Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS 2005), Stanford University, CA, USA, 2005.

A Game Theoretic Model of Spam E-Mailing

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

Abstract

We discuss how the interaction between spam senders and e-mail users can be modelled as a two-player adversary game. We show how the resulting model can be used to predict the strategies that the two opponent communities will eventually adopt, and how it can be employed to tune anti-spam filters.

1 Introduction

Spam e-mail messages, unsolicited messages sent blindly to very large numbers of recipients, are increasingly flooding mailboxes, undermining the usability of e-mail. Several types of counter-measures have been proposed, including special legislation, pricing policies, and technological responses, such as anti-spam filters; see, for example, Michelakis et al. (2004). We claim that game theoretic models can contribute to the study and further development of such counter-measures. As a first proof of concept, we demonstrate how the interaction between spam senders and e-mail users can be modelled as a two-player game when anti-spam filters are available. We show how the resulting model can be used to predict the behaviour that the two opponent communities will eventually adopt, and how it can guide the tuning of anti-spam filters that offer a tradeoff between two types of misclassification errors.

Section 2 below introduces our game theoretic model and discusses its parameters and assumptions. Section 3 shows how predictions about the behaviour of spam senders and e-mail users can be made by computing the Nash equilibria of the game. Section 4 then demonstrates how the model can be employed to tune anti-spam filters. Finally, section 5 summarizes our findings and proposes directions for further research.

2 Game theoretic model

We model the interaction between spam senders and other e-mail users as a two-player game between the community of spam senders (player I) and the community of e-mail users (player II). Figure 1 shows the game in what is known as extensive form.[1] The game is repeated whenever a user requests to obtain the next message from his incoming e-mail stream. At this point, the spam senders, who play first, can interfere: they may insert a spam message into the incoming e-mail stream of the user (action S in figure 1), which will cause the user to obtain a spam message, or they may do nothing (action L in figure 1), which will cause the user to obtain a non-spam, hereafter called *legitimate*, message. (If there is no legitimate message in the incoming stream, we wait until one arrives.) Thus, the frequency with which spam senders 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 in our model is that the community of spam senders controls the ratio of spam to legitimate messages the community of e-mail users receives, which is a reasonable assumption.

We focus on the scenario where all user mailboxes are fitted with anti-spam filters that flag messages they consider spam; this may be the effect of legislation that requires Internet service providers (ISPs) to provide such filtering facilities to their users. The filters can be modelled as chance nodes, labelled F in figure 1. On average, the filters misclassify spam messages as legitimate ($S \rightarrow$"L" in figure 1) with probability ε, and legitimate messages as spam ($L \rightarrow$"S") with probability η. The users are not aware of the true classes of the messages before they read them, so when they see

[1] For an introduction to game theory, see Davis (1983).

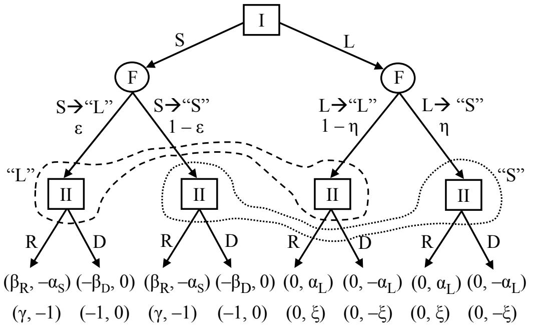
I
S
L
F
F
S→"L"
ε
S→"S"
$1-\varepsilon$
L→"L"
$1-\eta$
L→ "S"
η
"L"
"S"
II
II
II
II
R
D
R
D
R
D
R
D
$(\beta_R, -\alpha_S)$ $(-\beta_D, 0)$ $(\beta_R, -\alpha_S)$ $(-\beta_D, 0)$ $(0, \alpha_L)$ $(0, -\alpha_L)$ $(0, \alpha_L)$ $(0, -\alpha_L)$
$(\gamma, -1)$ $(-1, 0)$ $(\gamma, -1)$ $(-1, 0)$ $(0, \xi)$ $(0, -\xi)$ $(0, \xi)$ $(0, -\xi)$

Figure 1: Extensive form of the spam game

that their filter has classified a message as legitimate, they do not know which of the two U nodes in the set "L" of figure 1 the game is at. Similarly, they cannot distinguish between the two U nodes of the set "S".

When the users encounter a message that has been classified as spam ("S"), they can trust their filter's decision and delete the message without reading it (action D), or they can ignore the filter's decision and read it (action R). For completeness, we assume that the same actions are available with messages the filter has classified as legitimate ("L"): delete the message without reading it (D) or read it (R). Each user has to select a strategy of what he will do with incoming messages depending on the decisions of his filter; for example, read messages classified as legitimate and delete messages classified as spam (strategy RD); or read all messages, regardless of the filter's decision (RR). There are four such *pure strategies*, denoted RD, RR, DR, DD, where the first and second letters determine the user's actions when the filter has classified the message as legitimate or spam, respectively. DD corresponds to the case where a user abandons completely reading e-mail. DR, which does not make much sense, is included only for completeness, and will be disposed off with formal arguments later on.

More generally, a user may adopt a *mixed strategy* σ, whereby whenever the game is repeated he adopts one of the four pure strategies with probabilities $\sigma(RD)$, $\sigma(RR)$, $\sigma(DR)$, and $\sigma(DD)$, respectively, with $\sigma(RD) + \sigma(RR) + \sigma(DR) + \sigma(DD) = 1$. Similarly, we may assume that the overall community of users (player II) adopts a strategy σ, whose probabilities reflect the frequencies with which it adopts actions R and D for messages flagged as legitimate or spam, whenever the game is repeated. In the same manner, the spam posters (player I) adopt an overall strategy π, which specifies the probabilities $\pi(S)$ and $\pi(L) = 1 - \pi(S)$ of selecting actions S and L.

Whenever the game of figure 1 is repeated, the actions that players I and II select lead to a particular cost or benefit for each player. If player I selects action S and the filter misclassifies the message ($S \rightarrow$"L") and player II selects action R, then the game ends (at the leftmost leaf of figure 1) with a benefit of $\beta_R > 0$ for player I and a cost of $-\alpha_S$ for player II. The β_R quantity is the average benefit the spam poster obtains from each spam message that is read, taking into consideration factors such as the average profit from selling the products being advertised, the percentage of users that order products after reading spam messages, etc. The $-\alpha_S$ quantity is the average cost of reading a spam message, taking into consideration factors such as the average cost of downloading it, the average time it takes to read it, and the average cost of being persuaded occasionally to do business with the sender; the latter may be a positive factor, e.g., when a spam message advertises a true bargain, but with fraud messages it will be a negative one. In the remainder of this paper we assume that $-\alpha_S < 0$.

Returning to figure 1, deleting a misclassified spam message without reading it (second leaf from the left) incurs no loss to player II and a loss of $-\beta_D < 0$ to player I; $-\beta_D$ is the average cost of posting a spam message that is never read. Moving to the second II node from the left, reading or deleting without reading a correctly classified spam message incurs the same costs and benefits to the two players as in the first II node from the left, because the costs and benefits depend solely on the true class of the message, not the class assigned to it by the filter. There is a simplification here in the case of action R, because it might be argued that reading a spam message when the filter has correctly flagged it as spam (third leaf from the left) incurs lower cost to the user than when reading a spam message that the filter has mistakenly classified as legitimate (first leaf from the left). For example, a user may read messages flagged as spam late at night, when downloading is less expensive; hence, reading a spam message that has been classified as spam should cost less than reading a spam message that has been classified as legitimate. We will, nevertheless, adopt this simplification to make the analysis of section 3 more manageable.

When the message is legitimate (third and fourth II nodes from the left), there is no cost for player I, while the benefit for player II is $\alpha_L > 0$ if the message is read and $-\alpha_L$ if it is missed. This is another simplification, since the benefit from reading a legitimate message may not be exactly the opposite of the cost of missing it; for example, the benefit of reading it may be the sum of the information value i of the message minus the cost of downloading it, while the cost of missing it may be simply $-i$ if no downloading is involved. Here, we assume that i outweighs any other factor; then, it is reasonable to assume that the benefit of reading the

Table 1: Strategic form of the spam game

I\II	RR	RD	DD
S	$(\gamma, -1)$	$(-1 + \varepsilon(\gamma + 1), -\varepsilon)$	$(-1, 0)$
L	$(0, \xi)$	$(0, \xi - 2\xi\eta)$	$(0, -\xi)$

message is exactly the opposite of the cost of missing it. Also, in the case of action R, it could be argued again that when reading a legitimate message that has been classified as spam, the benefit for player II should be lower than when reading a correctly classified legitimate message; for example, the wrong flagging of the message may have led the user to delay its processing. A more elaborate form of our model could distinguish between two types of R action, read immediately and read with low priority, with different costs attached.

There could also be a fourth user action, for the case where a human interactive proof is requested. In that case, the message is returned to its sender, along with a request to repost it, this time including in the subject the answer to a riddle, to rule out automated spamming software; including the right answer guarantees that the filter will classify the message as legitimate. We leave such enhancements for future work.

We let $\xi = \alpha_L/\alpha_S$ and $\gamma = \beta_R/\beta_D$; following our assumptions, $\xi > 0$ and $\gamma > 0$. In other words, ξ measures how much worse it is for II to miss a legitimate message compared to reading a spam message, and γ is the ratio of player I's 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 (costs or benefits) 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.[2] Furthermore, we assume that the game is played repeatedly over a sufficiently short interval, such that ξ, γ, ε, and η can be treated as constants.

From the extensive form of figure 1 we can construct the game's strategic form. This is a 2 × 4 bimatrix showing the expected payoff of the two players for each combination of pure strategies they may select. Strategy DR of player II, however, is strictly dominated by strategy RD in the plausible case that $\varepsilon < 0.5$ and $\eta < 0.5$, which means that RD always leads to a greater payoff for II than DR, regardless of the strategy that player I selects. Hence, II would never use DR, and the 2 × 4 bimatrix is equivalent to the 2 × 3 bimatrix of table 1. We make the usual assumption that the table entries can be viewed as utilities, and, hence, it is legitimate to maximize expected payoffs.

3 Nash equilibria

Of particular importance in the analysis of games are Nash equilibria, hereafter called simply equilibria. In our case, an equilibrium is any pair (π^*, σ^*) of strategies of the two players, such that $u_I(\pi^*, \sigma^*) \geq u_I(\pi, \sigma^*)$ and $u_{II}(\pi^*, \sigma^*) \geq u_{II}(\pi^*, \sigma)$, for every π and σ, where u_I and u_{II} denote the payoffs to the two players. In other words, no player has an incentive to deviate unilaterally from (π^*, σ^*). When mixed strategies are allowed, every game has at least one equilibrium. In an infinitely repeated two-player game with a single equilibrium, we can expect the game to settle at the equilibrium, and, hence, we can predict the mixed strategies that the players will eventually adopt.[3] We show below that, with the exception of a particular situation, the spam game always has a single equilibrium, and, hence, we can predict the eventual behaviour of the players and their expected payoffs.

The determination of equilibria when mixed strategies are allowed is a computationally difficult problem. Still, in any game with $2 \times M$ pure strategies, as in our case, we can provide a complete listing of the equilibria of interest using a quasi diagrammatic procedure, in the spirit of the well known graphical solution for $2 \times M$ zero sum games (Owen, 1982; Hillier & Lieberman, 2001). We outline the procedure, apply it to a numerical example, and then apply it to the spam game of the previous section.

Consider a $2 \times M$ game whose strategic form is the bimatrix (A, B) with elements (a_{ij}, b_{ij}) representing the payoffs to players I and II respectively, player I selecting rows. Furthermore, assume that we have established that no equilibrium where player I adopts a single pure strategy exists. Hence, we only need to search for equilibria where player I selects his first pure strategy with probability p, with $0 < p < 1$. The best reaction of II to I's choice of p is any mixed strategy that constitutes a distribution on the set of best response pure strategies $J^*(p)$, where:

$$J^*(p) = \underset{j}{\operatorname{argmax}} \, [p \; b_{1j} + (1 - p) \; b_{2j}]$$

The pure strategies in $J^*(p)$ maximize II's expected payoff, given I's p. The expected payoff to II when he adopts any mixture of pure strategies in $J^*(p)$ is the following piecewise linear convex function:

$$G(p) = \max_{j} \, [p \; b_{1j} + (1 - p) \; b_{2j}]$$

On the linear parts of $G(p)$, the best response set $J^*(p)$ consists of a single pure strategy, while at the corners

[2] A better model would specify the outcomes in terms of a non-linear utility function of the costs and benefits.

[3] This is not the only interpretation of mixed strategy Nash equilibria; for an illuminating discussion see section 3.2 of Osborne and Rubinstein (1994).

of $G(p)$ the best response set $J^*(p)$ consists of as many pure strategies cross at that corner, typically two.

Consider first a value of p, with $0 < p < 1$, where $G(p)$ has a corner, and assume for ease of exposition that at that p there are exactly two best pure strategy responses for II in $J^*(p)$, namely j_1 and j_2. Let us also assume that II adopts j_1 and j_2 with probabilities s and $1-s$, respectively. If (p, s) is an equilibrium, player I must be indifferent between his two pure strategies (that he mixes with probabilities p and $1-p$) for the s that II has selected, because otherwise player I would be better off using only the pure strategy that gives him a better payoff, i.e., he would have an incentive to abandon (p, s) for $(1, s)$ or $(0, s)$. Player II must also be indifferent between the two strategies j_1 and j_2, that he mixes with probabilities s and $1-s$, but this is guaranteed by the fact that $j_1, j_2 \in J^*(p)$. Hence, a necessary condition for (p, s) to be an equilibrium is that player I must be indifferent between his two pure strategies. This is also a sufficient condition: if player I is indifferent between his two pure strategies, he has no incentive to change his mixture p; and player II has no incentive to change his mixture s of j_1 and j_2, because they lead to the same (best) payoff; nor does player II have any incentive to start using any other pure strategy outside $J^*(p)$, because by the definition of $J^*(p)$ it would lead to a lower payoff; hence, no player has an incentive to deviate unilaterally from (p, s) and, therefore, (p, s) is an equilibrium.

Therefore, we obtain an equilibrium at p if and only if there is a mixture s of j_1 and j_2, with $0 \leq s \leq 1$, that leaves player I indifferent between his two pure strategies. The latter can be written:

$$a_{1j_1}\ s + a_{1j_2}\ (1-s) = a_{2j_1}\ s + a_{2j_2}\ (1-s)$$

If $a_{1j_1} = a_{2j_1}$ and $a_{1j_2} = a_{2j_2}$, then the previous equality holds for any s, and, hence, we obtain a continuum of equilibria at p. Otherwise, the previous equality has a single solution for s:

$$s = \left(1 - \frac{a_{2j_1} - a_{1j_1}}{a_{2j_2} - a_{1j_2}}\right)^{-1}$$

and $0 \leq s \leq 1$ if and only if the following holds:

$$(a_{2j_1} - a_{1j_1})(a_{2j_2} - a_{1j_2}) \leq 0 \qquad (1)$$

Thus, when $a_{1j_1} \neq a_{2j_1}$ or $a_{1j_2} \neq a_{2j_2}$, there is a single equilibrium at p if inequality (1) holds, and no equilibrium otherwise. The inequality can be interpreted as stating that in the game restricted to columns j_1 and j_2, I's pure strategies are not strictly dominated.[4]

Table 2: Strategic form of an example game

I \ II	1	2	3
1	(7,2)	(1,7)	(1,6)
2	(2,7)	(6,2)	(3,5)

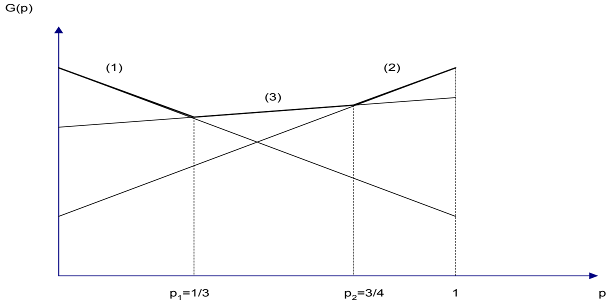

Figure 2: Best responses of II in the example game

For any p in a linear part of $G(p)$ where $J^*(p) = \{j^*\}$, with $0 < p < 1$, we obtain an equilibrium if and only if $a_{1j^*} = a_{2j^*}$; in fact we obtain a continuum of equilibria, for any mixture p of player I in the linear part.

To illustrate the procedure, we apply it first to a numerical example whose strategic form is shown in table 2. There is no equilibrium where player I uses a single pure strategy: if I uses only his strategy 1, then II's best response is his strategy 2, but then player I will be tempted to changes his strategy to 2; similarly, if I uses only strategy 2, then II's best response is strategy 1, but then player I will be tempted to change his strategy to 1. In this game, $G(p)$ is the following function, shown diagramatically in figure 2.

$$G(p) = \max\{7 - 5p; \quad 2 + 5p; \quad 5 + p\}$$

The corners of $G(p)$ are at $p_1 = \frac{1}{3}$, the intersection of strategies 1 and 3, and $p_2 = \frac{3}{4}$, the intersection of strategies 2 and 3, i.e., $J^*(\frac{1}{3}) = \{1, 3\}$ and $J^*(\frac{3}{4}) = \{2, 3\}$. There are no equilibria in the linear segments (1), (2), and (3) of $G(p)$, because $7 \neq 2$, $1 \neq 6$, and $1 \neq 3$. To check the first corner for equilibria, where $p = \frac{1}{3}$, we note that $a_{11} \neq a_{21}$, $a_{13} \neq a_{23}$, and that the game restricted to its first and third columns does not show any domination in I's strategies; hence we get a single equilibrium. II's equilibrium mixed strategy is found by determining a mixing parameter s among strategies 1 and 3 that makes I indifferent between his two strategies, i.e.:

$$7\ s + 1\ (1-s) = 2\ s + 3\ (1-s)$$

and thus $s = \frac{2}{7}$. Hence, the equilibrium mixed strategies at $p_1 = \frac{1}{3}$ are $(\frac{1}{3}, \frac{2}{3})$ for I and $(\frac{2}{7}, 0, \frac{5}{7})$ for II, with expected payoffs $\frac{19}{7}$ for player I and $\frac{16}{3}$ for player II. To check the second corner, where $p = \frac{3}{4}$, we note that $a_{12} \neq a_{22}$ and $a_{13} \neq a_{23}$. Furthermore, the

[4] Note that in a general $2 \times M$ game there must be at least one corner where the inequality is valid, for otherwise player I must have a strictly dominating strategy and the game can be simplified further. Hence we get for this class of games a constructive proof of Nash's existence theorem.

game restricted to its second and third columns shows a strictly dominating strategy for I, namely row two. Hence, we get no equilibrium at $p = \frac{3}{4}$.

If the expected payoffs for strategy 2 of I and 2 of II are changed to $(1, 2)$, instead of $(6, 2)$, there is still no equilibrium where player I uses only his second pure strategy ($p = 0$), but there is now an equilibrium where he uses only his first pure strategy ($p = 1$) and player II uses his second one. There are still no equilibria in segments (1) and (3), and there is still a single equilibrium at the corner $p = \frac{1}{3}$, which is not affected by the change. However, there is now also a single equilibrium at the corner $p = \frac{3}{4}$, because $a_{13} \neq a_{23}$ and no row strictly dominates the other for I in the game restricted to its second and third columns; requiring I to be indifferent between his two strategies leads to the conclusion that II must be using only his second strategy at this corner. In addition, since $a_{12} = a_{22} = 1$, we get a whole continuum of equilibria in the linear segment (2) of $G(p)$, where I uses any mixture p with $\frac{3}{4} < p < 1$ and II again uses only his second strategy. Overall, then, there is a continuum of equilibria for $\frac{3}{4} \leq p \leq 1$, where II uses only his second strategy; the expected payoffs are 1 for I and $2 + 5p$ for II.

Let us now apply the same procedure to the spam game of section 2. The reader is reminded that $\xi > 0$, $\gamma > 0$, and that we made the plausible assumptions that $\varepsilon < 0.5$ and $\eta < 0.5$ to simplify the strategic form of the game. As will be explained in section 4, the latter two assumptions together also imply that $\varepsilon > 0$ and $\eta > 0$. In the game of table 1, then, no equilibrium where player I adopts a single pure strategy exists: if I adopts S, II's best reaction is DD, but then player I will be tempted to change his strategy to L; and if I adopts L, II's best reaction is RR, but then player I will be tempted to change his strategy to S. Hence, at any equilibrium, player I must adopt his first strategy with a probability p, with $0 < p < 1$. Interestingly, at any equilibrium the expected payoff to player I is zero, because his second pure strategy gives zero payoff and he must be indifferent between his two strategies; i.e., the community of spam senders as a whole neither profits nor loses. This is, however, a collective figure; some individual spam senders may profit and others may lose, with their payoffs summing up to zero. The spam senders will never give up posting entirely, because this does not yield an equilibrium, as discussed above. Intuitively, if they all stop posting, the users will start reading all messages (RR), and this will tempt some spam senders to start posting again.

Player II's best response to I's choice of p is any mixed strategy that constitutes a distribution on the set of best response pure strategies $J^*(p)$. The expected outcome for II when he adopts such a mixed strategy is the following piecewise linear convex function, which is shown diagramatically in figure 3. The corners of $G(p)$ are at p_1, the intersection of RR and RD, and p_2, the intersection of RD and DD.

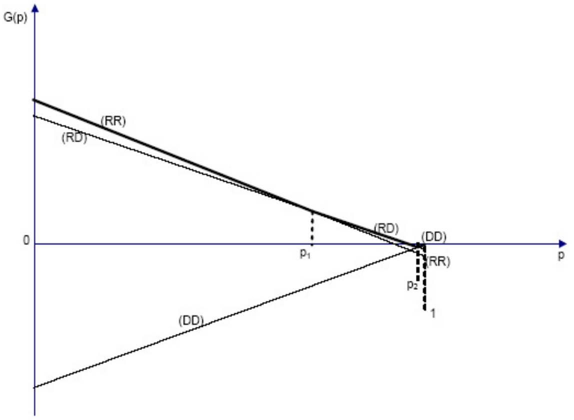

Figure 3: Function $G(p)$ in the spam game

$$G(p) = \max\{\xi - (\xi + 1)p; \;\; \xi - 2\xi\eta + p(2\xi\eta - \xi - \varepsilon); \\ -\xi + \xi p\}$$

Let us first examine the case where $-1 + \varepsilon(\gamma + 1) < 0$, or equivalently $\varepsilon < \frac{1}{\gamma+1}$. In this case, there is no equilibrium in the linear segments of $G(p)$. Furthermore, when the game is restricted to its RD and DD columns, row two gives a strictly dominating pure strategy for player I, and hence there is no equilibrium at the corner of p_2. However, there is no strictly dominating pure strategy for player I in the game restricted to its RR and RD columns, and $\gamma \neq 0$, $-1 + \varepsilon(\gamma + 1) \neq 0$. Hence, there is a single equilibrium at the corner of p_1, which occurs when player I mixes actions S and L with probabilities $(p_1, \; 1 - p_1)$ and player II mixes actions RR and RD with probabilities $(s_1, \; 1 - s_1)$. From the equations of the segments of $G(p)$ and the requirement that player I must be indifferent between his two pure strategies we obtain:

$$p_1 = \frac{2\xi\eta}{1 + 2\xi\eta - \varepsilon}, \quad s_1 = \frac{1 - \varepsilon(\gamma + 1)}{(\gamma + 1)(1 - \varepsilon)} \qquad (2)$$

The expected payoff to player II is then:

$$V_{II}{}^1 = \xi \, \frac{1 - 2\eta - \varepsilon}{1 + 2\xi\eta - \varepsilon} \qquad (3)$$

When $\frac{1}{\gamma+1} < \varepsilon < \frac{1}{2}$, it can be verified that we obtain only a single equilibrium at the corner of p_2, which occurs when player I adopts a mixed strategy $(p_2, 1 - p_2)$, while player II mixes actions RD and DD with probabilities $(s_2, 1 - s_2)$, where:

$$p_2 = \frac{2\xi\eta - 2\xi}{2\xi\eta - 2\xi - \varepsilon}, \quad s_2 = \frac{1}{\varepsilon(\gamma + 1)}$$

The expected payoff to player II is now:

$$V_{II}{}^2 = \xi \; \frac{\varepsilon}{2\xi\eta - 2\xi - \varepsilon} \qquad (4)$$

Finally, when $\varepsilon = \frac{1}{\gamma+1}$ we obtain single equilibria at the corners of p_1 and p_2, and a whole continuum of equilibria in the linear segment of $G(p)$ where II's best response is RD. In all of these equilibria, it can be verified that player II adopts the single pure strategy RD (which means that users always trust their filter's decision), and player I mixes actions S and L with probabilities $(p, 1-p)$, with $p_1 \leq p \leq p_2$. The expected payoff to player II is:

$$V_{II}{}^*(p) = \xi - 2\xi\eta + p(2\xi\eta - \xi - \frac{1}{\gamma+1}) \qquad (5)$$

where p can take any value in $[p_1, p_2]$. The extreme values of this range coincide with $V_{II}{}^1$ and $V_{II}{}^2$.

Note, also, that we have made the implicit assumption that $\frac{1}{\gamma+1} < \frac{1}{2}$, i.e., $\gamma > 1$, for otherwise the cases $\varepsilon \geq \frac{1}{\gamma+1}$ are impossible, since we have also assumed that $\varepsilon < \frac{1}{2}$. Hereafter, we focus on the case where $\gamma > 1$. The case $\gamma \leq 1$ is easier to analyze, since we obtain only the single equilibrium of equations (2)–(3).

4 Filter tuning

An immediate application of the above analysis is to spam filter tuning. In the framework of Statistical Decision Theory (DeGroot, 1970), spam filters are decision making mechanisms in two states of nature (S, L) and two actions ("S", "L"). We assume that the filter produces a scalar score x, normalized in $[0, 1]$, which indicates the filter's confidence that the incoming message is spam. Let $f_S(x) = P(x|S)$ be the distribution of x for spam messages, and $f_L(x) = P(x|L)$ the distribution for legitimate messages. We make the reasonable assumption that f_S is increasing in x and f_L decreasing. The filters that minimize the expected cost of the decision are characterized by the Neyman-Pearson lemma, which in our case states that a message should be classified as spam provided that $f_S(x)/f_L(x) > M$, where M is a function of the costs involved and the a priori probabilities of the two categories of messages (S, L). As illustrated in figure 4, under our assumptions on f_L and f_S the Neyman-Pearson criterion simplifies to classifying a message as spam provided its score x is greater than a value μ, which is determined by the two distributions and M. For this type of filters, we can express ε (probability of $S \rightarrow$ "L" error) and η (probability of $L \rightarrow$ "S" error) as:

$$\varepsilon(\mu) = \int_0^\mu f_S(x)dx \quad \text{and} \quad \eta(\mu) = \int_\mu^1 f_L(x)dx$$

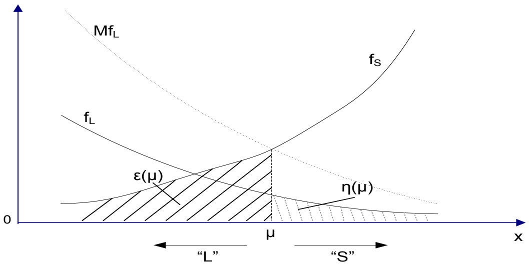

Figure 4: Applying the Neyman-Pearson lemma

All the above are shown in figure 4. In practice, the exact distributions f_S and f_L may be unknown, and we may only have estimates for some of the factors that influence M, such as the a priori probabilities of the categories. The ideal μ of figure 4, then, cannot be identified precisely, and typically filters let their users tune the value of μ by themselves. The selected μ value is used as a threshold on x, i.e., messages whose x is greater than μ are classified as "S", and those with $x \leq \mu$ are classified as "L". Using a μ value to the right of the ideal one in figure 4, causes fewer messages to be classified as "S" and more messages to be classified as "L"; at the same time ε increases, while η decreases. Similarly, using a μ value to the left of the ideal one, decreases ε and increases η, as more messages are classified as "S".[5] The users, then, are faced with a tradeoff between ε and η, which is controlled by the choice of μ and is characterized by the function $\eta = \eta(\varepsilon)$. The first derivative of this function is:

$$\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon} = \frac{\mathrm{d}\eta/\mathrm{d}\mu}{\mathrm{d}\varepsilon/\mathrm{d}\mu} = -\frac{f_L}{f_S}$$

which is negative, while the second derivative is:

$$\frac{\mathrm{d}^2\eta}{\mathrm{d}\varepsilon^2} = \frac{\mathrm{d}}{\mathrm{d}\mu}\left(\frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}\right)\frac{\mathrm{d}\mu}{\mathrm{d}\varepsilon} = -\frac{f'_L f_S - f'_S f_L}{f_S{}^3}$$

which is positive given our assumptions on f_S, f_L, and thus the tradeoff curve is convex. A reasonable form of this curve in the case of a 'good' and a 'bad' filter is demonstrated in figures 5 and 6 respectively, where ε^0 denotes the value of ε that corresponds to $\eta = 0.5$, and similarly for η^0. Note that the assumptions of section 2 that $\varepsilon < 0.5$ and $\eta < 0.5$ imply that both ε and η are bounded away from zero by ε^0 and η^0, respectively.

Returning to the spam game, we may assume that player II uses a single filter, whose f_S and f_L are the average distributions of the filters that are used by the

[5] In effect, the choice of μ allows users to change their filter's bias towards one of the two categories. An alternative biasing mechanism is to modify f_S and f_L, for example by retraining the filter on a biased collection of messages. Here we consider only the case where f_S and f_L are fixed and the only biasing mechanism is the choice of μ.

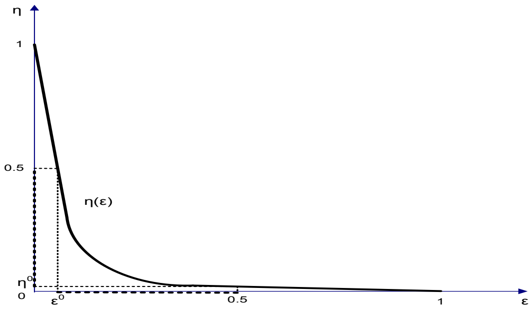

Figure 5: ε-η curve for a 'good' filter

individual users over the repetitions of the spam game. This average filter will be characterized by a fixed ε-η curve similar to those of figures 5 and 6. Player II, then, can only select an ε value, and the corresponding η value is determined by the ε-η curve.[6]

Player II should select the ε value that maximizes his expected payoff. We thus consider ${V_{II}}^1$, ${V_{II}}^2$, and ${V_{II}}^*(p)$ of section 3 with respect to ε. If II selects a relatively small error in the spam classification, i.e., $\varepsilon < \frac{1}{\gamma+1}$, then ${V_{II}}^1$ of equation (3) applies. The first derivative of ${V_{II}}^1$ is:

$$\frac{\mathrm{d}{V_{II}}^1}{\mathrm{d}\varepsilon} = -\frac{2(1+\xi)}{(1-\varepsilon+2\xi\eta)^2}\left(\eta'(1-\varepsilon)+\eta\right)$$

where $\eta' = \frac{\mathrm{d}\eta}{\mathrm{d}\varepsilon}$. The expression $-(\eta'(1-\varepsilon)+\eta)$ is decreasing with respect to ε and vanishes for $\varepsilon = 1$ (corresponding to $\eta = 0$, since $\mu = 1$). Hence the derivative of ${V_{II}}^1$ is positive for $\varepsilon^0 \leq \varepsilon < \frac{1}{1+\gamma}$, indicating that ${V_{II}}^1$ is increasing.

If the user selects a relatively large error in the spam classification, namely $\frac{1}{1+\gamma} < \varepsilon \leq \frac{1}{2}$, ${V_{II}}^2$ of equation (4) applies, which is always negative. Its derivative is:

$$\frac{\mathrm{d}{V_{II}}^2}{\mathrm{d}\varepsilon} = 2\xi^2\frac{\eta-1-\eta'\varepsilon}{(2\xi(\eta-1)-\varepsilon)^2}$$

The expression $\eta - 1 - \eta'\varepsilon$ vanishes for $\varepsilon = 0$ (corresponding to $\eta = 1$, since $\mu = 0$) and is decreasing with respect to ε, so it is negative for $\frac{1}{1+\gamma} < \varepsilon \leq \frac{1}{2}$, showing that ${V_{II}}^2$ is decreasing.

The case $\varepsilon = \frac{1}{\gamma+1} = \varepsilon^*$ is of particular importance, since it gives the supremum of ${V_{II}}^1$ and ${V_{II}}^2$. For this value of ε, we get a continuum of equilibria. The payoff to player II is given by ${V_{II}}^*(p)$ of equation (5), where p lies in $[p_1, p_2]$ while player II uses the pure strategy RD. As pointed out earlier, ${V_{II}}^*(p_1) = {V_{II}}^1(\varepsilon^*)$ and ${V_{II}}^*(p_2) = {V_{II}}^2(\varepsilon^*)$. Thus the payoff to player II is indeterminate, ranging between ${V_{II}}^1(\varepsilon^*)$ and ${V_{II}}^2(\varepsilon^*)$, depending on I's mix p. It is straightforward to show that ${V_{II}}^1 > {V_{II}}^2$ for the pair $(\varepsilon^*,\ \eta^*)$.

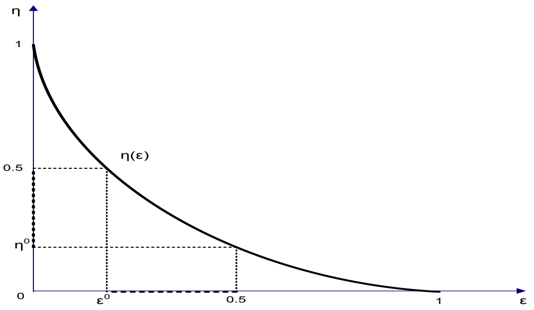

Figure 6: ε-η curve for a 'bad' filter

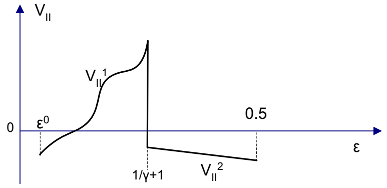

Figure 7: Expected payoff for player II (users)

Therefore, the optimal selection of ε presents a remarkable discontinuity; see figure 7. Player II, the community of e-mail users, should select an ε that is close but less than $\varepsilon^* = \frac{1}{\gamma+1}$, since $V_{II}(\varepsilon)$ tends to get its maximum when $\varepsilon \to \frac{1}{\gamma+1}^-$. Selecting $\varepsilon = \frac{1}{\gamma+1}$ precisely, leads to uncertainty on the outcome of the game, while any $\varepsilon > \frac{1}{\gamma+1}$ gives negative and certainly lower payoffs to II. Hence, a target for the community of e-mail users should be to approach ε^* from lower values, i.e., aim for a particular percentage of misclassified spam messages in the entire community. The users (or their administrators) can achieve this by agreeing to monitor the percentage of spam messages their individual filters misclassify, and tune their filters to keep the percentage close to (but less than) ε^*; then the average filter will also approach ε^* from lower values. The target ε^* depends only on $\gamma = \beta_R/\beta_D$. Hence, to identify ε^* precisely, the users need to know the spam senders' β_R and β_D (average benefit from each spam message read, average cost of each unread spam message). In the absence of accurate information on these parameters, the users should prefer to overestimate γ.

When $\varepsilon \to \varepsilon^{*-}$, equations (2) imply that $s_1 \to 0^+$. That is, the users' eventual (equilibrium) strategy will be to play RD coupled with an infinitesimal mix of RR, namely always read "L" messages and almost always delete "S" messages; an infinitesimal number of "S" messages will be read. (Intuitively, this provides an incentive for the spam senders to limit their traffic, since they have some hope of having their messages read without inundating the network.) Provided that $\varepsilon \to \varepsilon^{*-}$, equations (2) also predict that the percent-

[6] Player II selects μ, but he can get feedback on the resulting ε by collecting statistics on misclassified messages.

age of spam messages in the overall e-mail traffic will eventually approach ${p_1}^*$ from lower values, where:

$$ {p_1}^* = \frac{2\xi\eta^*}{1 + 2\xi\eta^* - \frac{1}{\gamma+1}} $$

${p_1}^*$ is increasing with respect to η^*. As filters improve, η^* decreases, and the predicted percentage of spam messages decreases, as one would expect, although the optimal ε remains unaffected.

It is perhaps surprising that ε^* does not depend on ξ (cost of missing a legitimate message compared to reading a spam message). This is against the intuition that as ξ increases, filters should be tuned for larger ε and smaller η, i.e., classify more easily messages as legitimate, to avoid missing legitimate messages when users trust their filters. This intuition, however, fails to take into account that the choice of ε also affects the behaviour of spam senders, and, thus, the percentage of spam messages users receive. For example, increasing ε may lead spam senders to post more messages, which will also be classified more easily as legitimate, and this may outweigh the users' benefit from misclassifying fewer legitimate messages. Hence, to determine the optimal ε, one must also take into account the reaction of the spam senders, which leads to $\varepsilon \rightarrow {\varepsilon^*}^-$.

An objection that can be raised against our modelling of the spam senders as a single player is that individual spam senders may act selfishly and ignore their community goal of sticking to ${p_1}^*$; a Prisoner's Dilemma situation.[7] For example, some individual spam senders may decide to increase their frequencies of posting spam, causing their community's frequency to exceed ${p_1}^*$ by Δp. This, however, will lead the users to switch to pure strategy RD, generating a $({p_1}^* + \Delta p)(-1 + \varepsilon(\gamma + 1))$ payoff for the community of spam senders, which is negative for $\varepsilon \rightarrow {\varepsilon^*}^-$, lower than the zero payoff of the equilibrium. If the cost parameters of the spam senders are sufficiently similar, this decrease will be felt by the selfish spam senders, who will view the increase of their frequencies as unprofitable, moving back to their original frequencies and restoring the collective frequency to ${p_1}^*$.

5 Conclusions and future work

We have shown how the interaction between spam senders and e-mail users can be modelled as an adversary game. We focused on the scenario where all user mailboxes are fitted with anti-spam filters, and the users can either read messages or delete them without reading, with their actions depending only on the verdicts of the filters. With the exception of a single point in the tradeoff between the filters' two types of error, the game always has a single Nash equilibrium, and, thus, always settles with players adopting particular strategies when repeated infinitely. We showed how the model can be used to determine the optimal point in the tradeoff, which e-mail users should adopt, and we provided a prediction of the eventual percentage of e-mail traffic that will be spam if the optimal point is adopted. Determining the tradeoff's optimal point requires only information on the costs of the spam senders. An immediate possibility, then, is to collect such information. An alternative is to extend our model with techniques from Bayesian games, where some of the opponents' costs are unknown.

We have already pointed out the possibility of extending our model with additional user actions, to distinguish between read-immediately and read-delayed, and allow requests of human-interactive proofs. Additional actions of the spam senders could also be included, to distinguish between sending fraud spam messages and genuine advertisements, blind spam postings and spam messages that are tailored to the interests of their recipients, and plain vs. disguised spam messages (e.g., messages with additional random words, intended to confuse statistical filters).[8] One may also study how ISPs and regulating bodies could influence the spam senders' β_R and β_D parameters, to lower the expected percentage of spam messages.

References

Dalvi, N., Domingos, P., Mausam, Sanghai, S., & Verma, D. (2004). Adversarial classification. *Proceedings of KDD-2004*. Seattle, WA.

Davis, M. (1983). *Game theory: A nontechnical introduction*. Dover Publications.

DeGroot, M. (1970). *Optimal statistical decisions*. McGraw Hill.

Fawcett, T. (2003). "In vivo" spam filtering: A challenge problem for KDD. *SIGKDD Explorations*, *5*, 140–148.

Hillier, F., & Lieberman, G. (2001). *Introduction to operations research*. McGraw Hill.

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G., & Stamatopoulos, P. (2004). Filtron: A learning-based anti-spam filter. *Proceedings of the 1st Conference on Email and Anti-Spam*. Mountain View, CA.

Osborne, M., & Rubinstein, A. (1994). *A course in game theory*. MIT Press.

Owen, G. (1982). *Game theory*. Academic Press.

[7] We thank an anonymous reviewer for raising this point.

[8] Disguised messages are part of an arms race between filter developers and spam senders (Fawcett, 2003). Dalvi et al. (2004) discuss how to model this race as a game.