Information Asymmetry and Thwarting Spam

Thede Loder, Marshall Van Alstyne, Rick Wash*

April 9, 2004

Abstract

We explore a novel approach to spam based on economic rather than technological or regulatory screening mechanisms. Our first point is that mechanisms designed to promote valuable communication can often outperform those designed merely to block wasteful communication. Our second is to shift focus from the information in the message to the information known to the sender. We can then use principles of information asymmetry to cause people who knowingly misuse communication to incur higher costs than those who do not. In certain cases, though not all, we can show this approach leaves recipients better off than even an idealized or "perfect" filter that costs nothing and makes no mistakes. Our mechanism also accounts for individual differences in opportunity costs, and allows for bi-directional wealth transfers while facilitating both sender signaling and recipient screening.[1]

I Introduction

Due to its low cost, speed, and freedom from geographical constraints, email has become a ubiquitous and arguably essential means of communication. Unfortunately, the same properties that make it so useful, combined with its openness and trusting design, enable unscrupulous marketers to broadcast email to untargeted audiences. The result is unnecessary and unwarranted costs for recipients.

Recent estimates indicate more than 50% of email is now spam, the volume of spam is growing rapidly, and worldwide costs exceed $20 billion annually(Hansell, 2003). The enormous quantity of unwanted communications has reduced the signal-to-noise ratio of email to such an extent that it has become an issue of national importance.

Legislative and technological solutions continue to be the primary means pursued to stop or limit spam. No fewer than eight bills[2] have been introduced into Congress over the last several years, and President Bush signed the CAN-SPAM Act into law in December, 2003. More than half the states (Spam Laws, 2003) have enacted laws to regulate email. Concurrently, the technology industry is mobilizing to provide products and

services intended to give back some measure of control to mailbox owners. In 2002, at least $54.4 million was invested in anti-spam startups, up 65% from the previous year (Hopkins, 2003).

Using principles of information economics, we develop an alternative to the popular mechanisms for filtering and banning communications as well as to challenge-response systems which verify sender identity. Based on a very simple model, we show that use of the right mechanism - one that facilitates communication rather than blocks it or bans it - can improve social welfare.

Our mechanism encourages selective targeting of messages, has dynamically adjusting prices (accounting for a recipient's value of time), depends on ex post verification – not ex ante classification – of content so that deceitful subject lines do not matter, transfers wealth directly to recipients and so requires neither rebate mechanisms nor government oversight, and it is incentive compatible. Recipients have reason to adopt it, not only to manage their incoming messages, but to receive wealth from those who would ask for their attention. Last, the mechanism is bi-directional, allowing both signaling and screening.

The remainder of this paper proceeds as follows. Section II references existing literature while critically examining several proposed solutions to the problem of spam. Section III graphically illustrates the effects of different policies on social welfare and offers intuition into why facilitating exchange can help recipients and senders more than blocking certain kinds of communication. It also introduces the use of "Attention Bonds" as an alternative screening mechanism. Section IV then develops a formal model and compares results from use of the attention bond to those of a perfect filter. Section VI relaxes certain assumptions and generalizes key results. It shows that signaling high value communication is also possible. Caveats and adoption issues appear in Section VII. Broader social implications are briefly examined in Section VIII. Section IX concludes.

II Existing and Proposed Solutions

The popular approaches to managing spam fall roughly into two categories, technological and regulatory. Following a review of these, we review work in the area of applying market mechanisms.

A Technological Solutions

Popular technological approaches include challenge-response, rule-based filters, Bayesian filters, and community classification.

Challenge-response systems initially block or hold email from unknown senders. The senders are notified of the block, then required to prove they are human by taking a quasi-Turing test. If they pass, the email is delivered (Mailblocks[3]).

Rule-based filters (some updated live, like those from Brightmail[4]), apply pattern matching rules to content or email headers. Similarly, Bayesian filters (Graham, 2002) use Bayesian networks, which can be trained over time with user feedback.

Community classification technology facilitates the harvest of human classification efforts across a large community (Cloudmark[5], Razor[6]). These systems block an email from further delivery if enough people identify it as spam.

The more complete solutions use a combination of technologies, but each of these above approaches have certain problems. For example, filters result in false positives and false negatives, and these failures can be time consuming for recipients and legitimate senders. Quasi-Turing tests lock out potentially useful and low-cost automated correspondence, such as account updates from online retailers and banks. Shared filters and human classification techniques result in a consensus definition of spam, cutting out potentially useful exchanges (one person's garbage is another's gold, and your neighbors end up deciding what you read).

A different, though promising approach is to incorporate the use of strong identities using digital signatures (Tompkins and Handley, 2003). This allows authentication of unrecognized senders and explicit granting of permission for email transactions. Authentication has the advantage that it prevents "spoofing," deliberate misuse of a third party identity to gain access, and it will inevitably become part of any realistic solution.

However, there are a few difficulties with strong authentication alone. The first is that the ease of obtaining new identities, however verifiable, makes it possible to start over each time the reputation capital of any given identity is spent beyond repair. Friedman and Resnick (Friedman and Resnick, 2001) show that newcomers will inevitably need to "pay their dues" in any open society (one that does not charge per access) that has low cost identities.

Related to the first problem, and in contrast with other forms of transaction where messages are used for negotiation purposes only, it is communication itself that is the subject of negotiation. If a recipient blocks email from all verifiable but unknown senders, email alone cannot be used to initiate valuable relationships of exchange, defeating one of its primary uses. Should the recipient use a different policy, for instance, one that lets through at most one email from each unknown sender, a spammer need merely generate a new identity every time he sends to be able to continually thwart the screen.

Unfortunately, no existing solution has been completely effective and most devolve to a technological arms race. As long as marginal costs of sending remain low, a spammer can simply send variation after variation to thwart a filter, eating up bandwidth along the way.

B Regulatory Solutions

The intent of the legal approaches is to regulate email communications. Various laws have been proposed to tax spam, force identification tags or labeling, create do-not-spam lists, and impose criminal charges on behavior outside prescribed guidelines. Eight states have passed anti-spam laws, including recent laws in California and Virginia which legalized substantial fines.

The CAN-SPAM law, recently passed, overrides the state laws, in effect, weakening them. Although a detailed analysis of CAN-SPAM is outside the scope of this paper, the national law strategy may fail for several reasons: a high cost of enforcement, a lack of incentive compatibility, issues of jurisdiction (spammers are already overseas), and the a nebulous definition of what constitutes spam. From an economic perspective, a legal consensus definition has the same outcome as the harvested human classification technologies: the one-size-fits-all approach has the potential to halt fruitful mutually-desired exchange.

Those involved are well aware of these issues. Timothy Muris, chairman of the FTC, made the following comment regarding the proposal to create a national do not email registry: "If such a list were established, I'd advise customers not to waste their time and effort. Most spam is already so clearly illegitimate that the senders are no more likely to comply with new regulations than with the laws they now ignore." (Firestone and Hansel, 2003)

C Market Solutions

A few articles have explored market-based mechanisms for allocating receiver attention (Kraut et al., 2003; Fahlman, 2002; Zandt, 2003). Such mechanisms include stamps, surcharges on communication, and auctions. These might work by shifting the burden of screening from recipients to senders who know more about message content. One such proposal, focusing on sender surcharges, benefits most recipients and all senders by forcing them to collectively stop over-exploiting receiver attention (Zandt, 2003). Senders always benefit if either surcharges are rebated to the community, or more accurate recipient profiling allows senders to target more selectively. Shortcomings of focusing on senders include voluntary participation in surcharge mechanisms and also the ability to lie about content ex ante in order to elevate interest.

An experimental investigation of pricing recipient attention via email postage found that charging does cause senders to be more selective and to send fewer messages (Kraut et al., 2003). In particular, variable rate usage charges reduced communication more than flat rate access charges. Interestingly, recipients did not see postage as a signal of value and the authors conclude that such systems show great promise but "need more work (p. 206)."

One well-designed mechanism is outlined by Fahlman in (Fahlman, 2002) and more casually in (Ayres and Nalebuff,

2003). The main observation is that communications media, such as email, telephone and instant messaging, allow one person to interrupt another and obtain their attention. Fahlman proposes giving recipients (or the target party of a communication) the ability to sell 'interrupt rights' to senders. Strangers must make a binding offer for the privilege of diverting a recipient's attention; they agree to pay an 'interrupt fee'.

III An Economic Approach

We extend early proposals in several ways. First, we introduce a formal model that allows incentive analysis and welfare comparisons across proposals, including the ability to explore different recipient policies regarding interruptions. Second, we extend the mechanism to make it bi-directional, that is, we allow welfare transfers in both directions, which further enhances the creation of markets for attention. By permitting screening *and* signaling, recipient choice or sender choice can help designate high value messages. Third, we compare this not just to the baseline case of no intervention, but to a 'perfect' filter, which we define as a filter that is costless to operate and makes no mistakes (no false positives or false negatives). Although no such filter exists, we use this as a proxy for any kind of filtering or banning technology. We then show that situations exist where an economic solution creates greater welfare and remains incentive compatible.

A Key Intuition

The pure technological and regulatory approaches limit unwanted communications by blocking or banning them. This goes against a classic principle of economics:

> *In terms of individual and aggregate social welfare, a system that facilitates valuable exchange and side payments will generally dominate a system that grants only unilateral veto power to either party.*

The improvement in exchange follows from mechanism design and the principles of information asymmetry. Our primary assumption is that the person who composes a message knows more about its content than a person who has not yet read it. This private information favors the sender, and standard mechanisms exist for screening out informed parties that would take advantage of uninformed parties. These include reputations and warranties. We therefore propose a screening mechanism that allows recipients to discriminate between classes of high and low quality senders or conversely a signaling mechanism that allows high quality senders to rise above the noise.

B Attention Bond Mechanism

In the case of any sender who has a prior relationship with a recipient, reputation systems work well. Such persons can simply be "whitelisted" and their messages passed through unchallenged. These lists could also be created for recipient inboxes based on the recipients own outbox or through "letters of introduction" based on the `CC:` field of known contacts.

In the case of strangers, the warranty mechanism is more suitable. Analogous to a standard bond mechanism, delivering email to an inbox requires an unknown sender to place a small pledge into escrow with a third party. In the case of screening, recipients determine the size of this bond, which they can dynamically adjust to their opportunity costs. The email is delivered only after the recipient receives suitable confirmation that the bond has been posted. When the recipient opens the email, she may act solely at her discretion to seize the pledge. Taking no action releases the escrow after a period of time.

If a recipient expects further communication with a particular sender and wishes to remove the bond requirement, they can add the sender to the whitelist, whereupon messages from the sender will pass through the screen unencumbered. The idea is simply to cause those who would misuse communication to signal their intention by their willingness to incur risk. Senders of valuable communications bare little exposure.

C Model

Intuition follows from a graphic representation of sender and recipient gains (or losses) due to acts of communication. For simplicity, let there be arbitrary maximum and minimum values $\bar{V}$ and $\underline{V}$ to a message, which are positive ($\bar{V} > 0$) and negative ($\underline{V} \leq 0$) respectively. These represent the range of value from welcome and unwelcome communication. To distinguish senders from receivers, subscript these by s and r. Introducing marginal cost c_s of composing and sending a message, cost c_r of reading and disposing of a message, then juxtaposing the distributions on a single plot gives Figure 1.

Under sender choice, messages will not be sent when the sender's value is negative. This eliminates messages to the left of c_s. Increasing sender costs, for example by taxing senders, provides one means of curtailing low sender value messages, and would be reflected in Figure 1 by moving c_s to the right.

Importantly, however, total value of communicating is the sum of sender and receiver value. Total surplus increases in the positive direction on both the s and R axes, in the region northwest of the welfare line W. Assuming that a filter stops all messages for which the value does not justify the cost of reading, the filter would eliminate messages south of c_r. Relative to the no-intervention case, this reduces recipient losses. It also, however, eliminates a region of positive social surplus. The triangular region below c_r but above W represents possible gains from trade. Within this region, unrecognized but legitimate marketing

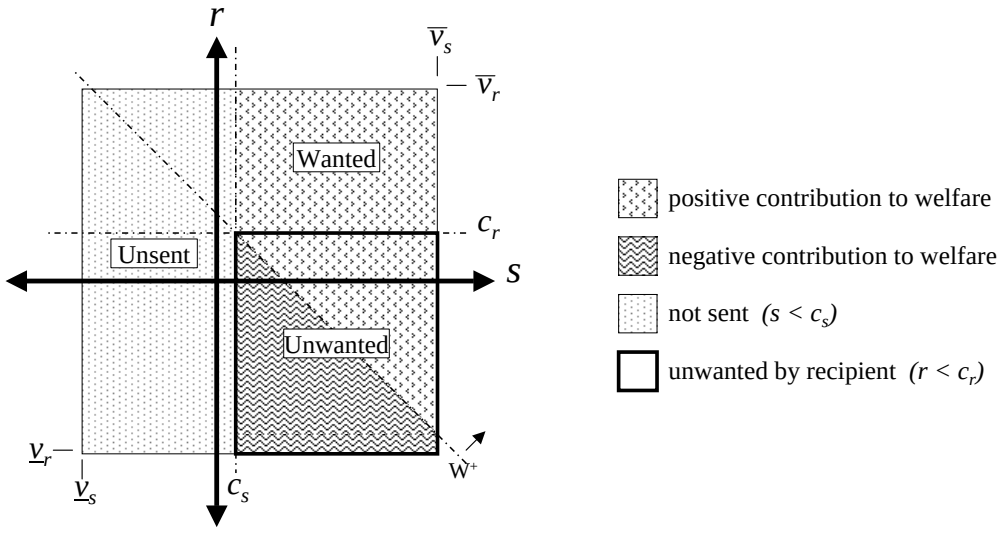


Figure 1: Distribution of Email

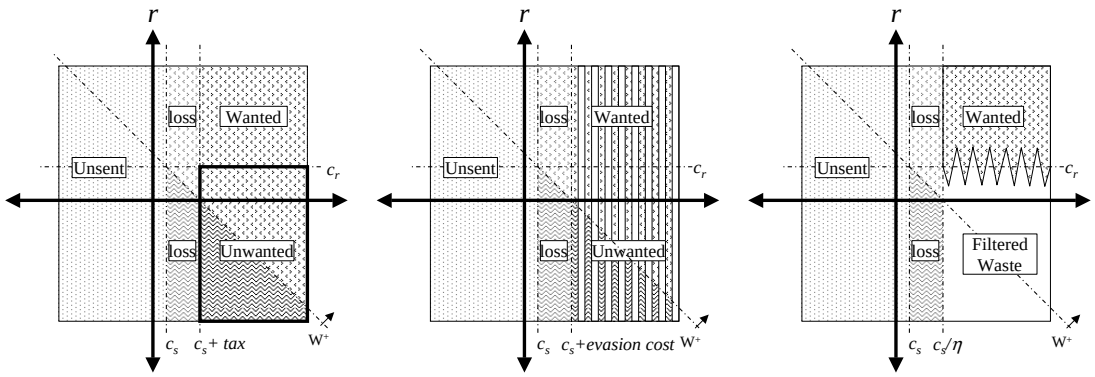
organizations, political campaigns, charities, persons seeking interviews, and remote contacts of one's social network might offer value in return for a recipient's attention. In economic terms, recapturing and dividing this surplus represents an opportunity for both parties.

In general, it will not be possible to recover this area perfectly. First, a recipient cannot know the value of a message from an unknown sender before seeing it. Second, realized value to senders and receivers can be private information, implying that the amount of surplus is unknown and subject to misrepresentation. Third, negotiating an acceptable division of surplus is complicated by the difficulty that the act of communication is itself the subject of the negotiation. A mechanism to substantially reclaim a measure of this surplus represents the bulk of this paper.

For comparison, Figure 2 illustrates social welfare effects of several proposed solutions explored in the previous section. Figure 2-a shows the effects of a levied tax in the form of fees, computation [7], challenge-response test taking effort, or time. This shows that a tax eliminates many wasteful messages but also cuts certain messages that are low value to senders and high value to recipients. It has no effect on messages that are of high value to senders but waste to recipients.

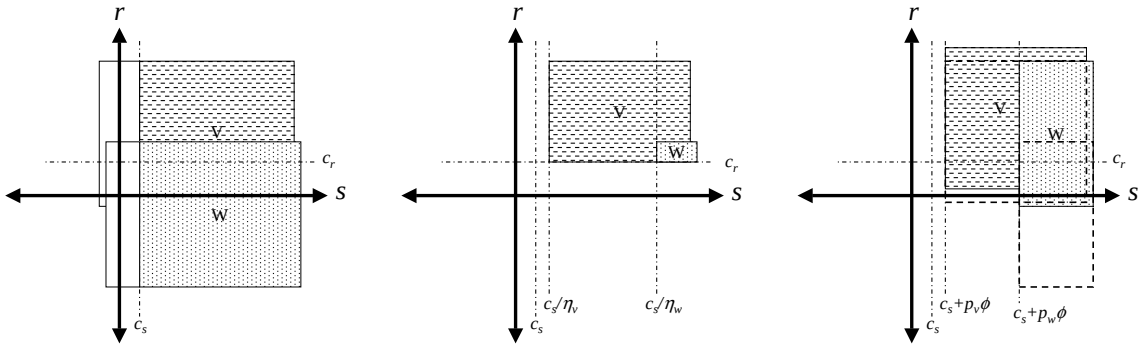
Figure 2-b shows the effects of bans, labeling, and "do-not-call" lists. Senders governed by relevant law are either enjoined or suitably filtered, reducing recipient waste while senders outside a given jurisdiction are unaffected.

From a recipient's perspective, Figure 2-c represents the more promising alternative of filters including, for example, rule-based, community-based, and Bayesian. Better filters learn recipient preferences and eliminate unwanted messages while suffering from fewer false positives (passing junk messages) and false



(a) Tax (b) Legal (c) Technological

Figure 2: Existing Solutions



(a) (b) (c)

Figure 3: Screening Example

negatives (screening valuable messages). Typically, however, all three systems focus on averting unwanted communication while overlooking the question of promoting valuable exchange.

D Screening Mechanism

Standard solutions to information asymmetry problems, where one party has information the other does not, include reputations (Shapiro, 1982), sunk costs of signaling (Spence, 1973), and bonding or warranties (Akerlof, 1970). In the context of communication from unknown senders, the absence of established reputation is a primary obstacle in dealing with spam. Pseudonymous sending is itself the problem. Sunk costs, in turn, are expensive and preferably avoided as socially inefficient. Thus we focus on bonding or warranties as an effective screen.

We also focus on the potential for a valuable relationship with a given sender as distinct from the potential for a specific valuable message. This avoids the considerable technical difficulty of classifying a particular message ex ante and instead relies on ex post verifiability of a sender's type.

Figure 3 shows the effects of filtering and attention bonds on two different email value distributions (modeled in later sections). Frame a shows the baseline case with valuable V and wasteful W distributions. For the baseline, email to the right of c_s is sent, and regions contributing to recipient surplus are shaded grey. The second frame, b shows regions of contribution to recipient surplus resulting from a perfect filter, which has two effects relative to frame a. The first effect is to reduce unwanted mail, eliminating messages below c_r. The second effect is to increase sender costs, for if $\eta = \frac{1}{5}$ messages reach their destination then $(1-\eta) = \frac{4}{5}$ messages are filtered, and in order to reach the same audience a sender incurs new costs $5c_s$. Alternatively, welfare in this region could simply be reduced by $(1-\eta)$.

In the final frame, c, recipients use the ABM and choose a bond size ϕ (an 'interrupt fee'). This shows sender costs increasing by an expected forfeit rate $p\phi$ but, as we will show, this rate differs for those sending valuable (p_v) and those sending unwanted (p_w) email. Unlike a filter, senders transfer surplus to recipients so that recipient surplus shifts up by $p\phi$, allowing interruptions to become valuable or at least value neutral.

IV Formal Model and Results

We will now introduce a formal model, with which we attempt to capture the value structure for email messages for both sender and recipient. The model is based on the value of a single message, the first communication from a stranger. We limit our focus to this as for subsequent email, information has already been revealed that can be used to whitelist the sender or, conversely, increase size of the demanded bond.

A Model

The value of an email to a sender, s, and the value of that same email to the recipient, r, we assume lie within a range of values. The maximum s is $\overline{v_s}$, minimum $\underline{v_s}$. Similarly, r is bounded by $\overline{v_r}$ and $\underline{v_r}$.

The probability of an email having value s to the sender is given by the function $F_S(s,r)$. The probability of a given r is $F_R(s,r)$. We assume $F_S(s,r)$ and $F_R(s,r)$ are continuous and uniform across their ranges of positive probability. Doing so allows us make the simplification of pulling $k = \frac{1}{\overline{v_s}-\underline{v_s}} \cdot \frac{1}{\overline{v_r}-\underline{v_r}}$ from under the integral of the weighted sum of net payoff which is used to determine sender and recipient surplus.

In addition to the distributions, there are marginal costs for sending and receiving email, c_s and c_r, which were introduced in the previous section. Both the sender and the recipient know their own costs.

For sender and recipient knowledge, we assume that the sender always knows their own value s for a given email but they do not know its value r to the recipient. The recipient knows r, but only after receiving and reading the email at cost c_r. Finally, we assume that the sender will not send an email when their expected return is negative (when $c_s > s$).

With these model parameters defined, we can determine the sender and recipient surplus (SS_0 and RS_0) for the baseline case where no spam solution is employed. Formally, this is

$$\mathrm{SS}_0 = k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s}^{\overline{v_s}} s - c_s \; \partial s \partial r \tag{1}$$

Likewise, the total expected surplus to the recipient for this email is

$$\mathrm{RS}_0 = k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s}^{\overline{v_s}} r - c_r \; \partial s \partial r \tag{2}$$

A.1 A Perfect Filter

To incorporate the perfect filter into the model, we consider its impact on recipient and sender surplus. Since the perfect filter costlessly eliminates any email where $r < c_r$ before the receiver receives it, the contribution to recipient surplus for such email is zero. Should the sender still send, their surplus is decreased by c_s, and any value s they might have gained from delivery is lost.

In addition, we expect senders to notice that a percentage of their emails are no longer getting through. Senders may endogenously correct this to reach the same number of recipients as before and endure a scaled increase in costs. Let η be the percentage of emails that get through the filter. If initially senders did not send messages of value $s < c_s$, with application of the perfect filter they will not send messages with values $s < \frac{c_s}{\eta}$.

Given these additions to the model, the total expected sender surplus for the perfect filter will be

$$\text{SS}_{PF} = k \cdot \int_{c_r}^{\overline{v_r}} \int_{\frac{c_s}{\eta}}^{\overline{v_s}} s - c_s \; \partial s \partial r - k \cdot \int_{\underline{v_r}}^{c_r} \int_{\frac{c_s}{\eta}}^{\overline{v_s}} c_s \; \partial s \partial r \tag{3}$$

Likewise, the total expected recipient surplus will be

$$\text{RS}_{PF} = k \cdot \int_{c_r}^{\overline{v_r}} \int_{\frac{c_s}{\eta}}^{\overline{v_s}} r - c_r \; \partial s \partial r \tag{4}$$

The second term in Equation 3 represents the sender losses due to emails that are sent and subsequently filtered before receipt. These losses, in effect, are what drive up the value of threshold marginal cost $\frac{c_s}{\eta}$. However, for simplicity, we will ignore this term for later results (doing so favors the perfect filter).

If senders do not correct for the filtering of messages, a static interpretation for recipient surplus is

$$\text{RS}^{'}_{PF} = (\eta)k \cdot \int_{c_r}^{\overline{v_r}} \int_{c_s}^{\overline{v_s}} r - c_r \; \partial s \partial r$$

where surplus is due to the fraction of messages reach their destination.

The formulation in Equation 4 allows spammers to adjust their behavior and so will be used for analysis. The static analysis could be used with results that are qualitatively unchanged. Note that as $c_s \to 0$, the endogenous formulation shows *no* effect on sender costs, favoring the perfect filtering case.

A.2 Attention Bond Mechanism

The Attention Bond Mechanism (ABM) is represented by two parameters: a bond value ϕ and probability of seizing the bond p. The bond value ϕ is chosen ex ante and will be a constant in the model. Probability p is a proxy for the recipient's "policy", or their decision process to seize a bond based upon factors such as the value received from communication. Policy together with the bond, $p \cdot \phi$, is the expected benefit to the recipient and the extra cost of the bond mechanism to the sender. The cost increase for the sender has the secondary effect of reducing the total emails sent to those where $s > c_s + p\phi$ is true.

Putting all of this together, we find that the sender's total expected surplus is

$$\text{SS}_{\phi,p} = k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s + p\phi}^{\overline{v_s}} s - c_s - p\phi \; \partial s \partial r \tag{5}$$

Likewise, the recipient's total expected surplus is now

$$\text{RS}_{\phi,p} = k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s+p\phi}^{\overline{v_s}} r - c_r + p\phi \; \partial s \partial r \tag{6}$$

B Results

B.1 Single Distribution

We start with results where the sender and recipient draw from a single distribution, and this will form the basis of later analysis. First, we must calculate what bond the recipient should charge to receive the optimal payoff.

Lemma 1 *The optimal bond ϕ^+ that a recipient can choose is*

$$\phi^+ = \frac{1}{2p}\left((\overline{v_s} - c_s) - \left(\frac{\overline{v_r} + \underline{v_r}}{2} - c_r\right)\right)$$

Proof. Given our assumption of uniform value distributions, we integrate Equation 6, which yields

$$\frac{\overline{v_s} - c_s - p\phi}{\overline{v_s} - \underline{v_s}}\left(\frac{\overline{v_r} + \underline{v_r}}{2} - c_r + p\phi\right) \tag{7}$$

Then we set the derivative of Equation 7 to zero and solve for ϕ. By inspection of Equation 7, the second derivative is $-p^2$, which is always negative. Therefore, this value for ϕ is a maximum.

This optimal bond is exactly one half of the total sender surplus minus the average recipient surplus. This makes sense, as it forces the sender to split her surplus with the recipient. It is also inversely proportional to the probability of collecting the bond. As a recipient collects the bond less frequently, they need to increase the size of the bond to have the same effect.

With the optimal bond from above, we can compare the ABM to the baseline case and see if and when it is better.

Proposition 1 *Given a recipient-chosen optimal bond ϕ^+ and a corresponding p, the recipient surplus with the ABM is always at least that of the baseline case or*

$$\text{RS}_{\phi^+,p} \geq \text{RS}_0$$

To prove this, we first set up the equations:

$$k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s+p\phi^+}^{\overline{v_s}} r - c_r + p\phi^+ \, \partial s \partial r \geq k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s}^{\overline{v_s}} r - c_r \, \partial s \partial r \tag{8}$$

When evaluated, it turns out that this is always true. This means that a recipient is always better off using the attention bond mechanism.

Next, we begin our comparisons with a perfect filter. We ask if the bond mechanism can actually do better than a perfect technological filter.

Proposition 2 *For certain distributions, using the ABM with a recipient-chosen optimal bond creates greater recipient surplus than when using a perfect filter.*

$$\mathrm{RS}_{\phi^+,p} > \mathrm{RS}_{PF}$$

Proof. We set up an inequality similar to Equation 8 using Equations 4, and 6. Then we incorporate Lemma 1 and use $\eta = \frac{\overline{v_r}-c_r}{\overline{v_r}-\underline{v_r}}$ and ask under what conditions, if any, the inequality is true. It turns out to be true when

$$\frac{\left(\overline{v_r}-\underline{v_r}\right)\left(\overline{v_s}-c_s+\left(\frac{\overline{v_r}+\underline{v_r}}{2}-c_r\right)\right)^2}{\cdot\left(\overline{v_r}-c_r\right)\left(\overline{v_r}\left(\overline{v_s}-c_s\right)-\overline{v_s}\cdot c_r+\overline{v_r}\cdot c_s\right)} \geq 2 \tag{9}$$

Equation 9 is not always true, but it is for some important situations. In particular, this is true when the sender value is fairly high and the recipient value is low or negative. (For example, when $\overline{v_s} > 2 \cdot \overline{v_r}$ and $\underline{v_r} \ll 0$, or in distribution W, from Figure 4.)

B.2 Social Welfare

Now that we have looked at the effects of the Attention Bond Mechanism for the recipient, the question remains "Is this a good idea overall?" To answer this question, we must calculate the total social welfare of the system.

Definition 1 *The total social welfare contribution of an email is the sum the recipient and sender surplus.*

$$\mathrm{W} = \mathrm{RS} + \mathrm{SS}$$

With Definition 1, we can compare the social welfare contribution with the Attention Bond Mechanism to that of the perfect filter.

Proposition 3 *For certain value distributions, the total social welfare with a recipient-chosen optimal bond is greater that of a perfect filter.*

$$\mathrm{W}_{\phi^+,p} > \mathrm{W}_{PF}$$

To prove, we use Definition 1, with Equations 3, 4, 5, 6, Lemma 1, and $\eta = \frac{\overline{v_r}-c_r}{\overline{v_r}-\underline{v_r}}$ then simplify the inequality. We find that the ABM is superior when

$$\begin{aligned} \frac{4}{3} \leq & \frac{\left(\overline{v_r}-c_r\right)\left(\overline{v_r}-\underline{v_r}\right)}{\left(\overline{v_r}\left(\overline{v_s}-c_s\right)-\overline{v_s}\cdot c_r+\overline{v_r}\cdot c_s\right)} \\ & \times \frac{\left(\overline{v_s}-c_s+\frac{\overline{v_r}+\underline{v_r}}{2}-c_r\right)^2}{\left(c_r^2+\overline{v_r}^2-c_r\overline{v_s}+\overline{v_r}\left(-2c_r-c_s+\overline{v_s}\right)+c_s\underline{v_r}\right)} \end{aligned} \tag{10}$$

Similarly to Equation 9, this is not always true, but it is in certain common situations. For example, if the distribution is the "Unwanted" box from Figure 1 (where most of the recipient value is negative), or distribution W from Figure 1, then this equation will be true.

An alternative method of choosing the bond is to attempt to maximize the social welfare that is produces.

Lemma 2 *The social welfare maximizing bond size is*

$$\phi^\star = \frac{1}{p}\left(\frac{\overline{v_r}+\underline{v_r}}{2}-c_r\right)$$

Proof. To prove, find $\mathrm{W}_{\phi,p}$ using Definition 1 and Equations 5 and 6. Maximize with respect to ϕ.

Note that the result for $\phi^\star$is proportional to the average recipient surplus, but unrelated to sender surplus. The bond essentially compensates the recipient for their inconvenience.

Now that we know the social welfare maximizing bond, we can ask when and if this bond produces greater social welfare than the perfect filter.

Proposition 4 *For certain distributions, the social welfare when using the ABM with a socially-optimal bond $\phi^\star$ is greater than the social welfare when using a perfect filter*

$$\mathrm{W}_{\phi^\star,p} > \mathrm{W}_{PF}$$

Proof. We begin with the setup

$$k \cdot \int_{\underline{v_r}}^{\overline{v_r}} \int_{c_s+p\phi^\star}^{\overline{v_s}} (r - c_r + p\phi^\star) + (s - c_s - p\phi^\star)\, \partial s \partial r \geq k \cdot \int_{c_r}^{\overline{v_r}} \int_{\frac{c_s}{\eta}}^{\overline{v_s}} (r - c_r) + (s - c_s)\, \partial s \partial r \tag{11}$$

Filling in $\phi^\star$and η and solving, we find that this is true when

$$1 \leq \frac{(\overline{v_r} - c_r)\left(\overline{v_r} - \underline{v_r}\right)}{(\overline{v_r}\,(\overline{v_s} - c_s) - \overline{v_s} \cdot c_r + \overline{v_r} \cdot c_s)} \times \frac{\left(\overline{v_s} - c_s + \frac{\overline{v_r}+\underline{v_r}}{2} - c_r\right)^2}{\left(c_r^2 + \overline{v_r}^2 - c_r\overline{v_s} + \overline{v_r}\,(-2c_r - c_s + \overline{v_s}) + c_s\underline{v_r}\right)} \tag{12}$$

Note that this is the same as Equation 10 from Proposition 3 with a different constant. Similar to Equation 10, the inequality will be true if the distribution is similar to "Unwanted" box from Figure 1, or the W distribution in Figure 3.

V Policy Independence

Recall that a policy is the basis of a recipients' decision regarding the bond, modeled as a probability p of seizing it. Example policies include (i) seize the bond for low value messages $r < c_r$ (ii) always seize the bond (iii) seize the bond only for the most offensive communication $r \approx \underline{v_r}$ or (iv) seize randomly. Unusual but permissible policies could include seizing just from the top, the middle, or both ends of the value distribution.

If senders do not know recipients' private information, i.e their true value for a message, but only know the seize rate, then the welfare results have the property of being "policy independent."

Policy independence states that, subject to the boundary conditions $p^+\phi^+ = k$ and $0 < p^+ \leq 1$, a recipient is free to choose any policy she wishes without affecting her maximum expected surplus. This property arises because recipient surplus depends on an optimal and constant bond benefit $p^+\phi^+$. Since

expected net payoffs from sending a message are also constant, this implies that sender surplus is also policy independent and, at the optimum, total welfare is independent of a recipients' individual choice of seizure policy p.

An alternate interpretation of policy independence is that a rational recipient can freely choose any size screen ϕ by reducing the seize rate, subject to the same boundary conditions. Individual tailoring based on factors outside the model, such as risk aversion, also follows.

To state policy independence formally, we have the following lemma:

Proposition 5 *The recipient surplus using a bond is* ***policy independent****. That is, for any bond collection policy that collects with probability p, there exists a bond ϕ which gives the recipient the maximum surplus.*

$$\forall p, \text{ with } 0 < p \leq 1, \exists \phi \text{ such that } \mathrm{RS}_{\phi,p} = \mathrm{RS}_{\max}$$

Likewise, given any bond ϕ above a certain minimum, there exists a collection policy which collects with probability p that gives the recipient the maximum surplus.

$$\forall \phi, \text{ with } \phi > \phi_{min}, \exists p \ 0 < p \leq 1 \text{ such that } \mathrm{RS}_{\phi,p} = \mathrm{RS}_{\max}$$

Proof. To prove the first half of the proposition, we look at the definition of $\mathrm{RS}_{\phi,p}$. If we take p as given, we can then use the definition of $\phi^+$to calculate an optimal ϕ. Filling in this p and $\phi = \phi^+$ into the equation for $\mathrm{RS}_{\phi,p}$, we see that the resulting surplus is independent of both p and ϕ, and therefore a constant.

To prove the second half, we first need a lemma about the optimal collection policy given a bond ϕ.

Lemma 3 *Given a ϕ, the optimal recipient-chosen policy p^+will be*

$$p^+ = \frac{1}{2\phi}\left((\overline{v_s} - c_s) - \left(\frac{\overline{v_r} + \underline{v_r}}{2} - c_r\right)\right)$$

This can be derived in a manner similar to $\phi^+$by taking the derivative of $\mathrm{RS}_{\phi,p}$ with respect to p.

Now we can prove the second half of the proposition. Looking at the definition of $\mathrm{RS}_{\phi,p}$ we can fill in $p^+$for p and see that, like the first half, the surplus is independent of both ϕ and p and therefore a constant.

Importantly, a recipient might choose a non-optimal policy of choosing both p and ϕ but this becomes endogenously self-correcting. Either a low expected screen $p\phi$ encourages too many low value communications or a high expected screen discourages too many high value senders.

VI Relaxing Assumptions

Now we consider two cases where certain assumptions in the model are relaxed. First, we consider using this mechanism in reverse as a signal to senders. This allows people to signal their interest in a given topic. Then, we explore the consequences of placing email messages into multiple distributions. The results provide intuition for performance of the attention bond mechanism when one distribution is mostly spam while another is mostly valuable (e.g. distributions V and W).

A Signaling & Reverse Signaling

To this point, we have assumed the uninformed recipient chooses the size of ϕ. An alternative possibility is for the informed sender to signal their interest in a potentially valuable communication. If this is the case, the choice of ϕ is no longer optimal from the recipient's perspective but merely chosen to avoid negative expected recipient surplus. That is, a sender minimizes ϕ such that $E[\mathrm{RS}_{\phi,p}] \geq 0$, where $E[\mathrm{RS}_{\phi,p}]$ is given by Equation 6.

Integration of Equation 6 yields $\frac{(\overline{v_s}-c_s-p\phi)}{\overline{v_s}-\underline{v_s}}\left(\frac{\overline{v_r}+\underline{v_r}}{2}-c_r+p\phi\right)$ for which the roots are $p\phi \in \left\{\overline{v_s}-c_s, -\frac{\overline{v_r}+\underline{v_r}}{2}+c_r\right\}$. Thus the expected signal is either the maximum surplus to the sender or the expected value to the recipient, whichever is less. Given that the distribution draws from unrecognized senders, one might expect $\frac{\overline{v_r}+\underline{v_r}}{2}-c_r<0$ so that pledging this amount raises the recipient's expected value of communication from unknown persons to at least ≥ 0.

This analysis extends further to bilateral initiative. The attention bond mechanism permits reverse signaling to occur from a recipient who could *solicit* messages of high net value to themselves but low net value to senders – messages that would otherwise have gone unsent. Such communication could involve high sender costs, high recipient personalization, or sender reputation effects from poorly directed communication. Real life analogues include recipient who front the costs of expensive catalogues that are rebated on the first purchase, and credit applications from persons interested in tailored information based on loan amounts and credit ratings. This leads to

Proposition 6 *A reverse signaling mechanism strictly increases recipient surplus and also total welfare.*

Proof. Similar to the previous case, a recipient chosen signal solves for the minimum ϕ such that $E[\mathrm{SS}_{\phi,p}] \geq 0$, where $E[\mathrm{SS}_{\phi,p}]$ is given by:

$$k \cdot \int_{c_r+p\phi}^{\overline{v_r}} \int_{\underline{v_s}}^{c_s} s - c_s + p\phi \; \partial s \partial r$$

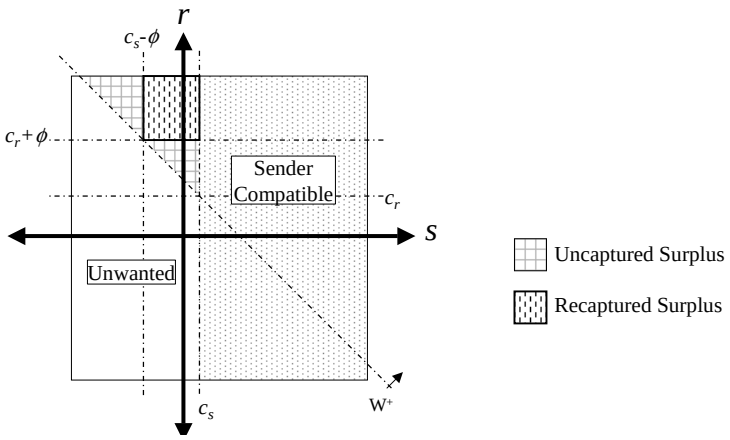


Figure 4: Signaling

Integration yields $\frac{(\overline{v_s}-c_s)}{(\overline{v_s}-\underline{v_s})(\overline{v_r}-\underline{v_r})}\left(c_r-\overline{v_r}+p\phi\right)\left(\frac{\underline{v_s}-c_s}{2}+p\phi\right)\geq 0$ implying that $p\phi\in\left\{\overline{v_r}-c_r,\frac{c_s-\underline{v_s}}{2}\right\}$. Assuming a sender keeps the pledge $p=1$, a recipient fronts an amount equal to her own surplus or the sender's average losses, whichever is less. This leaves senders at least as well off as under the perfect filter.

For any region in which $r>\frac{c_s-\underline{v_s}}{2}$, new communications take place with the attention bond mechanism that never occur with the perfect filter or the baseline case of no intervention. In graphic terms, this recaptures the region of uncaptured surplus in Figure 4. Total welfare strictly improves.

B Perfectly Correlated Values

In the baseline model we assumed that recipient and sender values are uncorrelated, and found that the attention bond dominates in some but not all cases. Here we extend these results to linearly correlated values $r=\lambda s$ such that $\lambda\in(-\infty,0)$ implies negative and $\lambda\in(0,\infty)$ implies positive correlation. As an aside, linear correlation means that, having written (or read) any specific message, senders and receivers can infer each other's value. With correlation, an attention bond almost always dominates a perfect filter.

Proposition 7 *When sender and receiver values are linearly correlated, receiver surplus from an attention bond is almost always at least as great as that from a perfect filter. Social welfare of the bond almost always exceeds that of a perfect filter.*

$$r=\lambda s \text{ implies } \mathrm{RS}_{\phi^+,p}\geq \mathrm{RS}_{PF} \text{ and } \mathrm{W}_{\phi^+,p}\geq \mathrm{W}_{PF}$$

Proof. The case of negative correlation is straightforward Consider first the perfect filter. For $r > 0$ but $\lambda < 0$, a recipient wants communication but a sender will not send. For $s > 0$ but $\lambda < 0$, a sender would communicate but being blocked avoids the cost of a message. Total welfare is 0. In contrast, the attention bond passes certain messages. Let values be such that $\frac{s}{-\lambda} - (c_s + \phi) > r + c_r$. Then sender net surplus exceeds attention bond ϕ, the bond is always seized, and neither party has negative payoffs. Note also that reverse signaling becomes feasible. A recipient with surplus $r - c_r - c_s + \phi > 0$ can choose $\phi = s - c_s = \frac{r}{\lambda} - c_s < 0$ for sender surplus $\frac{r}{\lambda} - c_s - \phi \geq 0$, allowing the recipient to subsidize interaction. A pure Pareto improvement follows.

In the case of positive correlation $\lambda > 0$ and a perfect filter, a sender never sends if $s < c_s$ or if $\frac{s}{\lambda} < c_r$. The latter is blocked and costly to send without payoff. For the attention bond, if $c_s > \lambda c_r$ then the recipient safely chooses $\phi = 0$ since the sender is self-policing. Otherwise, let $\phi = \frac{c_r}{\lambda} - c_s$ and all messages that were previously sent and passed through the filter will also be sent and pass the attention bond. To see that there exist new messages that pass through, let sender surplus be $s - \phi > c_s$ with recipient surplus $\lambda s < c_r$ then after substitution $\frac{c_r}{s+2c_s} < \lambda < \frac{c_r}{s}$. Then a sender has sufficient surplus to permit the bond to be seized, yielding positive welfare across both parties.

C Multiple Distributions

Up to this point, we have modeled a single email value distribution. Now we extend this model to handle multiple simultaneous distributions of sender and recipient values, and create distributions sets $i = 1 \ldots n, V_s^1, \ldots, V_s^n, V_r^1, \ldots, V_r^n$, combined as $V^1, \ldots, V^n$, having bounds $\overline{v_s}^i, \underline{v_s}^i, \overline{v_r}^i, \underline{v_r}^i$ and therefore k^i. We make the assumption that a sender knows in which distribution an email they send belongs. A recipient does not know the source distribution of an email ex ante, and while they do know the relative likelihood α_i of an email coming from a given distribution i, they can only choose a single bond size ϕ for use across all of them. After they have read an email, a recipient is assumed to know its distribution and can use different bond seize policies $(p_1, \ldots, p_N)$.

Putting this together, we see that a sender's total average surplus is

$$\text{SS}_{\phi,p_1,\ldots,p_N} = \sum_{i=1}^{N} \alpha_i k^i \cdot \int_{\underline{v_r}^i}^{\overline{v_r}^i} \int_{c_s+p_i\phi}^{\overline{v_s}^i} s - c_s - p_i\phi \; \partial s \partial r \tag{13}$$

Likewise, the total expected recipient surplus will be

$$\text{RS}_{\phi,p_1,\ldots,p_N} = \sum_{i=1}^{N} \alpha_i k^i \cdot \int_{\underline{v_r}^i}^{\overline{v_r}^i} \int_{c_s+p_i\phi}^{\overline{v_s}^i} r - c_r + p_i\phi \; \partial s \partial r \tag{14}$$

C.1 Choosing the Bond under Multiple Distributions

Now we revisit the policy in the case where there are multiple distributions (V_s^1,...,V_s^N) that a sender can draw from. Again we assume that the sender knows which distribution they are drawing from. Also, each sender distribution has a corresponding recipient distribution (V_r^1, ..., V_r^N). Finally, the recipient can only set one bond amount ϕ for all distributions (since it has to be set ex ante), but can collect with different policies for each distribution (p_1, ..., p_N).

Lemma 4 *Given N distributions V^1, ..., V^N, the recipient-chosen optimal bond ϕ^+ is*

$$\phi_N^+ = \frac{\sum_{i=1}^{N} \frac{\alpha_i}{\overline{v_s}^i - \underline{v_s}^i} \cdot p_i \cdot \left(\overline{v_s}^i - c_s - \left(\frac{\overline{v_r}^i + \underline{v_r}^i}{2} - c_r\right)\right)}{\sum_{i=1}^{N} \frac{\alpha_i}{\overline{v_s}^i - \underline{v_s}^i} p_i^2}$$

Each distribution V^i has a relative contribution of α_i to the total surplus, where $\sum_{i=1}^{N} \alpha_i = 1$. Therefore, the total recipient surplus is

$$\text{RS}_{\phi,p_1,\ldots,p_N} = \sum_{i=1}^{N} \alpha_i k^i \cdot \int_{\underline{v_r}^i}^{\overline{v_r}^i} \int_{c_s+p_i\phi}^{\overline{v_s}^i} r - c_r + p_i\phi \; \partial s \partial r \tag{15}$$

Evaluating this equation, taking the first derivative, setting equal to zero and solving for ϕ yields ϕ_N^+.

Lemma 5 *For any distribution V^i, given a bond ϕ, the optimal policy p_i^+ is*

$$p_i^+ = \frac{1}{2\phi}\left(\overline{v_s}^i - c_s - \left(\frac{\overline{v_r}^i + \underline{v_r}^i}{2} - c_r\right)\right)$$

Proof. We take $\text{RS}_{\phi,p_1,\ldots,p_N}$ and calculate the first derivative with respect to p_i. This is simple, as p_i only appears in one term, so the rest of the terms are irrelevant. Solving this for p_i yields the above equation for p_i^+.

With the optimal policy and bond for any one distribution (from Lemma 5), we can can compare expected seize costs across multiple distributions. As one might suspect, the greater the inconvenience to the recipient, the higher the cost to the sender.

Proposition 8 *Let one distribution of email value be V (for 'valuable'), and a second distribution be W (for 'waste'). When the average recipient surplus from email in V is sufficiently greater than that obtained from email in W, the recipient-optimal imposed costs to the sender are higher for the W distribution than for the V distribution. That is,*

$$p_v^+\phi < p_w^+\phi$$

Filling in the values, it is easy to see that this is true when

$$\left(\overline{v_s}^V - c_s - \left(\frac{\overline{v_r}^V + \underline{v_r}^V}{2} - c_r\right)\right) < \left(\overline{v_s}^W - c_s - \left(\frac{\overline{v_r}^W + \underline{v_r}^W}{2} - c_r\right)\right) \tag{16}$$

$$\overline{v_s}^V - \overline{v_s}^W < \frac{\overline{v_r}^V + \underline{v_r}^V}{2} - \frac{\overline{v_r}^W + \underline{v_r}^W}{2} \tag{17}$$

It easy to see that if the upper end of the two distributions is the same for the sender ($\overline{v_s}^V = \overline{v_s}^W$), then this is true whenever the average recipient value for V is greater than that for W. For this to remain true when the sender distributions are different, then there there must be a proportional difference in the average recipient values.

D Policy Independence with Multiple Distributions

Next, we extend our results from Proposition 5 to include the multiple distribution case.

Proposition 9 *Considering distributions V^1,...,V^N, with associated policies p_1,...,p_n and a common bond ϕ, achieving the maximum recipient surplus $\mathrm{RS}_{\phi,p_1,\ldots,p_N}$ is **policy independent**. That is,*

$$\exists \mathrm{RS}_{\max}\ \forall\phi,\ \phi > \phi_{\min}\ \exists p_1,\ldots,p_N\ 0 < p_i \leq 1$$
$$\textit{such that}\ \mathrm{RS}_{\phi,p_1,\ldots,p_N} = \mathrm{RS}_{\max}$$

Likewise,

$$\exists \mathrm{RS}_{\max}\ \forall i\ \forall p_i,\ 0 < p_i \leq 1\ \exists \phi, p_1,\ldots,p_{i-1},p_{i+1},\ldots,p_N$$
$$0 < p_j \leq 1\ \textit{such that}\ \mathrm{RS}_{\phi,p_1,\ldots,p_N} = \mathrm{RS}_{\max}$$

Proof. We maximize $\mathrm{RS}_{\phi,p_1,\ldots,p_N}$ for all of the parameters, and then put all of these in an array. Without actually solving for ϕ in the last equation, we can keep it simpler. This array looks like

$$\begin{aligned}
p_1 &= \frac{1}{2\phi}\left(\overline{v_s}^1 - c_s - \left(\frac{\overline{v_r}^1 + \underline{v_r}^1}{2} - c_r\right)\right) \\
&\vdots = \vdots \\
p_N &= \frac{1}{2\phi}\left(\overline{v_s}^N - c_s - \left(\frac{\overline{v_r}^N + \underline{v_r}^N}{2} - c_r\right)\right) \\
0 &= g_1[p_1,\phi] + \cdots + g_N[p_N,\phi]
\end{aligned}$$

where

$$\begin{aligned}
g_i[p_i,\phi] = \frac{\alpha_i}{\overline{v_s}^i - \underline{v_s}^i}\Big(&p_i \cdot \left(\overline{v_s}^i - c_s - p_i\phi\right) \\
&-p_i \cdot \left(\frac{\overline{v_r} + \underline{v_r}}{2} - c_r + p_i\phi\right)\Big)
\end{aligned} \tag{18}$$

The major observation here is that this system of equations is not independent. Fill in all of the p_i's into the last equation, and you find that every term cancels out. This means that there remains one degree of freedom (within the constraints of the variables) to choose and still maximize surplus. Therefore, choose any one variable, and use this system of equations to set the remaining N variables, and the recipient surplus will still be maximized.

E Virtues of Policy Independence

With respect to recipient surplus, incentive compatibility regarding the seizure of the bond represents a significant virtue of policy independence. A recipient cannot unilaterally improve his or her surplus by always seizing the bond. That is, given an optimal $p_i\phi$ pair, policy independence means that to increase his own surplus a recipient cannot raise the sender's risk of loss p_i without also decreasing ϕ in exact proportion. Given a non-optimal pair, the recipient could choose both high p_i and ϕ but this endogenously reduces willingness to send, and the recipient suffers a welfare loss from messages that are never sent.

Policy independence also means that, subject to the boundary conditions, there is no social efficiency loss in setting a single screen ϕ for 1 distribution or for N arbitrary distributions. That is, the sum of maximum recipient surplus for independent distributions v_r^i individually is the same as total surplus for all $v_r^1, \ldots, v_r^N$ maximized collectively. Further, since the expected recipient gain – or conversely the expected sender

exposure – $p_i\phi$ is constant and independent of other distributions $v^j, j \neq i$, sender surplus in distribution v_s^i is also independent of v_r^j and v_s^j. Therefore aggregate recipient and sender surplus is unaffected. This property is unlikely to hold true for most filtering technology where consideration of additional distributions worsens the problems of false positives and false negatives by introducing confusion among the distributions.

F Dynamic Bond

It is possible to have the bond amount be variable depending on any number of factors. The basic insight here is that the bond request (with the value amount in it) is sent back after the server receives an email. The server can then use information in that email to set the bond amount required to pass the email through.

There are a number of ways this information can be used. First of all, information about the sender can be used. This will allow the recipient to charge a higher (or lower) bond to some senders than others. A good example of this would be the fact that I know most of my family has aol.com email addresses, and few spammers use such address. So I could charge a smaller bond for an aol.com sender than a normal sender, hoping that that would make it easier for distant family members to reach me without significantly increasing the amount of spam I receive.

More significantly, it is possible to adjust the bond value based on the content of the message being sent. If the message appears to be something of value to me, I can charge a lower bond amount than if the message is on a topic I care little about. This can function as a signal to some marketers that I am interested in a given topic also.

A method of doing this would be to use some sort of machine classification system (such as bayesian filters) to classify the incoming email into multiple buckets (car ads, sci-fi ads, viagra ads, etc.). Then allow the user (recipient) to assign a value (weight) to each of these classifications. The server then would use this machine classification to classify the incoming email, combine the user's specified weights, and come up with a bond value. For example, given the three categories "car ads", "sci-fi ads" and "viagra ads", I can assign these categories weights of (0.10, -0.50, 0.20) indicating that I like sci-fi, but am not interested in cars or viagra. The server would then increase the bond by $0.10 if it things the email is a car ad, decrease the bond by $0.50 if it thinks it is a sci-fi ad, and increase the bond by $0.20 if it thinks it is a viagra ad. It can also do a combination of the above, like decreasing the bond by $0.40 ($0.10 - $0.50) if it thinks it is an ad for a sci-fi car. (Other combinations, such as sci-fi viagra ads, are left up to the reader).

It is also possible for the machine classification system to assign a "probability" value to each category. For example, this classification system could say that the email is a car ad with probability 10ad with probability 1would be the base bond + (0.10*0.10 + 0.01*-0.50 + 0.80*0.20) which would be the base bond

+ $0.165.

Also note that as a signal, it is possible for the recipient to change these values. When my car was stolen last year, it would have been nice to be able to change the "car ads" weight from a positive to a negative value, signalling to potential marketers that I am now interested in purchasing a car. (I did sign up for a number of car ads, and when I did finally buy a car, it would have also been nice to return the "car ads" value back to the old value after buying the car to hopefully stop the flow of the now unwanted ads.)

It is important that all of this variable bond adjustment happen automatically on the server without any user intervention, as requiring user interaction in this would defeat the purpose of the attention bond mechanism.

G Contact Management

There are a number of additional ways of automating the use of the whitelist that goes along with the attention bond mechanism that could be very useful. The primary ones we consider here are expiration dates, letters of reference and sliding windows.

The whitelist can have "expiration dates" associated with entries in the whitelist, where after that date the sender is no longer whitelisted. This is useful for interactions with are temporary in nature and not necessarily permanent.

Additionally, it could be possible to have "letters of reference" that allow people to temporarily be placed on your whitelist. For example, if someone on your whitelist that you trust sends you an email and CC's a third party, that third party can gain a temporary whitelist entry which would allow them to communicate with you for at least a little while (the duration of the conversation hopefully). This would function as a letter of reference from a trusted third party.

Also, it could be possible to have a sliding window in use with the expiration dates of entries. This sliding window would have two facets. First of all, the window would have a base length (say for example one month) where once a person gets added to the whitelist, the expiration date is set to the length of that window in the future. Then, each time the person interacts with you that window moves (slides) forward, setting a new expiration date one length of that window ahead of the last interaction. That way continued interaction is allowed, and a person can stay on the whitelist as long as the interaction continues.

Secondly, the length of this window can also be a function of the interactions. If you have a very dense (timewise) set of interactions (you correspond a lot in a short period of time), then possible the length of the window can be elogated such that the person can go longer in time without contacting you before their whitelist entry expires. Someone you correspond with very frequently, such as family, will end up with very

long windows, and people who you correspond with infrequently will have fairly short windows.

VII Caveats

While we believe the use of attention bonds represents the best approach to spam, we have several caveats. First, the mechanism does not dominate the perfect filter with all value distributions. It is particularly strong when the bulk of the distribution is of negative value and there is significant sender surplus to be transferred to the recipient, but it can do harm for primarily desirable distributions.

Due to risk aversion with some senders, some email that is potentially valuable to the recipient but of little perceived value to the sender will not be sent. (In favor of attention bonds, a positive correlation between values will reduce this loss.) Organizations often have sales or marketing related inquiry addresses, suggestion drop boxes (anonymous tip lines), or may otherwise value inbound information to a degree where creating **any** restriction or additional barrier for the sender is unacceptable. The ABM may not be appropriate for these addresses.

While we do not specifically analyze transaction costs – the bond exchange process is assumed to have negligible costs – the model can easily accommodate them. Either or both c_r and c_s can be increased by a multiplier $\Delta > 1$ in Equation 6. Alternatively, a constant cost or a function of bond size can be split between sender and receiver in some proportion. The results are qualitatively unchanged; if transaction costs are non-zero, then additional fees reduce the size of the region where the attention bond dominates the perfect filter.

Adoption, protocol, and infrastructure issues are likely to be significant, but in the interest of brevity are not discussed here.

VIII Social Benefits

The main benefits of the Attention Bond Mechanism are the ability to cause those who would misuse communications channels to reveal their intentions and the ability to improve to social welfare. Those individuals intending to send spam are unlikely to warrant that their messages are not spam. As modeled, the screening mechanism offers a strict Pareto improvement relative to no intervention and a range of potential improvement relative to even an ideal or perfect filter. By making markets instead of foreclosing them, however, the ABM has several benefits beyond the scope of the analytic model.

A Availability of Contact Information

Spiders and web crawlers mine web pages for legitimate addresses in order to send them spam (Tompkins and Handley, 2003). If communications from strangers are mostly wasteful, then recipients may prefer to hide their contact information. Survey statistics bear this out. According to the Pew Internet survey (Fallows, 2003), 73% of the 2200 people polled stated they avoid giving out their email address, while 69% avoid posting their email address on the web. In contrast to the alternatives, a successful screen raises the expected value of communication from unrecognized senders, motivating email users to publish their contact information. This reduces search costs and facilitates valuable interaction among strangers.

B Generality of Mechanism

The ABM is a general economic mechanism for allocating attention and should be applicable to many forms of interrupt-capable communications media, such as email, telephone, instant messaging, and SMS (mobile phone) messaging. For instance, Fahlman (Fahlman, 2002) provides the primary example as the use of a binding interrupt fee to block telemarketing calls. A telco could implement the bond negotiation and collection mechanism right on its switches. The existing infrastructure could then be used to 'bill' the caller and transfer the bond amount to the account of the subscriber. By brokering transactions, the telco's switches effectively become a marketplace.

C Cheaper Channel Costs

The economics of the Attention Bond Mechanism compare favorably with those of competing communications channels. Direct marketing through traditional mail, for example, incurs costs of printing and postage. For a simple postcard, total costs average 31-37¢[8]. Assuming that a reasonable but modest bond were always seized, the proposed mechanism might nevertheless have these advantages: (i) total costs could be lower (ii) recipients receive the benefit of an expected bond instead of dissipating its value through transaction costs in printing and postage (iii) there is reduced environmental waste (iv) a recipient's response, if it occurs, is easier via email and the Internet than traditional mail, and (v) senders learn whether a given message was opened, providing useful information for future correspondence.

D Reduced Arms Race

Even as filters improve, clever misspellings, unrelated subject lines, text hidden in images, and programs designed to thwart challenge-response systems all show the escalating resilience of spam technologies to spam filters. Rather than relying on ex ante classification of message type, the proposed incentive mechanisms

relies on ex post verification of a message's value. Since ex post value is harder to fake in the eyes of a recipient, the problem of defeating a filter becomes less of a technological arms race. For filters, diversity of communications content also increases problems of false positives and false negatives. In contrast, the ABM scales with little social efficiency loss to accommodate more diverse distributions.

E Political Speech

One little explored consequence of filtering technology is its effect on speech. Successful filters have the potential consequence of indiscriminantly eliminating the good along with the bad and the tasteless. For example, in contrast to the 89% percent who consider unsolicited commercial email to be spam, only 65% consider unsolicited email from non-profits or charities to be spam (Fallows, 2003). Because filters provide unilateral veto without subsequent negotiation of favorable exchange, even legitimate and motivated political movements or interest groups would find it difficult to bypass or negotiate filtering. Use of the ABM, however, allows a motivated group to reach an audience. We grant that political speech represents one form of communication where the no intervention case may be more attractive although it may mean incessant spam. Relative to filters, however, attention bonds offer less disruption in valuable communication.

F Individual Tailoring

Although not modeled explicitly, it is easy to see that the size of screen and the seize policy can be functions of other variables. This allows the mechanism to adjust dynamically to individual tolerance of interruptions, opportunity costs of time, sizes of social networks, a desire to inconvenience the fewest senders, etc. Such external factors must be learned in the case of filters and are difficult to incorporate in the case of certain taxes and other spam proposals.

IX Conclusions

Our principal finding is that for a wide variety of plausible conditions, signaling and screening mechanisms dominate mechanisms whose chief purpose is to block or ban email exchanges. In particular, welfare can improve both collectively and for those recipients that filters and legislation are designed to protect.

The mechanism works by forcing unrecognized senders to act on their private knowledge of their own distribution, valuable or wasteful, as it applies to their intended communication. For communication of low value to recipients, the mechanism enforces a wealth transfer from senders to recipients. Communication of high value to recipients is delivered with little exposure to the sender. The net result is that well-targeted communications behave analogously to direct mail advertisements. An added benefit, however, is

that resources consumed in the physical mail channel as transaction costs – marginal costs of printing and posting bulk mail – are instead captured as value to recipients. This is incentive compatible for recipients while bi-directionality allows information and wealth transfers in either direction, helping to promote valuable transactions rather than veto them. The mechanism also depends on ex post verification not ex ante classification so it suffers less from deceitful, and costless, signals in subject lines. Finally, the mechanism allows dynamically adjusted prices, accounting for recipients' value of time.

Acknowledgements

We are very grateful for the assistance of Mark Benerofe in revising and clarifying the ideas presented here. We also appreciate the feedback on early versions of this paper from Bob Marinier, Terrance Kelly, Dan Silverman, Jussi Keppo, Peter Honeyman, Paul Laskowski, and Lori Cranor.

References

Abadi, Martin; Birrell, Andrew; Burrows, Mike; Dabek, Frank and Wobber, Ted, "Bankable Postage for Network Services." In "Advances in Computing Science ASIAN 2003," Springer-Verlag, 2003, Lecture Notes in Computer Science.

Akerlof, George A, "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics*, *84*(3), pp. 488–500, 1970.

Ayres, I and Nalebuff, B, "Want to Call Me? Pay Me!" In "Wall Street Journal, October 8 2003," 2003.

Benerofe, Mark. 2003, personal Communications.

Brightmail Inc., "Spam Attacks and Spam Categories." 2003, `http://www.brightmail.com/spamstats.html`.

Fahlman, Scott E, "Selling Interrupt Rights: A Way to Control Unwanted E-Mail and Telephone Calls." *IBM Systems Journal*, *41*(4), pp. 759–766, 2002.

Fallows, Deborah, "Spam: How it is hurting email and degrading life on the Internet." 2003, `http://www.pewinternet.org/reports/ pdfs/PIP_Spam_Report.pdf`.

Firestone, David and Hansel, Saul, "Senate Votes to Crack Down on Some Spam." In "New York Times, October 23, 2003," 2003.

Friedman, E and Resnick, Paul, "The Social Cost of Cheap Pseudonyms." *Journal of Economics and Management Strategy*, *10*(2), pp. 173–199, 2001.

Glassman; Manasse; Abadi; Gauthier and Sobalvarro, "Escrow services are financially viable for sub penny transactions: The Millicent Protocol for Inexpensive Electronic Commerce." In "Fourth International WWW Conference," 1995.

Graham, Paul, "A Plan for Spam." 2002, `http://www.paulgraham.com/spam.html`.

Hansell, Saul, "Diverging Estimates of the Costs of Spam." In "New York Times, August 28, 2003," 2003.

Hermalin, Benjamin and Katz, Michael, "Sender or Receiver: Who Should Pay to Exchange an Electronic Message?", 2003.

Hopkins, Jim, "Anti-spam Start-Ups Cash In as Junk E-Mail Grows." In "USA Today, July 21, 2003," 2003.

Kraut, R; Sunder, Shyam; Morris, J; Cronin, M and Filer, D, "Markets for Attention: Will Postage for Email Help?" In "ACM Conference on CSCW," 2003, pp. 206–215.

Schwartz, Evan I., "Spam Wars." *Technology Review*, *106*(6), pp. 32–39, 2003.

Shapiro, Carl, "Consumer Information, Product Quality, and Seller Reputation." *The Bell Journal of Economics*, *13*(1), pp. 20–35, 1982.

Spam Laws. 2003, `www.spamlaws.com/state/index.html`.

Spence, Michael, "Job Market Signaling." *The Quarterly Journal of Economics*, *87*(3), pp. 355–374, 1973.

Tompkins, Trevor and Handley, Dan, "Giving E-mail back to the users: Using digital signatures to solve the spam problem." *First Monday*, *8*(9), 2003, `http://firstmonday.org/issues/issue8_9/tompkins/index.html`.

Zandt, Timothy Van, "Information Overload in a Network of Targeted Communication.", 2003, RAND.

Notes

*tloder@umich.edu, mvanalst@umich.edu, rwash@umi ch.edu This material is based upon work supported by the National Science Foundation under Grant No. 0114368.

[1]This material is based upon work supported by the National Science Foundation under Grant No. 0114368.

[2]Criminal Spam Act of 2003 (S.1293), SPAM Act (S.1231), REDUCE Spam Act of 2003 (S.1327), Wireless Telephone Spam Protection Act (H.R.122), Reduction in Distribution of Spam Act of 2003 (H.R. 2214), CAN-SPAM Act of 2003 (S.877), Anti-Spam Act of 2003 (H.R. 2515), Computer Owners' Bill of Rights (S.563)

[3]http://www.mailblocks.com/

[4]http://www.brightmail.com/

[5]http://www.cloudmark.com/

[6]http://razor.sourceforge.net/

[7]http://www.hashcash.org/

[8]via personal correspondence with Mark Benerofe, former SVP, Priceline.com