follows: in an equilibrium without HIPs, a user reads a message provided that his/her expected net benefit (benefit from reading a legitimate message minus reading a spam) outweighs the loss resulting from outright deleting it. However, when HIPs are available, the user is not required to delete a message, but can send a HIP at a small cost instead. This means that the comparative advantage of reading a message versus not reading it is reduced. Hence, if the benefit from reading used to be marginal to the user, the introduction of the HIP makes it negative. Thus users will opt for a decrease in the outright read option, and start using HIPs instead. This action from the part of the users will

reduce the small equilibrium profits of the spammers, since these profits result from the outright read actions. Hence the spammers have to decrease their spam rate in order to make profitable for the users not to use HIPs outright. The above arguments can be verified by comparing the expressions for the users' benefits in the HIP vs. NO-HIP region.

Our results on the beneficial nature of HIPs are valid even when there is a reasonable but nonzero cost of say $-\alpha'_H$ whenever a HIP is requested for a spam message. Setting $\phi = \frac{\alpha'_H}{\alpha_S}$ and assuming that $\mu < 2\xi$, the conditions $\phi \leq \frac{2\xi - \mu}{2\xi}$ and $\phi \leq \frac{(2\xi - \mu)\eta\varepsilon}{(2\xi - \mu)\eta\varepsilon + (1-\varepsilon)(1-\eta)\mu}$ when $\varepsilon < \frac{1}{1+\gamma}$ and $\varepsilon > \frac{1}{1+\gamma}$ respectively, can be shown to guarantee that users will adopt HIPs and spam traffic will decrease with respect to the NO-HIP region. These bounds on ϕ depend crucially on the net cost of the HIP scheme, the quantity $2\xi - \mu$.

To sum up, making HIPs available to users reduces the volume of spam on the Internet and increases the benefit that users obtain from using e-mail, provided that we are in the HIP region. How easily can the latter be achieved? Recall that the HIP region is $\mu < 2\xi$, or equivalently $\alpha_H < 2\alpha_L$. In other words, we are in the HIP region if the average cost of requesting a HIP for a legitimate message (which includes the average cost in time, irritation etc. to reply to the HIP, the average cost of delaying the message from being read by its recipient, the average cost of occasionally missing a legitimate message that incurs when the sender does not reply to a HIP, etc.) is smaller for the community of legitimate users than the average cost of missing two legitimate messages. We believe that this condition is trivially satisfied. For example, if all legitimate senders respond (correctly) to HIPs, most (if not all) users would select (b) to (a).

(a) You miss two legitimate (ham) messages.

(b) You have to respond to one HIP; your message for which a HIP was requested will be delayed until you respond to the HIP.

Even in an extreme case where senders almost never respond to HIPs and, thus, requesting a HIP for a legitimate message almost guarantees that the message will be missed, we are still in the HIP region, because $\alpha_H \approx \alpha_L$, if we make the reasonable assumption that the (almost certain) cost of missing a legitimate message is much higher than all the other costs included in a_H (e.g., the irritation, loss of time, etc. of the senders, who nevertheless almost never reply to HIPs). Hence, in practice we will always be in the HIP region.

5. CONCLUSIONS AND FUTURE WORK

We showed how our previous game theoretic model of spam e-mailing can be extended to include the possibility of employing HIPs in conjunction with filters that classify incoming messages as legitimate or spam. Our extended model indicates that making HIPs widely available to e-mail users in practice always reduces the volume of spam on the Internet and increases the benefit that legitimate users obtain from using e-mail.

One limitation of our model is that it does not consider the classification confidence of the filters. Most learning-based filters, for example, do not simply return classification decisions ("L" or "S"); they also return numeric scores reflecting their confidence to their decisions. These confidence scores can be used to formulate additional usage scenarios;

for instance, users may wish to request HIPs for all messages that their filters were uncertain of how to classify. Scenarios of this type, however, cannot be investigated with our current model. To support them, our model would have to allow users to formulate their strategies as functions of both the decisions and the confidence scores of their filters, leading to a continuum of strategies for player II. We hope to investigate this issue in further work.

Another possible extension is to consider different populations of users, some with more accurate (but perhaps more expensive) filters than others, and study the effect of HIPs on the incoming spam traffic and expected payoff of each population. This possible extension is inspired by the work of Reshef and Solan [19], who modelled two populations of users, but without considering HIPs and assuming that users always follow the decisions of their filters.

We can also consider the case of a Bayesian game, where some of the opponents' cost parameters are unknown. In this framework of incomplete information, we can also model the possibility of severe changes in operating conditions; for example, a temporary significant decrease in filter accuracy, as a result of a new obfuscation method discovered by the spammers, until an appropriate counter-measure is incorporated in the filters. Finally, a possible extension of our model is to include the spammers' ability to contract out, with some cost, the solving of the HIP challenges.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] I. Androutsopoulos, E. Magirou, and D. Vassilakis. A game theoretic model of spam e-mailing. In *2nd Conference on Email and Anti-Spam*, Stanford, CA, 2005.

[2] H. Baird and D. Lopresti, editors. *2nd International Workshop on Human Interactive Proofs*, number 3517 in Lecture Notes in Computer Science. Springer, 2005.

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

[4] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Computers beat humans at single character recognition in reading based human interaction proofs (HIPs). In *2nd Conference on Email and Anti-Spam*, Stanford, CA, 2005.

[5] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Designing human friendly human interaction proofs (HIPs). In SIGCHI *conference on Human Factors in Computing Systems*, Portland, OR, 2005.

[6] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *10th* ACM KDD *Conference*, Seattle, WA, 2004.

[7] M. Davis. *Game Theory: A Nontechnical Introduction*. Dover Publications, 1983.

[8] H. D. Drucker, D. Wu, and V. Vapnik. Support Vector Machines for spam categorization. IEEE *Transactions*

On Neural Networks, 10(5):1048–1054, 1999.

[9] J. Goodman and W.-T. Yih. Online discriminative spam filter training. In *3rd Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[10] J. T. Goodman and R. Rounthwaite. Stopping outgoing spam. In *5th* ACM *Conference on Electronic Commerce*, New York, NY, 2004.

[11] F. Hillier and G. Lieberman. *Introduction to Operations Research*. McGraw Hill, 2001.

[12] H. W. Kuhn. Extensive games and the problem of information. *Contributions to the Theory of Games*, 2(28):193–216, 1953.

[13] B. Medlock. An adaptive, semi-structured language model approach to spam filtering on a new corpus. In *3rd Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[14] V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam filtering with Naive Bayes – which Naive Bayes? In *3rd Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[15] E. Michelakis, I. Androutsopoulos, G. Paliouras, G. Sakkis, and P. Stamatopoulos. Filtron: a learning-based anti-spam filter. In *1st Conferece on Email and Anti-Spam*, Mountain View, CA, 2004.

[16] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[17] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

[18] G. Owen. *Game Theory*. Academic Press, 1982.

[19] E. Reshef and E. Solan. The effects of anti-spam methods on spam mail. In *3rd Conference on Email and Anti-Spam*, Mountain View, CA, 2006.

[20] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization – Papers from the* AAAI *Workshop*, pages 55–62, Madison, WI, 1998.

[21] P. Simard, R. Szeliski, J. Benaloh, J. Couvreur, and I. Calinov. Using character recognition and segmentation to tell computer from humans. In *7th International Conference on Document Analysis and Recognition*, Edinburgh, UK, 2003.

[22] E. Voorhees. The TREC question answering track. *Natural Language Engineering*, 7(4):361–378, 2001.

APPENDIX

We illustrate the analysis behind Table 1, with the upper left entry. The payoff functions of the users, $U_{II.i}$, $i = 1, 2$, can be written as:

$$U_{II.i}(t, p_i, q_i) = r_i(t) + (\mu - 2\xi)P_{iL}(t) \ q_i + u_i(t) \ p_i,$$

where:

$$
\begin{aligned}
r_i(t) &= P_{iL}(t)(\xi - \mu) \\
u_i(t) &= P_{iL}(t)\mu - P_{iS}(t).
\end{aligned}
$$

For the cheap HIP case ($\mu < 2\xi$), the coefficient of q_i is negative, so it is optimal (and intuitively obvious) that $q_i = 0$ in equilibrium.

We inquire whether there is an equilibrium with the reasonable values of $0 < t < 1$, i.e., some but not all messages are spam, and users do not read all messages classified as spam, i.e., $0 < p_2 < 1$. Now since $0 < p_2 < 1$, it must be $u_2(t) = 0$, namely:

$$(1 - t)\eta\mu - t(1 - \varepsilon) = 0, \text{ or:}$$

or:

$$t = \frac{\eta\mu}{1 - \varepsilon + \eta\mu} \equiv t_1'.$$

For this value of t_1', it can be shown that $u_1(t_1')$ is positive, hence $p_1 = 1$. The expected payoff to the spammers for $p_1 = 1$ becomes:

$$V(t, 1, p_2) = t \left[\varepsilon\gamma + 1 - \varepsilon + p_2(1 - \varepsilon)(1 + \gamma) \right].$$

Since $0 < t_1' < 1$, it must hold that:

$$\varepsilon\gamma + 1 - \varepsilon + p_2(1 - \varepsilon)(1 + \gamma) = 0, \text{ or:}$$

or:

$$p_2 = \frac{1 - \varepsilon - \varepsilon\gamma}{(1 - \varepsilon)(1 + \gamma)} \equiv \sigma_1,$$

where $0 < p_2 < 1$ requires $\gamma > 0$ and $\varepsilon < \frac{1}{1+\gamma}$.