

Voluntary Delay in Communication Buffers to Reduce Spam

Benjamin Chiao and Guang Shi

Peking University

benjamin.chiao@gmail.com, shiguangpku@gmail.com

Jan. 10, 2010

Updated version:

<http://benjaminchiao.org/papers/buffer.pdf>

Abstracts: Some communication methods are close to instantaneous but this is not entirely necessary. We evaluate a patent pending technology that provides an option for senders to voluntarily delay delivery of their messages. The delayed messages will be stored in a buffer, in which malicious messages could be tagged. Senders could signal credibility by choosing to use the buffer (or more generally a signaling time). Recipients could set a waiting time as a screen; only messages that have been buffered longer than the waiting time will be delivered to them. Recipients, on receiving a message, will be notified about whether the message is tagged as spam in the buffer, and the time length in which the messages have been stored in the buffer without being tagged. Recipients could make use of the buffered time to tell whether a message is likely to be malicious or not. We characterize conditions for some separating and pooling pure strategy Nash equilibria and for recipients to be better off.

Acknowledgments: We especially want to thank Jeffrey MacKie-Mason who offered guidance in the early phase of this work. We also benefited from the discussions with Rahul Telang, Michael Smith, members of the Incentives-Centered Design Lab at the University of Michigan and of the Information Society group at the Peking University Guanghua School of Management.

1. INTRODUCTION

We evaluate theoretically an anti-spam mechanism called communication buffers. This mechanism also applies to communication methods other than electronic mail. It is an implementable technology pending patent approval. The patent pending number is PCT/CN2009/071969.

Some communication methods are close to instantaneous but this is not entirely necessary. For example, commercial solutions abound in which firms take advantage of previous electronic mail circulating on the network to identify future spam¹. The problem is that they need to let go of a lot of spam to generate enough statistics over time to identify and remove spam. In the anti-spam mechanism we study in this paper, there is an option for a sender to voluntarily delay the delivery of his messages to signal credibility by selecting whether to use a buffer. A recipient has to select her waiting time to screen messages. It is the time length she would like messages addressed to her to be delayed before being forwarded to her. The delayed messages will be stored in a buffer. In the buffer, malicious messages are more likely to be detected as spam as time increases. Hereafter, for clarity, senders are male and recipients are female, both in singular or plural forms.

Buffered messages will be forwarded to the recipient at end of the waiting time. The recipient will be notified two pieces of information of a message. The first is whether the message is forwarded from the buffer. The second is whether the message has been tagged as spam in the buffer. Recipients could make use of such information to help decide whether a message is likely to be spam or not.

There are a few works that study the spam problem using traditional economic frameworks. Pavlov et al. (2005) view the unsolicited commercial email market as a tragedy of commons. Melville et al. (2006) continue this prior study using simulation. Pavlov et al. (2008) model the underlying dynamics of the email marketing ecosystem. All of these studies conclude that filtering spam could lead to the unintended consequence of increasing the spam volume.

Some research works go further to propose mechanisms to solve the spam problems. Fahlman (2002) proposes a mechanism called “selling interrupt rights”, where a sender interrupting a recipient by sending a message could pay the recipient. Hermalin and Katz (2004) study whether the senders or recipients should pay for messages. Kraut et al. (2005) test experimentally the effectiveness of the email stamps. Loder et al. (2006) propose that the senders could send a bond to a third party, which will only be refunded if the recipients approve. van Zandt (2004) proposes to make use of a Vickrey auction to price recipients’ attention. von Ahn et al. (2003) analyze the Challenge-Response system, under which unrecognized senders have to pass a test in order to deliver their messages. Chiao and MacKie-Mason (2006) propose an uncensored communication channel to entice the spam demanders and suppliers to trade in that channel.

None of these prior works, however, focus on using time to solve the information asymmetry problem in communication.

This paper is organized as follows. In section 2 we of-

¹Ironport, for example, monitors a quarter of the world’s email and web traffic from over 100,000 participating organizations to identify spam. Url: http://www.ironport.com/technology/ironport/_senderbase/network.html.

fer a theoretical model of voluntary delay in communication buffers and derive all the pure strategy Nash equilibria when senders or recipients (or both) are sufficiently patient. Section 3 is welfare comparison. Implementation of the mechanism is discussed in Section 4. Section 5 concludes.

2. THEORY

We use a representative agent framework to analyze two major stakeholders: senders and recipients. It involves value judgment to say that a message or its sender is universally hated. We do not give a universal definition of spam here. A spam message in this paper is one in which it has been classified as spam in the buffer².

2.1 Basic Setup

2.1.1 Senders

In this model, the senders choose the message volume and whether to use a buffer. We divide senders, indexed by s , into two types: spammers (denoted as the set \otimes) and non-spammers. Spammers are distributed on the interval $[0,1]$ with the probability distribution function f . Non-spammers are distributed on the interval $[0,1]$ with the probability distribution function g . Let $N_s \geq 0$ be the volume of messages sent by sender s . $\phi_s = 1$ indicates that a buffer will be used, else $\phi_s = 0$. $\rho \in [0,1]$ is the sender's discount rate. $f_s \in [1, \infty)$ is the filter strength sender s perceives his messages will be subject to if a buffer is not used. For each message responded, sender s receives p_s dollars. The price p_s is normalized to one now.

$B(N_s, t)$ is the total volume of sender s 's messages delivered to the recipient but not yet tagged as spam in the buffer by the buffered time t , which is set by the recipient as her waiting time in the model. Its properties are listed in Assumption 1:

ASSUMPTION 1. (*Buffer Technology*) The buffer *i*) outputs more if there is more messages sent to it ($\frac{\partial B(N_s, t)}{\partial N_s} \geq 0$), *ii*) needs time to detect spam ($B(N_s, 0) = N_s, \forall s$), *iii*) detects more spam as time increases ($\frac{\partial B(N_s, t)}{\partial t} \leq 0$), *iv*) makes no mistake in detecting non-spammers ($B(N_{s \notin \otimes}, t) = N_{s \notin \otimes}, \forall t, \forall s \notin \otimes$)³.

For tagged messages, we assume that the corresponding response rate of recipients is zero. This is consistent with *iv*) of the above assumption. For untagged messages, let $\theta(\phi_s, t, \kappa) \in [0, 1]$ be the corresponding response rate. It depends on whether a buffer is used by sender s , ϕ_s , the waiting time t , and the number of recipients reading unbuffered messages κ . Its properties are listed in Assumption 2:

ASSUMPTION 2. (*Response Rate*) The response rate is *i*) higher when the buffer is used ($\theta(1, t, \kappa) > \theta(0, t, \kappa)$), *ii*) increasing in waiting time ($\frac{\partial \theta(\phi_s, t, \kappa)}{\partial t} \geq 0$), *iii*) increasing in the number of recipients reading unbuffered messages ($\frac{\partial \theta(\phi_s, t, \kappa)}{\partial \kappa} > 0$), *iv*) zero when the buffer is not used and no recipient reads unbuffered messages ($\theta(0, t, 0) = 0$).

²See the Implementation section for some of the means for classification.

³(*iv*) is of course not realistic. Type II errors are common in filtering system. However, when option D in the Implementation section is turned on, (*iv*) is becoming more realistic.

(*i*) above is consistent with that using the buffer increases the credibility of a message. When the buffer is used, a recipient is more willing to respond as time increases. This could be true if a recipient has more confidence about the true identity of a sender as buffer time increases. This does not hold if the value of a message is perishable.

$C(N_s)$ is the cost function of sending messages. Its properties are listed in Assumption 3:

ASSUMPTION 3. (*Sending Cost*) *i*) The sending cost is convex ($C'(N_s) \geq 0; C''(N_s) \geq 0$). *ii*) There is no fixed cost ($C(0) = 0$).

Message sending is not costless. Although replication and transport costs are low especially for electronic messages, there are many other costs. For example, it takes costs to gather the information of targeted recipients. Content creation and disguising costs are usually high for spam, since they need to pass through many anti-spam filters. Message sending also involves high labor costs or spambot costs, because most spambots are not owned by senders themselves. Also, spam sending may incur potential severe legal penalties.

Sender s chooses (N_s, ϕ_s) to maximize the expected profit:

$$\pi_s(N_s, \phi_s) = \theta(\phi_s, t, \kappa) (\rho^t B(N_s, t))^{\phi_s} \left(\frac{N_s}{f_s}\right)^{1-\phi_s} - C(N_s) \quad (1)$$

The first term on the right hand side of (1) is the sender's revenue. It is the product of price, response rate and the volume of message received, the latter of which depends on whether a buffer is used. When the buffer is used, sender's profit is time-discounted because his messages are forwarded to recipients at the end of the waiting time t .

2.1.2 Recipients

There are infinite homogeneous recipients, indexed by r , distributed on the interval $[0,1]$ with the probability distribution function h . Let $N_r^B \geq 0$ be the volume of buffered messages addressed to recipient r , $N_r^U \geq 0$ be the volume of unbuffered messages addressed to r . $N_r = N_r^B + N_r^U$ is the total volume of messages addressed to r . $t \geq 0$ is recipient r 's waiting time for buffered messages addressed to her. $\delta \in [0, 1]$ is her discount rate, and $\kappa_r \in \{0, 1\}$ is recipient r 's decision variable of whether to read unbuffered messages. Denote \bar{t} as the time required to tag all the spam in N_r . $B(N_r^B, \bar{t})$ is the volume of buffered non-spam addressed to r . N_r can then be classified into:

- Buffered messages: N_r^B
 - Non-spam: $B(N_r^B, \bar{t})$
 - Spam: $N_r^B - B(N_r^B, \bar{t})$
 - * Untagged: $B(N_r^B, t) - B(N_r^B, \bar{t})$
 - * Tagged: $N_r^B - B(N_r^B, t)$
- Unbuffered messages: N_r^U

The recipient's problem is to choose the waiting time, t , for buffered messages and whether to read unbuffered message, κ_r , to maximize her utility:

$$U_r = U_r(v_r^a, v_r^b, \kappa_r v_r^c) \quad (2)$$

where:

- $v_r^a = \delta^t B(N_r^B, \bar{t})$ measures the time-discounted utility from wanted buffered messages *read*, where $B(N_r^B, \bar{t})$ is the volume of buffered non-spam addressed to r
- $v_r^b = B(N_r^B, t) - B(N_r^B, \bar{t})$ is the volume of unwanted buffered messages *read*. These are the Type I errors in the buffer
- $v_r^c = (a - a^\lambda)N_r^U$. $a \in [1, \infty)$ is a constant, which measures the recipient's preference of spam of the unbuffered messages. The recipient extremely hates spam if $a \rightarrow 1$, under which $v_r^c \rightarrow 0$ for $\forall \lambda > 0$, where $\lambda \in [0, 1]$ is the ratio of spam in the unbuffered messages. v_r^c is decreasing in λ . When all the unbuffered messages are spam ($\lambda = 1$), $v_r^c = 0$. When all the unbuffered messages are not spam ($\lambda = 0$), $v_r^c = aN_r^U$, which measures the utility of wanted unbuffered messages. When $\lambda \in (0, 1)$, $0 \leq v_r^c \leq aN_r^U$.

By definition of the desirability of v_r^a , v_r^b and v_r^c , we have $\frac{\partial U_r}{\partial v_r^a} > 0$ and $\frac{\partial U_r}{\partial v_r^b} < 0$.

2.2 Scenarios

We consider two scenarios: status quo and voluntary buffer. Status quo (SQ) refers to the current messaging system without buffers. Voluntary buffer (VB) refers to the case which a buffer is optional for senders. One key difference for these two cases lies in the response rates. In this section, we analyze the best responses of senders and recipients.

2.2.1 Status Quo

In this scenario there is no buffer as an option for the sender. Senders need not to choose whether to use a buffer. Recipients need not to choose the waiting time. We denote $\phi_s = 0$ and $t = 0$ for notational consistency. Thus the response rate of recipients in the status quo can be written as $\theta(\phi_s, t, \kappa) = \theta(0, 0, \kappa)$. Under the status quo, senders only choose the volume of messages sent, N_s , and recipients only choose whether to read the messages received⁵, κ_r . We derive the best responses of senders and recipients in Results 1 and 2.

RESULT 1. (*Sender's BR, SQ*) The best response for sender s is $\{N_s^* \text{ s.t. } \frac{\theta(0,0,\kappa)}{f_s} = C'(N_s^*)\}$.

PROOF. See Appendix 6.1. \square

RESULT 2. (*Recipient's BR, SQ*) The best response for recipient r is $i) \kappa_r^* = 1$ if $\lambda \in [0, 1)$, $ii) \kappa_r^* = 0$ or 1 if $\lambda = 1$.

PROOF. See Appendix 6.2. \square

Result 2 implies that recipient r will read the received messages if not all of them are spam. Otherwise, it is indifferent for her to read them or not.

⁴One could alternatively define v_r^c as consisting of two categories of messages: the volume of wanted unbuffered messages *read*, and the volume of unwanted unbuffered messages *read*. Since these messages are unbuffered, the recipients cannot separate the two categories.

⁵Under the status quo, all the received messages are unbuffered since there is no buffer option. Thus recipient's choice variable κ_r is consistent with its definition in the basic setup.

2.2.2 Voluntary Buffer

In this scenario, senders have the option to choose whether to use a buffer. Recipients can choose the waiting time to screen spammers. Recipients have the option to read buffered, unbuffered messages or both. In other words, all messages might be responded ($\theta(\phi_s, t, \kappa) \geq 0, \forall \phi_s, \forall t, \forall \kappa$). We derive the best responses of non-spammers, spammers and recipients in Results 3, 4 and 5.

RESULT 3. (*Non-Spammer's BR, VB*) The best response for sender $s \notin \otimes$ is $\{\phi_{s \notin \otimes}^* = 1 \text{ and } N_{s \notin \otimes}^* \text{ s.t. } \theta(1, t, \kappa)\rho^t = C'(N_{s \notin \otimes}^*)\}$ if sender s is sufficiently patient.

PROOF. See Appendix 6.3. \square

Result 3 indicates that under the voluntary buffer, if a non-spammer is sufficiently patient, his best response is to use a buffer and the optimal message volume is $N_{s \notin \otimes}^*$ above.

RESULT 4. (*Spammer's BR, VB*) The best responses for sender $s \in \otimes$ are

- $\{\phi_{s \in \otimes}^* = 0 \text{ and } N_{s \in \otimes}^* \text{ s.t. } \frac{\theta(0,t,\kappa)}{f_s} = C'(N_{s \in \otimes}^*)\}$, if $\frac{\theta(0,t,\kappa)}{f_s} > \theta(1, t, \kappa)\rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*}$,
- $\{\phi_{s \in \otimes}^* = 1 \text{ and } N_{s \in \otimes}^* \text{ s.t. } \theta(1, t, \kappa)\rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*} = C'(N_{s \in \otimes}^*)\}$,
- $\{\phi_{s \in \otimes}^* = 0 \text{ or } 1 \text{ and } N_{s \in \otimes}^* \text{ s.t. } \frac{\theta(0,t,\kappa)}{f_s} = \theta(1, t, \kappa)\rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*} = C'(N_{s \in \otimes}^*)\}$, otherwise.

PROOF. See Appendix 6.4. \square

Result 4 says that the spammer's best responses above depend on the relationship between $\frac{\theta(0,t,\kappa)}{f_s}$ and $\theta(1, t, \kappa)\rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*}$. $N_{s \in \otimes}^*$ in the latter term is spammer's optimal message volume when the buffer is used, which satisfies the FOC (20)⁶. The condition implies that when choosing whether to use the buffer, the spammer will consider the trade-off among the response rate, unfiltered rate and discount rate⁷.

The recipient chooses the waiting time t for buffered messages and whether to read unbuffered messages κ_r . We have the following results:

RESULT 5. (*Recipient's BR, VB*) The set of best responses for recipient r is

- $\{t^* \text{ s.t. } \frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) B(N_r^B, \bar{t}) + \frac{\partial U_r}{\partial v_r^b} \frac{\partial B(N_r^B, t^*)}{\partial t^*} = 0 \text{ and } \kappa_r^* = 1\}$ if $N_r^U > 0$ and $\lambda \in [0, 1)$,
- $\{t^* \text{ s.t. } \frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) B(N_r^B, \bar{t}) + \frac{\partial U_r}{\partial v_r^b} \frac{\partial B(N_r^B, t^*)}{\partial t^*} = 0 \text{ and } \kappa_r^* = 0 \text{ or } 1\}$ if $N_r^U = 0$, or if $N_r^U > 0$ and $\lambda = 1$.

⁶To avoid confusion, we emphasize here that it is the $N_{s \in \otimes}^*$ that satisfies the FOC (20), instead of (18). We do not further distinguish the notations of spammer's optimal message volume when the buffer is used and when the buffer is not used for notational consistency. Else, we can distinguish by defining $N_{s \in \otimes}^{U*}$ in equation (18) and $N_{s \in \otimes}^{B*}$ in equation (20) for clarity. The superscript U/B indicates whether the buffer is used.

⁷When the buffer is not used, the response rate is $\theta(0, t, \kappa)$, the unfiltered rate is $\frac{1}{f_s}$, and there is no time discount. When the buffer is used, the response rate is $\theta(1, t, \kappa)$ the unfiltered rate is $\frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*}$, and the time discount is ρ^t .

PROOF. See Appendix 6.5. \square

Result 5 implies that if the spam ratio in the unbuffered messages is strictly less than 1, the recipient's best response is to read them since there are some valuable messages. However, if there is no unbuffered message, or if all the unbuffered messages are spam, it is indifferent for her to read them or not. Her optimal waiting time for the buffered messages is t^* .

2.3 Competitive Equilibrium

A competitive equilibrium is defined to satisfy:

- The utility of each recipient is maximized:

$$U_r^*(t^*, \kappa_r^*) = \underset{\{t, \kappa_r\}}{\text{Max}} U_r(v_r^a, v_r^b, \kappa_r v_r^c) \quad (3)$$

- The profit of each sender is maximized:

$$\pi_s^*(N_s^*, \phi_s^*) = \underset{\{N_s, \phi_s\}}{\text{Max}} \pi_s(N_s, \phi_s) \quad (4)$$

- The supply of and demand for messages are equal:

$$\int_0^1 N_{s \in \otimes}^* f(s) ds + \int_0^1 N_{s \notin \otimes}^* g(s) ds = \int_0^1 (N_r^{U^*} + N_r^{B^*}) h(r) dr \quad (5)$$

By homogeneity of each agent type, we have the following aggregation rule on the unit interval:

$$\int_0^1 x_i f(i) di = x \int_0^1 f(i) di = x. \quad (6)$$

That is, the aggregate value equals to the average value. Thus the above condition is equivalent to:

$$N_{s \in \otimes}^* + N_{s \notin \otimes}^* = N_r^{U^*} + N_r^{B^*} \equiv N_r^* \quad (7)$$

- The number of recipients reading unbuffered messages is aggregated across all recipients:

$$\kappa = \int_0^1 \kappa_r h(r) dr = \kappa_r \text{ following (6)}$$

- Assuming free-entry, senders obtain zero profits:

$$\pi_s^*(N_s^*, \phi_s^*) = 0, \forall s \quad (8)$$

With the above equilibrium conditions and the intersection of the best responses presented, we derive the pure strategy Nash equilibria when senders or recipients (or both) are sufficiently patient in the two scenarios:

PROPOSITION 1. (NE, SQ) *The pure strategy Nash equilibrium under the status quo is $\{N_s^* \text{ s.t. } \frac{\theta(0,0,1)}{f_s} = C'(N_s^*) \forall s; \kappa_r^* = 1\}$.*

PROOF. See Appendix 6.6. \square

The Nash equilibrium in Proposition 1 implies that senders cannot signal credibility by using a buffer since there is no such option, while recipients can only passively read all messages (including spam) received.

When senders or recipients (or both) are sufficiently patient, there are two sets of pure strategy Nash equilibria under the voluntary buffer. One set is separating equilibria and the other is pooling equilibria.

PROPOSITION 2. (NE1, VB) *If sender $s \notin \otimes$ is sufficiently patient and $\frac{\theta(0,t^*,1)}{f_s} > \theta(1,t^*,1)\rho^{t^*} \frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*}$, the set of pure strategy separating Nash equilibria under the voluntary buffer scenario is $\{\phi_{s \notin \otimes}^* = 1 \text{ and } N_{s \notin \otimes}^* \text{ s.t. } \theta(1,t^*,1)\rho^{t^*} = C'(N_{s \notin \otimes}^*) \text{ for } s \notin \otimes; \phi_{s \in \otimes}^* = 0 \text{ and } N_{s \in \otimes}^* \text{ s.t. } \frac{\theta(0,t^*,1)}{f_s} = C'(N_{s \in \otimes}^*) \text{ for } s \in \otimes; \kappa_r^* = 1 \text{ and } t^* \text{ s.t. } \frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) N_{s \notin \otimes}^* = 0\}$.*

PROOF. See Case 1 in Appendix 6.7. \square

The separating equilibria imply that if recipients choose to read unbuffered messages, spammers will not use the buffer, while non-spammers will use the buffer under the equilibrium existence conditions. Thus recipients can discern spam depending on whether the buffer is used⁸.

PROPOSITION 3. (NE2, VB) *If sender s is sufficiently patient, the set of pure strategy pooling Nash equilibria under the voluntary buffer scenario is $\{\phi_{s \notin \otimes}^* = 1 \text{ and } N_{s \notin \otimes}^* \text{ s.t. } \theta(1,t^*,0)\rho^{t^*} = C'(N_{s \notin \otimes}^*) \text{ for } s \notin \otimes; \phi_{s \in \otimes}^* = 1 \text{ and } N_{s \in \otimes}^* \text{ s.t. } \theta(1,t^*,0)\rho^{t^*} \frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*} = C'(N_{s \in \otimes}^*) \text{ for } s \in \otimes; \kappa_r^* = 0 \text{ and } t^* \text{ s.t. } \frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) B(N_r^{B^*}, \bar{t}) + \frac{\partial U_r}{\partial v_r^b} \frac{\partial B(N_r^{B^*}, t^*)}{\partial t^*} = 0\}$.*

PROOF. See Case 2 in Appendix 6.7. \square

The pooling equilibria indicate that if recipients choose not to read unbuffered messages, both spammers and non-spammers will choose to use buffer under the equilibrium existence condition. So spammers cannot be separated from non-spammers. However, the message volume sent by spammers is smaller than non-spammers. If the filter strength of the buffer is strong enough ($\frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*} \rightarrow 0$), spammers will choose not to send messages at all ($N_{s \in \otimes}^* \rightarrow 0$), because all spam are filtered out by the buffer. This can be interpreted as a degenerated case. It is equivalent to the scenario that there is no buffer, but the filter strength of the message server is so strong that all spam are filtered out. Spam disappears in equilibrium.

Since $C'(N_s)$ is monotone and if sender s is sufficiently patient, the pure strategy Nash equilibrium above is unique under the status quo. However, the Nash equilibrium might not be unique in the voluntary buffer scenarios because there could be multiple t^* .

3. WELFARE COMPARISON

We consider the welfare implications of the above Nash equilibria. In the voluntary buffer scenario, we only consider the separating equilibria above since the pooling equilibria is a degenerated case. We only need to compare the welfare of recipients because senders always obtain zero profits in the competitive equilibrium.

PROPOSITION 4. (Welfare, VB v. SQ) *The utility of recipient r is weakly higher under the voluntary buffer scenario*

⁸We do not consider mixed strategies for senders in the voluntary buffer scenario because of two reasons. First, if automatic message forwarding technologies are more available so that recipients will forward all messages to the buffer themselves, this is equivalent to the pure strategy Nash equilibria. Second, if they don't, the mixed strategies are going to depend on the distribution of the discount rate, which greatly complicates the analysis.

than that under the status quo if i) sender $s \notin \otimes$ and recipient r are sufficiently patient and ii) recipient r hates spam sufficiently.

PROOF. See Appendix 6.8. \square

4. IMPLEMENTATION

We provide the benchmark model above to prove the effectiveness of voluntary delay in reducing spam and improving recipient’s welfare. There are some extensions in implementing the mechanism. Here is a sketch of several implementation methods depending on which options are turned on. We have analyzed the mechanism above without turning on any options below. Option B empowers senders to signal by choosing a continuous time variable instead of whether to use the buffer. Turning on options C and D could introduce more errors in spam detection but possibly increases the speed of detection—the net result of such trade-off is largely an empirical issue so we do not study its theoretical properties here.

1. Suppose sender eBay wants to send an email to customer@gmail.com and the recipient customer@gmail.com has set her waiting time.

Figure 1. A Hypothetical Screenshot for Senders

2. Suppose the buffer option has already been built in the email system. To use a buffer, eBay sends the mail to customer@gmail.com, and clicks on the buffer button. See Figure 1 for the screenshot.

Figure 2. A Hypothetical Screenshot for Senders under Option A

3. Option A: suppose the buffer option is not built in the current email system, the buffer function can be implemented in this way: eBay sends the mail to customer@gmail.com, and carbon copies it to a buffer of its choice⁹, such as customer_at_gmail.com@ebuffer.com¹⁰, where ebuffer.com is an independent buffer service provider¹¹. See Figure 2 for the

⁹Choosing which buffer company to send to allows competition between buffers.

¹⁰The buffer’s name rule for each recipient is the recipient’s email address followed by ”@ebuffer.com”. It ensures that there is a credible one-to-one correspondence in the buffer for each recipient.

¹¹Under Option A, one problem is that recipients will receive the same message twice. The first is sent by the message server instantaneously. The second is sent with the tags by the buffer after delay. However, it can be technologically solved by deleting the instantaneous message on receiving the buffered message.

case.

4. Option B: Instead of only choosing whether to use buffer, the sender could specify the signaling time, say 20 minutes¹², using syntax like:

customer_at_gmail.com_20min@ebuffer.com in Step 3¹³. The buffer time (t_b) is some function of the recipient’s waiting time (t_r) and the sender’s signaling time (t_s), for example, $t_b = \min\{t_r, t_s\}$.

5. The buffer starts analyzing the mail for malicious content and tag mail as spam as it sees fit during its detection process. There are various ways for the buffer to detect spam, including technological filtering and non-technological solutions, such as Option C and D.

6. Option C: All recipients are entitled to tag the message.

7. Option D: Since existing tools allow senders to change their names on the email headers, the buffer does not know if the message is really from eBay. The buffer will confirm receipt of the message if eBay elects beforehand to receive a notification when a message in the buffer is purported to have been sent from it and tag such purported messages as spam as it sees fit (alternatively, eBay could be given an account to log in the buffer without notifications). eBay could in addition give permission to anyone eBay wants to help tag messages.

8. The buffer will forward the email at the end of the buffer time and inform the recipients its judgement of the message based on its analysis during the delay. The mail is tagged ”Risky” if the buffer believes it is spam. It is tagged ”Clean” otherwise.

Figure 3. A Hypothetical Screenshot for Recipients

9. On the recipient’s side, she has the option to choose or change her waiting time. She will receive the buffered mail with the tags after Step 8. See Figure 3. The ”Risky” and ”Clean” tags could help her to discern spam. There is no tag if a sender did not choose to use buffer. It is up to her to read it or not.

5. CONCLUSION

We have evaluated an anti-spam mechanism that utilizes time in a buffer to facilitate the removal of spam. Our key insight relies on the observation that it is not necessary for messages to be sent instantaneously. A voluntary delay could signal that a sender is credible since malicious messages are more likely to be detected in a buffer over time.

In our model, we provide the conditions under which the welfare of recipients will improve in the voluntary delay. The most important result we have proved is that under certain conditions the set of separating pure strategy Nash equilibria [when senders or recipients (or both) are sufficiently patient] exists if a buffer is provided as an option. That is, spammers will choose not to use a buffer and non-spammers

¹²Allowing senders to choose the signaling time is an extension to the basic model, where he could only choose whether to use buffer. The signal now becomes continuous instead of discrete.

¹³It can also be implemented in Step 2 by replacing the buffer button in Figure 1 with a text box where a sender could type in his signaling time.

will choose to use a buffer. Thus when a recipient receives an unbuffered message, she will immediately know that the message is spam. However, it is up to her to decide whether to read it, depending on her own preference.

There are several key limitations. First, if the impatient factor of each recipient varies, we might not be able to separate spammers from non-spammers. This is because spammers might still buffer the messages a short time and the impatient recipients could mistakenly treat them as non-spam. Second, we did not consider the effects of whether to read unbuffered messages on the recipient's response rate wrt buffered messages. If a recipient chooses not to read unbuffered messages, her response rate wrt buffered but untagged messages could possibly be higher under budget or time constraints. Third, we ruled out the case in which a sender could divide the message into two parts (such as header and body), which the recipient could preview one part to infer the other so the impatient factor of the recipients might be affected. For example, a recipient might be attracted by the email header to read the whole email message. Lastly, we have not analyzed mixed strategies. This is because we have argued there is no need to model a large class of them especially when automatic message forwarding technologies become more available.

6. APPENDICES

6.1 Proof of Result 1

For sender s , no matter $s \in \otimes$ or $s \notin \otimes$, the maximization problem becomes

$$\text{Max}_{N_s} \pi_s(N_s, 0) = \theta(0, 0, \kappa) \frac{N_s}{f_s} - C(N_s) \quad (9)$$

$$\text{FOC} : \frac{\theta(0, 0, \kappa)}{f_s} = C'(N_s^*) \quad (10)$$

We rule out the corner solution ($N_s = 0$) because it happens only when $f_s \rightarrow \infty$, in which case no buffer is necessary in the first place as all spam are filtered out.

6.2 Proof of Result 2

For recipient r , the maximization problem is

$$\text{Max}_{\kappa_r} U_r = U_r(v_r^a, v_r^b, \kappa_r v_r^c) = U_r(0, 0, \kappa_r(a - \alpha^\lambda)N_r^U) \quad (11)$$

Since $\frac{\partial U_r}{\partial v_r^c} > 0$, and v_r^c brings zero utility to r if $v_r^c = 0$, there are two cases. If $\lambda \in [0, 1)$, $\kappa_r^* = 1$. Else, $\kappa_r^* = 0$ or 1 .

6.3 Proof of Result 3

For sender $s \notin \otimes$, if $\phi_{s \notin \otimes} = 0$,

$$\text{Max}_{N_{s \notin \otimes} \geq 0} \pi_s(N_{s \notin \otimes}, \phi_{s \notin \otimes}) = \pi_s(N_{s \notin \otimes}, 0) = \theta(0, t, \kappa) \frac{N_{s \notin \otimes}}{f_s} - C(N_{s \notin \otimes}) \quad (12)$$

$$\text{FOC} : \frac{\theta(0, t, \kappa)}{f_s} = C'(N_{s \notin \otimes}^*) \quad (13)$$

else if $\phi_{s \notin \otimes} = 1$,

$$\begin{aligned} \text{Max}_{N_{s \notin \otimes} \geq 0} \pi_s(N_{s \notin \otimes}, \phi_{s \notin \otimes}) &= \pi_s(N_{s \notin \otimes}, 1) = \theta(1, t, \kappa) \rho^t B(N_{s \notin \otimes}, t) - C(N_{s \notin \otimes}) \\ &= \theta(1, t, \kappa) \rho^t N_{s \notin \otimes} - C(N_{s \notin \otimes}), \text{ by Assumption 1 (iv)} \end{aligned} \quad (14)$$

$$\text{FOC} : \theta(1, t, \kappa) \rho^t = C'(N_{s \notin \otimes}^*) \quad (15)$$

Since $\theta(1, t, \kappa) \rho^t N_{s \notin \otimes}$ and $\theta(0, t, \kappa) \frac{N_{s \notin \otimes}}{f_s}$ are linear in $N_{s \notin \otimes}$ and $C(N_{s \notin \otimes})$ is convex, $\pi_s(N_{s \notin \otimes}^*, 1)$ being the vertical distance between these two curves is larger than $\pi_s(N_{s \notin \otimes}^*, 0)$ if the following condition holds:

$$\theta(1, t, \kappa) \rho^t > \frac{\theta(0, t, \kappa)}{f_s} \quad (16)$$

By Assumption 2 (i) and $f_s \geq 1$, condition (16) holds if sender s is sufficiently patient ($\rho \rightarrow 1$).

6.4 Proof of Result 4

For sender $s \in \otimes$, if $\phi_{s \in \otimes} = 0$,

$$\text{Max}_{N_{s \in \otimes} \geq 0} \pi_s(N_{s \in \otimes}, \phi_{s \in \otimes}) = \pi_s(N_{s \in \otimes}, 0) = \theta(0, t, \kappa) \frac{N_{s \in \otimes}}{f_s} - C(N_{s \in \otimes}) \quad (17)$$

$$\text{FOC} : \frac{\theta(0, t, \kappa)}{f_s} = C'(N_{s \in \otimes}^*) \quad (18)$$

else if $\phi_{s \in \otimes} = 1$,

$$\text{Max}_{N_{s \in \otimes} \geq 0} \pi_s(N_{s \in \otimes}, \phi_{s \in \otimes}) = \pi_s(N_{s \in \otimes}, 1) = \theta(1, t, \kappa) \rho^t B(N_{s \in \otimes}, t) - C(N_{s \in \otimes}) \quad (19)$$

$$\text{FOC} : \theta(1, t, \kappa) \rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*} = C'(N_{s \in \otimes}^*) \quad (20)$$

There are three cases. $\pi_s(N_{s \in \otimes}^*, 0) > \pi_s(N_{s \in \otimes}^*, 1)$ if $\frac{\theta(0, t, \kappa)}{f_s} > \theta(1, t, \kappa) \rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*}$, where $N_{s \in \otimes}^*$ satisfies condition (20).

Else if $\frac{\theta(0, t, \kappa)}{f_s} < \theta(1, t, \kappa) \rho^t \frac{\partial B(N_{s \in \otimes}^*, t)}{\partial N_{s \in \otimes}^*}$, $\pi_s(N_{s \in \otimes}^*, 0) < \pi_s(N_{s \in \otimes}^*, 1)$. Else $\pi_s(N_{s \in \otimes}^*, 0) = \pi_s(N_{s \in \otimes}^*, 1)$.

6.5 Proof of Result 5

For recipient r ,

$$\text{Max}_{\{t, \kappa_r\}} U_r(v_r^a, v_r^b, \kappa_r v_r^c) = U_r(\delta^t B(N_r^B, \bar{t}), B(N_r^B, t) - B(N_r^B, \bar{t}), \kappa_r(a - \alpha^\lambda)) \quad (21)$$

$$\text{FOC w.r.t. } t : \frac{\partial U_r}{\partial v_r^a} \delta^{t*} (\ln \delta) B(N_r^B, \bar{t}) + \frac{\partial U_r}{\partial v_r^b} \frac{\partial B(N_r^B, t^*)}{\partial t^*} = 0 \quad (22)$$

There are three cases for κ_r^* :

If $N_r^U = 0$, that is, there is no unbuffered message, $\kappa_r^* = 0$ and $\kappa_r^* = 1$ are indifferent for the recipient since $v_r^c = 0$.

If $N_r^U > 0$ and $\lambda = 1$, $\kappa_r^* = 0$ and $\kappa_r^* = 1$ are indifferent for the recipient since $v_r^c = 0$.

If $N_r^U > 0$ and $\lambda \in [0, 1)$, $\kappa_r^* = 1$ is the recipient's dominant strategy since $v_r^c > 0$.

6.6 Proof of Proposition 1

There are two cases:

Case 1: if $\kappa_r^* = 1$, the response rate is $\theta(0, 0, 1)$, which is positive following Assumption 2 (iii) and (iv). The best response of sender s (both $s \in \otimes$ and $s \notin \otimes$) is N_s^* satisfying $\frac{\theta(0, 0, 1)}{f_s} = C'(N_s^*)$, from which we have $N_s^* > 0$. So

$\lambda = \frac{\int_0^1 N_{s \in \otimes}^* f(s) ds}{C(N_{s \in \otimes}^*) + \int_0^1 N_{s \notin \otimes}^* g(s) ds} = \frac{1}{2}$ since $N_{s \in \otimes}^* = N_{s \notin \otimes}^* = N_s^*$. Given $\lambda = \frac{1}{2}$, recipient r 's best response is $\kappa_r^* = 1$. Therefore, $\{N_s^* \text{ s.t. } \frac{\theta(0, 0, 1)}{f_s} = C'(N_s^*) \forall s; \kappa_r^* = 1\}$ is a Nash equilibrium.

Case 2: if $\kappa_r^* = 0$, the response rate is $\theta(0, 0, 0)$, which is zero following Assumption 2 (iv). The best response of sender s (both $s \in \otimes$ and $s \notin \otimes$) is $N_s^* = 0$. That is, no sender sends messages and no recipient reads messages. The market disappears. We do not consider this case since it is trivial.

6.7 Proof of Proposition 2 and 3

There are two cases:

Case 1: Given t^* , if $\kappa_r^* = 1$, we have the following results. Non-spammer s 's best response is $\phi_{s \notin \otimes}^* = 1$ and $N_{s \notin \otimes}^*$ s.t. $\theta(1, t^*, 1)\rho^{t^*} = C'(N_{s \notin \otimes}^*)$ if $s \notin \otimes$ is sufficiently patient. Spammer s 's best response is $\phi_{s \in \otimes}^* = 0$ and $N_{s \in \otimes}^*$ s.t. $\frac{\theta(0, t^*, 1)}{f_s} = C'(N_{s \in \otimes}^*)$ if $\frac{\theta(0, t^*, 1)}{f_s} > \theta(1, t^*, 1)\rho^{t^*} \frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*}$.¹⁴ It is straightforward that $N_{s \notin \otimes}^*, N_{s \in \otimes}^* > 0$.

Given $\{\phi_{s \notin \otimes}^* = 1, N_{s \notin \otimes}^* > 0; \phi_{s \in \otimes}^* = 0, N_{s \in \otimes}^* > 0\}$, $N_r^{B*} = N_{s \notin \otimes}^*$, $N_r^{U*} = N_{s \in \otimes}^* > 0$. All the unbuffered messages are spam, that is, $\lambda = \frac{\int_0^1 N_{s \in \otimes}^* f(s) ds}{\int_0^1 N_{s \in \otimes}^* f(s) ds + \int_0^1 0g(s) ds} = 1$. Recipient r 's best response is $\kappa_r^* = 0$ or 1 and t^* s.t. $\frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) N_{s \notin \otimes}^* = 0$.

With the intersection of the best responses above, we obtain the separating Nash equilibria in Proposition 2.

Case 2: Given t^* , if $\kappa_r^* = 0$, we have the following results. Non-spammer s 's best response is $\phi_{s \notin \otimes}^* = 1$ and $N_{s \notin \otimes}^*$ s.t. $\theta(1, t^*, 0)\rho^{t^*} = C'(N_{s \notin \otimes}^*)$ if $s \notin \otimes$ is sufficiently patient. If spammer $s \in \otimes$ is also sufficiently patient, his best response is $\phi_{s \in \otimes}^* = 1$ and $N_{s \in \otimes}^*$ s.t. $\theta(1, t^*, 0)\rho^{t^*} \frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*} = C'(N_{s \in \otimes}^*)$, since the condition $\theta(1, t^*, 0)\rho^{t^*} \frac{\partial B(N_{s \in \otimes}^*, t^*)}{\partial N_{s \in \otimes}^*} > \frac{\theta(0, t^*, 0)}{f_s} = 0$ holds following Assumption 2 (iv). It is straightforward that $N_{s \notin \otimes}^*, N_{s \in \otimes}^* > 0$.

Given $\{\phi_{s \notin \otimes}^* = 1, N_{s \notin \otimes}^* > 0; \phi_{s \in \otimes}^* = 1, N_{s \in \otimes}^* > 0\}$, there is no unbuffered message, that is, $N_r^{U*} = 0$. Recipient r 's best response is $\kappa_r^* = 0$ or 1 and t^* s.t. $\frac{\partial U_r}{\partial v_r^a} \delta^{t^*} (\ln \delta) B(N_r^{B*}, t^*) + \frac{\partial U_r}{\partial v_r^b} \frac{\partial B(N_r^{B*}, t^*)}{\partial t^*} = 0$, where $N_r^{B*} = N_{s \notin \otimes}^* + N_{s \in \otimes}^*$ following condition (7).

With the intersection of the best responses above, we obtain the pooling Nash equilibria in Proposition 3.

6.8 Proof of Proposition 4

From Proposition 1 and 2, the equilibrium message volume sent by the sender is:

$$N_s^{SQ*} \text{ s.t. } \frac{\theta(0, 0, 1)}{f_s} = C'(N_s^{SQ*}) \quad \forall s \text{ under SQ} \quad (23)$$

$$N_{s \notin \otimes}^{VB*} \text{ s.t. } \theta(1, t^*, 1)\rho^{t^*} = C'(N_{s \notin \otimes}^{VB*}) \quad \forall s \notin \otimes \text{ under VB} \quad (24)$$

$$N_{s \in \otimes}^{VB*} \text{ s.t. } \frac{\theta(0, t^*, 1)}{f_s} = C'(N_{s \in \otimes}^{VB*}) \quad \forall s \in \otimes \text{ under VB} \quad (25)$$

We add the superscript SQ/VB to N_s^* to avoid confusion. When ρ is sufficiently large, both $N_{s \notin \otimes}^{VB*}$ in (24) and $N_{s \in \otimes}^{VB*}$ in (25) are weakly larger than N_s^{SQ*} in (23) by Assumption 3 (i) and Assumption 2 (i), (ii).

¹⁴If the condition does not hold, there is a set of pooling equilibria which is the same as Case 2. We do not further discuss it here.

The equilibrium utility of the recipient under the status quo is:

$$U_r^{SQ*}(v_r^{a, SQ}, v_r^{b, SQ}, \kappa_r v_r^{c, SQ}) = U_r(0, 0, (a - a^{\frac{1}{2}})N_r^{SQ*}) \quad (26)$$

where $v_r^{a, SQ} = v_r^{b, SQ} = 0$ since there is no buffer under the status quo.

The equilibrium utility of the recipient under the voluntary buffer is:

$$U_r^{VB*}(v_r^{a, VB}, v_r^{b, VB}, \kappa_r v_r^{c, VB}) = U_r(\delta^{t^*} N_{s \notin \otimes}^{VB*}, 0, 0) \quad (27)$$

where $v_r^{a, VB} = \delta^{t^*} N_{s \notin \otimes}^{VB*}$ and $v_r^{b, VB} = 0$ since only non-spammers use a buffer under the voluntary buffer scenario in equilibrium and there is no Type II errors by Assumption 1 (iii). $v_r^{c, VB} = 0$ because all the unbuffered messages are spam ($\lambda = 1$).

It is straightforward that

$$v_r^{b, VB*} = v_r^{b, SQ*} = 0. \quad (28)$$

If the recipient is sufficiently patient ($\delta > 0$), we have

$$\delta^{t^*} N_{s \notin \otimes}^{VB*} = v_r^{a, VB*} > v_r^{a, SQ*} = 0. \quad (29)$$

However,

$$0 = v_r^{c, VB*} \leq v_r^{c, SQ*} = (a - a^{\frac{1}{2}})N_r^{SQ*}. \quad (30)$$

Thus the relationship between U_r^{VB*} and U_r^{SQ*} is determined by the marginal utility of buffered non-spam (condition (29)) and of unbuffered messages (condition (30)). If the recipient sufficiently hates the spam in the unbuffered messages ($a \rightarrow 1$), the marginal utility of buffered non-spam under the voluntary buffer ($v_r^{a, VB*}$) dominates that of unbuffered messages under the status quo ($v_r^{c, SQ*}$), and we obtain the result:

$$U_r^{VB*} \geq U_r^{SQ*}.$$

7. REFERENCES

- Benjamin Chiao and Jeffrey MacKie-Mason (2006). Using Uncensored Communication Channels to Divert Spam Traffic. *Net Institute Working Paper* No. 06-20. URL: <http://benjaminchiao.org/papers/openchannel.pdf>.
- Scott Fahlman. Selling Interrupt Rights: A Way to Control Unwanted E-Mail and Telephone Calls. *IBM Systems Journal*, 41 (4), pp.759-766, 2002.
- Benjamin Hermalin and Michael Katz. Sender or Receiver: Who Should Pay to Exchange an Electronic Message? *The RAND Journal of Economics*, 35 (3), pp.423-447, 2004.
- Robert Kraut, Shyam Sunder, Rahul Telang, and James Morris (2005). Pricing electronic mail to solve the problem of spam. *Human-Computer Interaction*, 20(1-2):195-223.
- Theodor Loder, Marshall Van Alstyne, and Rick Wash (2006). An economic response to unsolicited communication. *Advances in Economic Analysis and Policy*, 6(1). Article 2.
- Melville, Nigel P., Stevens, Aaron, Pllice, Robert K. and Pavlov, Oleg V., Unsolicited Commercial E-Mail: Empirical Analysis of a Digital Commons (July 1, 2006). *International Journal of Electronic Commerce*, Vol.10, No. 4, pp. 143-168, 2006.
- Pavlov, Oleg V., Melville, Nigel P. and Pllice, Robert K.. Mitigating the Tragedy of the Digital Commons: The Problem of Unsolicited Commercial E-Mail. *Communications of*

the Association for Information Systems, Vol. 16, pp. 73-90, 2005.

Pavlov, Oleg V., Melville, and Plice, Robert K. Toward a sustainable email marketing infrastructure. *Journal of Business Research*, 2008.

L Von Ahn, M Blum, NJ Hopper, J Langford. CAPTCHA: Using hard AI problems for security. *Lecture notes in computer science*, 2003.

T Van Zandt. Information overload in a network of targeted communication. *The RAND Journal of Economics*, Vol. 35, No. 3, Autumn 2004 pp. 542-560.