2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51.123

A Game Theoretic Model of Spam E-Mailing

ZHAO Jie[1,+], LIAO dabao[2] and ZHENG jiasheng[3]

[1] School of Experiment and Practice Training Management Center. Zhejiang Police Vocational Academy.
Hangzhou, Zhejiang Province, China

[2,3] School of Information and Electronic Engineering Zhejiang University of Science and Technoloy
Hangzhou, Zhejiang Province, China

Abstract. Discussed 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.

Keywords: Spam filters; Nash equilibria; Spam sender

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 trade of between two types of misclassification errors[1].

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 form1 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

[+] ZHAO Jie
Zhaojie@zjjy.com.cn

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 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"[2].

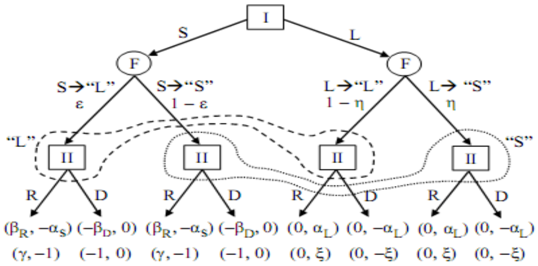
I
S
L
F
F
S→"L"
ε
S→"S"
1 − ε
L→"L"
1 − η
L→ "S"
η
"L"
"S"
II
II
II
II
R
D
R
D
R
D
R
D
$(\beta_R, -\alpha_S)$ $(-\beta_D, 0)$ $(\beta_R, -\alpha_S)$ $(-\beta_D, 0)$ $(0, \alpha_L)$ $(0, -\alpha_L)$ $(0, \alpha_L)$ $(0, -\alpha_L)$
$(\gamma, -1)$ $(-1, 0)$ $(\gamma, -1)$ $(-1, 0)$ $(0, \xi)$ $(0, -\xi)$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

Figure 1. Extensive form of the spam game

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 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)[2,3]; 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 FS completely reading e-mail. DR, which does not make much sense, is included only for completeness, and will be disposed of with formal arguments later on.

When the message is legitimate , 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 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.

TABLE I. STRATEGIC FORM OF THE SPAM GAME

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

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[5]. 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 figure1.Furthermore, we assume that the game is played repeatedly over a suffciently short interval, such that ξ, γ, ε, and η can be treated as constants.

3. Nash equilibria

Of particular importance in the analysis of games are Nash equilibria, hereafter called simply equilibria[4]. In our case, an equilibrium is any pair (π*, σ*) of strategies of the two players, such that uI (π*, σ*) ⩾uI (π, σ *) and uII (π*, σ*) ⩾ uII (π*, σ), for every π andσ , where uI and uII 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 atthe equilibrium, and, hence, we can predict the mixed strategies that the players will eventually adopt3 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 diffcult problem. Still, in any game with 2 ×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 ×M zero sum games,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 × M game whose strategic form is the bimatrix (A,B) with elements (aij , bij) 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:

$$J^*(p) = \arg\max[pb_{1j} + (1-p)b_{2j}] \tag{1}$$

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:

$$G(p) = \max[pb_{1j} + (1-p)b_{2j}] \tag{2}$$

On the linear parts of G(p), the best response set J*(p) consists of a single pure strategy, while at the cornersof 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 j1 and j2. Let us also assume that II adopts j1 and j2 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 that II has selected, because otherwise player

I would be better off using only the pure strategy that gives him a better payof, 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 j1 and j2, that he mixes with probabilities s and 1 − s, but this is guaranteed by the fact that j1, j2 ∈ 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[5,6]. This is also a suffcient condition: if playerI 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 j1 and j2, 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 j1 and j2, with 0 ⩽ s ⩽ 1, that leaves player I indifferent between his two pure strategies. The latter can be written:

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

If $a_{1j1} = a_{2j1}$ and $a_{1j2} = a_{2j2}$, 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:

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

and 0 ⩽ s ⩽ 1 if and only if the following holds:

$$(a_{2j_1}s - a_{1j_1})(a_{2j_2}s - a_{1j_2}) \leq 0 \qquad (3)$$

Thus, when $a_{1j1} = a_{2j1}$ or $a_{1j2} = a_{2j2}$, there is a single equilibrium at p if inequality (3) holds, and no equilibrium otherwise. The inequality can be interpreted as stating that in the game restricted to columns j1 and j2, I's pure strategies are not strictly dominated.

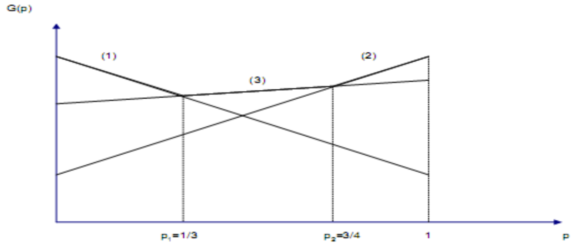

Figure 2. Best responses of II in the example game

TABLE II. 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)

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.

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 fS(x) = P(x|S) be the distribution of x for spam messages, and fL(x) = P(x|L) the distribution for legitimate messages. We make the reasonable assumption that fS is increasing in x and fL 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 fS(x)/fL(x) > M, where M is a function of the costs involved and the a priori probabilities of the two categories of messages(S, L)[6].

Returning to the spam game, we may assume that player II uses a single filter, whose fS and fL are the average distributions of the filters that are used by the 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 p1∗; a Prisoner's Dilemma situation.

For example, some individual spam senders may decide to increase their frequencies of posting spam, causing their community's frequency to exceed p1∗ by Δp. This, however, will lead the users to switch to pure strategy RD, generating a(p1∗ + Δp)(−1 + ε(γ + 1)) payoff for the community of spam senders, which is negative for ε → ε∗−, lower than the zero payoff of the equilibrium. If the cost parameters of the spam senders are suffciently similar. individual users over the repetitions of the spam game.This average filter will be

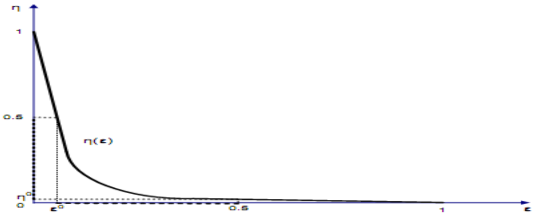

Figure 3. ε-η curve for a 'good' filter

characterized by a fixed ε-η curve similar to those of figures 2 and 3. Player II,then, can only select an ε value, and the correspondingη value is determined by the ε-η curve.

5. Conclusion

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 tradeof 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 tradeof, which e-mail users should adopt, and we provided a prediction of the eventual percentage of e-mail trafic that will be spam if the optimal point is adopted. Determining the tradeof'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.

6. Reference

[1] 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.

[2] United Nations Conference on Trade and Development (2003). E-Commerce and Development Report 2003 Internet edition prepared by the UNCTAD secretariat Chapter 3: ICT strategies for development.

[3] Philippe Gerard (2005)'Co-operating internationally a gainst spam' -ASEM London 4th Conference on eCommerce Tackling Spam.

[4] MessageLabs. Intelligence Annual Email Security Report 2004. http://www.messagelabs.com/binaries/LAB480 endofyear v2.pdf.

[5] The Spamhaus Project. Increasing Spam Threat from Proxy Hijackers. www.spamhaus.org/news.lasso?article=156.

[6] Nicola Lugaresi. European Union vs. Spam: A Legal Response. In Proceedings of the First Conference on E-mail and Anti-Spam, 2004.