Impact of Spam Exposure on User Engagement

Anirban Dasgupta[†], *Kunal Punera*[‡], *Justin M. Rao*[∓], *Xuanhui Wang*[§]
[†]*Yahoo! Labs, Sunnyvale CA*
anirban@yahoo-inc.com
[‡] *RelateIQ Inc., Palo Alto, CA*
kunal.punera@utexas.edu
[∓]*Microsoft Research, New York, NY*
justin.rao@microsoft.com
[§]*Facebook, Menlo Park CA*
xuanhui@gmail.com

Abstract

In this paper we quantify the effect of unsolicited emails (spam) on behavior and engagement of email users. Since performing randomized experiments in this setting is rife with practical and moral issues, we seek to determine causal relationships using observational data, something that is difficult in many cases. Using a novel modification of a user matching method combined with a time series regression on matched user pairs, we develop a framework for such causal inference that is particularly suited for the spam exposure use case. Using our matching technique, we objectively quantify the effect that continued exposure to spam has on user engagement in Yahoo! Mail. We find that indeed spam exposure leads to significantly, both statistically and economically, lower user engagement. The impact is non-linear; large changes impact users in a progressively more negative fashion. The impact is the strongest on "voluntary" categories of engagement such as composed emails and lowest on "responsive" engagement metrics. Our estimation technique and results not only quantify the negative impact of abuse, but also allow decision makers to estimate potential engagement gains from proposed investments in abuse mitigation.

1 Introduction

Over the last several years, as email has steadily become the dominant mode of text-based online communication, unsolicited bulk email, generally referred to as "email-spam" or simply "spam", has increased in lockstep [33]. By some estimates the total fraction of all emails that can be considered spam is higher than 90% [33, 10]. Moreover, while email-spam began as a way for unscrupulous marketers to advertise their products, it has now become the main vector for phishing [4, 14], installing malware, and stealing information [22]. In short, email-spam has morphed from being a mild irritant to an outright danger to the users.

This has led to major efforts both in the industry and the research community to develop better spam filters [5, 12, 13, 39, 40]. However, spammers are known to quickly adapt their email messages in order to circumvent these filters [16]. This has resulted in an adversarial game of "cat-and-mouse" between email service providers (ESPs) and spammers: (1) Spammers send out bulk emails designed to bypass the spam filters of major email service providers; (2) In time, spam filters adapt using machine learning and crowdsourcing techniques and block the offending emails; (3) Spammers re-tune message content, change the sending locations and so forth, and the cycle continues. This results in email-spam reaching user inboxes for the duration between the bulk mails being sent and the spam-filters adapting. Unfortunately, even though filters have improved dramatically, spam is so cheap to send that the required conversation rates for profitability, which are below 1 in 5 million, can still be sustained [22].

Barring some fundamental change in the spam market (such as legal or technological solutions), the chief way to combat spam is to invest more resources to make the spammers' response cycle less economically viable, which would force some spammers out of the market. Characterizing this ecosystem is thus essential not just for making both policy decisions but also in making decisions that on the surface seem to be purely machine learning in nature—e.g. how to design spam filters that exploit signatures that are the hardest to game.

Although qualitative arguments about spam being a negative social externality have been often made, it is much harder to quantify the intuited numbers [1, 21, 27]. Since botnets form the main spam-delivery infrastructure, researchers interested in understanding the economics of spam have made significant efforts in understanding the market behind the creation and renting of botnets [32, 41, 3]. Kanich et al. [23] measure how successful product-oriented spam ultimately is in marketing

and selling the corresponding products. Similar studies have provided quantitative estimates on the economics of account phishing [17], the market behind "human-farms" [31] and malware distributions [6]. Rao and Reiley [34] review a large fraction of this literature from an economic perspective. Such quantitative studies have collectively thrown valuable light on various aspects of the underground economy, thereby providing guidance to both the policy-designers and designers of spam-filters.

Given this extensive literature, it is perhaps surprising that seemingly little attention has been paid to the interplay between email users and email service providers and the associated responses to problems of email-spam. For example, we are not aware of any work that quantifies the long-term effects of spam reaching the inbox on user engagement. In terms of the interplay, changes in user engagement have a direct impact on ESP revenue and are thus an important decision metric for anti-spam investment. Economic theory tells us that a profit maximizing firm will invest in anti-spam technology only if there is a compensating return in terms of increased user engagement or retention. For instance, simply because we all think spam is a bad thing does not mean service providers will go broke fighting it! Being able to provide a quantitative estimate on how the long-term user engagement is affected as a result of spam would provide an added concrete incentive for the ESP to fight spam.

Some econometric studies [7, 42] have approached the problem from the firm perspective (the client of the email provider) and have shown that spam has a significant cost in terms of the working time spent by users in dealing with email. In particular, Caliendo et al. [7] use a survey approach and find that the average employee in their sample spent 1200 minutes per year in dealing with spam. However, these small-scale studies cannot quantify the effect of spam on *longer-term user engagement*. Does getting more spam cause a user to stop using the email service? It seems intuitive to assume "yes". However, it has never been established whether this causal effect exists, *how strong the effect is* if it exists, *what types of engagement* would it affect, and *how to measure this in a statistically robust manner*. More explicitly, answering these questions is useful for multiple reasons—it helps our broad understanding of the total negative externality of spam, which could potentially have implications in deciding how to deal with spam at the policy-level. Also, as spam filters get better, making additional improvements in spam catch-rate becomes harder and hence more expensive, and often involves difficult trade-offs either regarding total investment or about false-positive rates (i.e. in deciding the operating point of the spam classifiers). In terms of social efficiency spam is clearly a negative [34]—the consensus view is that spam should be mitigated far below current levels in order to raise social welfare because the social costs of spam clearly outweigh the monetary returns from spamming. However, since the government cannot compel ESPs to invest more heavily in anti-spam technology, obtaining estimates of the negative impact of spam, such as ones in this paper, is important. Accurately quantifying the impact of spam allows firms to make informed, well-targeted investments. In turn, these investments can potentially lead to improvements in service quality for the end-users. While our study does not provide authoritative answers to all these questions, it certainly builds many of the tools and the necessary formalizations for it.

The gold standard for estimating causal effects is randomized experimentation, also referred to as "A/B testing" [24]. If we can expose users to spam completely at random, then we can safely assume that any effect we observe is due to spam. In the real world, however, performing such experimentation is difficult because exposing users to spam is problematic for both user experience and the ESP's reputation. Estimating causal effects is typically difficult in the absence of randomized experiments because most actions reflect something about the user in terms of their type or future intentions. These circumstances lead to the classic problem of correlation in the absence of causation. For example, since users tend to get spam when they give out their emails to third party services and active users tend to do so more often than less active users, a naive plot of engagement-vs-spam would show activity and spam exposure being positively correlated.

An alternate method of estimating such effects is to conduct in-depth surveys or in-lab tests of a smaller set of users. In-lab methods are inadequate for our problem as we are looking to estimate potentially small, but long-term effects. The size of the surveys or lab studies is necessarily limited by cost, which makes it hard to estimate small and long-term effects. More importantly, what users report in a survey may not be reflected in their actual behavior. In particular, rounding error can severely bias estimates. For example, answering in a survey that one spends 5 minutes a day dealing with spam might seem like a "small" amount, but over the course of a year, that is 1250 minutes, or about 20 hours. For a $30 an hour employee, this means it is a $600 per year problem. If the true value was 1.5 minutes, but the user rounded up, the resulting estimate could be off by a wide margin.

An extensive literature in econometrics has focused on developing techniques such covariate matching, regression-coefficient methods, bias reduction, neighbor matching, propensity score matching (PSM) etc. [35, 20, 9, 29] to deal with selection bias in observational data. Among these, PSM and neighbor matching techniques are considered more robust in estimating effects

of a categorical treatment variable [29] than regression-coefficient methods—in both of these the intuition is to be able to match a untreated user with a treated one based on a set of pre-defined user attributes. PSM creates the matching using only a single propensity score that is obtained by a weighted combination of the user attributes—the weights are learnt by modeling the exposure treatment as a categorical variable directly using a first stage logistic (or similar) regression. For nearest neighbor methods user matching is done by treating them as points in a high dimensional space. It is commonly believed that PSM is more robust than nearest neighbor matching methods when the number of user attributes is large since finding nearest neighbors in high dimensions is not robust (see e.g. [29] for detailed discussion). Yet, for PSM one has to assume that the first stage regression is correctly specified. The non-parametric nature of nearest neighbor matching methods makes them more reliable with respect to the fact that one does not have to correctly specify a first stage regression—in small samples and with high data dimensionality, the benefits of PSM outweigh the drawbacks.

In our setting, the popularity of Yahoo! Mail gives us a huge set of users to match over, compared to the number of user attributes. Also, existing PSM methods typically assume the ability to model the probability that a particular user falls into the *categorical* "treatment" group. However, in our application, spam exposure is a continuous variable, leaving the treatment group ill-defined, and hence this assumption fails. For both reasons, the nearest neighbor matching is more appropriate in this setting.

In this paper we describe a large-scale *nearest neighbor matching* method to infer causal relationship from observational data for which the exposure is a continuous variable. We apply this technique to the spam-engagement setting. Overall the results provide strong empirical support for the commonsense notion that spam has a negative impact on user engagement. We provide quantitative estimates that show that the impact of spam in the inbox can have serious revenue implications and can contribute to a large percentage drop in user engagement. The effect is largest for more "volitional" user activities such as composing and sending emails. The function mapping spam changes to engagement appears to be convex, with the marginal impact increasing with the size of the exposure change. User characteristics are not particularly informative in predicting the response to spam — notably light users are equally affected in absolute terms by a piece of spam in the inbox, meaning that percentage-wise the impact is far greater for these users. Thus, although the intuition that spam causes decreased user engagement is commonplace, the main insight supplied by this study is to extend and formalize this intuition in a quantitative way.

Our Contributions.

- We conduct a principled and thorough study of the causal relationship between spam exposure and long-term user engagement. We find that, indeed, exposure to spam results in long-term reduction in user engagement in terms of logins, page views, and emails sent. As far as we know, this is the first such study to *quantitatively* establish this link between spam exposure and user engagement.

- We propose the use of a variant of propensity score matching, namely *nearest neighbor matching*, in combination with regression based techniques in establishing causal relationships in large-scale observational data settings when the exposure metric is continuous. This contribution of our paper is of interest *independent* of its particular application in this study. Our simulations (described in the Appendix) indicate that this method is indeed superior to (variants of) propensity score matching for continuous exposure metrics.

Organization. In Section 2 we present our approach for estimating causal relationships in large-scale observational data settings. Then in Section 3 we instantiate our proposed approach to the case study of estimating the effect of spam exposure on long-term user engagement. The results of this case study are given in Section 4. In Section 5 we review prior work in causality estimation and spam exposure studies. In Section 6 we conclude. Finally, in the Appendix we compare our proposed methodology with variants of propensity score matching and on simulated data show that our approach performs better at estimating a hidden relationship between variables.

2 Measuring the Effect of Spam on User Engagement

In this section, we first define the problem of estimating the effect of spam exposure on user engagement. We start with a description of the aspects of the problem that make it unique from other works in measuring effects. We then present a formalization of the continuous exposure setting and describe how to map our problem to this formalization.

2.1 Aspects of the Problem Setting

Our problem of measuring engagement as a function of spam exposure has the following characteristics that make it unique, and hence requiring modifications to established methodology.

Continuous Exposure: In our problem, the exposure variable is continuous—there is no clear definition of a "treatment" vs. "control" group. We cannot identify a set of users and consider them as "treated," i.e. having been sufficiently exposed to spam because nearly everyone is exposed to some degree. One solution would be using an arbitrary threshold to define a treatment class. But in some sense this is just asking the same question back again: what is a critical level of spam such that a person receiving that amount can be considered to be sufficiently exposed? Thus, the continuous exposure is not just an artifact of the data, incorporating that into the modeling and estimating process is absolutely essential.

Engagement as a function of Exposure: Having defined exposure to be a continuous variable, computing a single number as the expected size of the effect is not meaningful any more. Instead we want to answer the following question: what is the expected effect if the amount of exposure is increased by an amount Δs. We intend to approximate the function that captures the change in the effect as a result of the change in the exposure for an average user.

Infeasibility of Randomized Testing: Randomized experiments are clearly the gold standard for measuring effects. Suppose we intend to estimate the effect on a user receiving Δs more spam messages in a month. Ideally, we would be able to select a small random set of users, and then tune their spam filters such that they receive Δs more spam for this month. We could then measure the resulting effect against a randomized control group.

For the spam-setting, however, performing such experimentation is difficult on many levels: (1) exposing users to spam is problematic from both a user experience and Yahoo!'s reputation point of view. The negative effects of spam does in fact often extend beyond a minor nuisance, since a majority of these messages contain URLs that tempt users to either conduct commercial transactions or to give out their personal information; (2) even if we could filter out the most pernicious types of spam, the revenue risk associated with user defection would cause the size of the study to be limited, both in terms of the amount of exposure and the number of users; 3) spam that does leak into inbox is, by definition, currently undetectable before the user has interacted with it. Thus, any randomized experiments would have to account for exposure of this kind anyway.

2.2 Formal Problem Definition

We now define the problem formally and point out the empirical quantities for which we would like to create unbiased estimators. Suppose for each user i, $\mathbf{x}_i$ denotes the set of features we observe. Let s_i denote her exposure variable and y_i denote the response (or effect) variable. Note that s_i is continuous. If we want to study the impact of spam on the user, then the exposure variable would be the amount of spam received by the user in a particular time period, the same for all users — we call this the *exposure period*. Abusing notation, we write $y(\mathbf{x}, s)$ to denote that the response is a function of the user features and the exposure. Let Δs denote a certain amount of change in the exposure variable, and $\Delta y(\Delta s)$ denote the function that measures the *average* change in y due to an increase Δs in the exposure. Formally, we define $\Delta y(\Delta s)$ as follows. Let $\mathrm{E}[\cdot]$ denote the expectation operator.

$$\Delta y(\Delta s) = \mathrm{E}_{(\mathbf{x},s)}[y(\mathbf{x}, s+\Delta s) - y(\mathbf{x}, s)]. \quad (1)$$

The expectation in the above expression is taken over all the user features and all the previous value of exposure. This of course is not an observable quantity, since one user has only one value of s. Thus, a more feasible quantity to measure is the following – difference over pairs who differ only in exposure, but have the same feature vector.

$$\Delta y(\Delta s) = \mathrm{E}_{i,i'}[y_i - y_{i'} | \exists(\mathbf{x}, s_i, y_i), (\mathbf{x}, s_{i'}, y_{i'}), s_i - s_{i'} = \Delta s] \quad (2)$$

Note how this quantity generalizes the effect measurement for binary treatment variables. If $s \in \{0,1\}$, then the standard question of measuring the average treatment effect would be

$$\begin{aligned} & y(s=1) - y(s=0) \\ & = \mathrm{E}_{\mathbf{x}}[y_i - y_{i'} | \exists(\mathbf{x}, s=1, y_i), (\mathbf{x}, s=0, y_{i'})] \end{aligned}$$

In our case, we are thus interested in the function $\Delta y(\Delta s)$ instead of a single value that measures the treatment vs. non-treatment. This makes the application of the standard propensity score matching techniques [35] impossible: we can no longer define a treatment class.

One naive way of creating the estimate would be to compute the following difference—essentially just take the differences in the effect levels of users whose exposure is s and those whose exposure is $s + \Delta s$.

$$f(\Delta s) = \mathrm{E}_i[y_i \mid s_i = s] - \mathrm{E}_i[y_i \mid s_i = s + \Delta s]$$

But this would be the wrong quantity, since conditioning on the fact $s_i = s + \Delta s$ is different from conditioning on $s_i = s$ (the corresponding distributions of $\mathbf{x}$ and hence $y(\mathbf{x}, s)$ are different), and thus the above difference does not measure what would happen to the average person if the exposure suffered by that person increased by Δs.

Nearest Neighbor Matching. The essence of nearest neighbor matching is that we can approximate the equation 2 by the following one.

$$\Delta y(\Delta s) = \mathrm{E}_{i,i'}[y_i - y_{i'} | s_i - s_{i'} = \Delta s, \mathbf{x} \approx \mathbf{x}'] \quad (3)$$

where $\mathbf{x} \approx \mathbf{x}'$ denotes that $\mathbf{x}$ and $\mathbf{x}'$ are approximately similar, instead of being exactly same. The variants of this definition of approximate similarity define the different variants of the nearest neighbor matching algorithm.

Suppose we have a particular matching function, in which, for each given user $i = (\mathbf{x}_i, s_i, y_i)$, we can find out a set of users N_i such that for each $j \in N(i)$ satisfies $\mathbf{x}_j \approx \mathbf{x}_i$. Further define $\mathbf{1}(X)$ to be the indicator vector for the event X, in particular let $\mathbf{1}(s, s')$ denote the indicator vector such that $|s - s'| = \Delta s$. If there are n users overall, our empirical estimator for the quantity in equation 3 is then given by

$$n(i, \Delta s) = \sum_i \sum_{j \in N(i)} \mathbf{1}(s_i, s_j)$$

$$\Delta y(\Delta s) = \frac{1}{n} \sum_i \frac{1}{n(i, \Delta s)} \sum_{j \in N(i)} \mathbf{1}(s_i, s_j)(y_i - y_j)$$

Essentially, in each neighborhood $N(i)$, we compute the average effect due to an increase of Δs exposure and then average these effects over all the points to get the average effect.

3 Data, Features and Matching for Spam

In this section we describe how to apply the above matching technique for the spam exposure case study. We start with a summary of our overall method. In order to measure how engagement is affected by spam exposure we first need to specify how to measure *user engagement* and *spam exposure* for a user. We then describe *how to create matchings* between users based on user behavior features.

3.1 Technique Summary

In order to measure the effect of spam exposure on user engagement we first create a set of behavioral features per user for a 2 month period, called the "matching period." These features are then used to create matchings between users. We then observe the spam exposure of these users on the exposure month (month 3) immediately following the matching period. Due to random variation in spam, the two users in a match are often exposed to different amounts of spam (Δs). We then examine how Δs impacts behavior in the observation period immediately following the exposure month. We look at difference in engagement for both the short-run (only month 4) and long-run (months 5-6), while controlling for how these differences persisted within the pair (e.g. higher month 3 spam likely means higher month 4 spam; in estimating month 4 engagement, we will control for this difference).

The attribution of causality depends on the assumption that within each pair of users, month 3 spam exposure is random. This is known as the "selection on observables" assumption. In general, spam exposure is correlated with user activity. Using your account more actively tends to get the email address "out there" more, making exposure to spam non-random. For example, in a cross-section of users, light users tend to get less spam than heavy users. This is precisely the reason we need to use the matching methodology to estimate causal effects (and overcome spurious correlation). In our case, we match on both the level and linear trend of usage. So the identifying assumption stated more precisely is: conditional upon the level and trend of usage (on all 14 matching criteria) over two months, the spam exposure difference between users within a pair in the following month is related to future usage in only the following ways (a) the direct impact of past spam exposure; (b) the indirect impact of past spam exposure (higher spam today, might mean higher spam tomorrow, which we must control for).

3.2 Data Description and Matching Attributes

Our data comes from the Yahoo! Mail logs of user activity.[1] To ensure accurate results, we first cleaned the data of accounts that were potentially corrupted by phishing attempts or spambots. We dropped any user who showed a change in more than 4 sent messages a day (in average) between the matching months (months 1-2) and the target months. This number was chosen based on an analysis of the distribution to determine what qualified as an improbable outlier. We also dropped a pair of users that had a Euclidean match distance of greater than 0.1 to ensure that we were always very close matches. Finally, we dropped all users that showed near zero mail page views in the matching month(s) and outliers (+3 standard deviations). The former is to increase the strength of our estimator, as it is unreasonable to assume spam impacted a user that never logged in, the latter to reduce the influence of high leverage anomalies.

After performing all the cleaning operations, we took a large random sample of 500,000 users for 6 months, and generated the following features per user per day: *all inbound mail, classified spam, total sent mail, composed mail, replies, forwards, mail time spent, all page views on Yahoo! site, all time spent on Yahoo! site, delete without reading (messages that are removed from the inbox without reading), deletes, spam votes and non-spam*

[1] Note that this is purely observational data, no active experimentation or bucket-testing was involved. Furthermore, we use only behavioral statistics aggregated at the anonymized user level. Thus there are no privacy issues related to email content, or the graph of user-user communication.

votes.

To ensure that the matching generated very similar users, we used all the 14 features over 2 months to compute nearest neighbors. In addition, we also ensured that the matched user accounts were registered in the same year. We performed the matching process over the entire mail sample, thus enabling a small enough distance threshold. As a result of the matching, we end up with 486,102 matched pairs (one user could be considered in multiple pairs, and not user-user pairs qualify for matching, as we see below). Using the first two months of data for the matching period ensures that each pair of users had the same level of usage and the same (first order linear) trend.

3.3 Metrics for User Engagement

Yahoo! Mail users interact with the web user interface in a variety of ways. Users can login into the interface and just glance at the list of emails in the various folders ("boxes"), can click on individual emails to open them in a separate panel for reading or delete it without reading. Other email related actions that are instrumented include replying to individual emails, or forwarding them, composing new emails and marking emails as spam or non-spam. Each of these actions represents a different kind of engagement, and naturally certain forms of engagement are more significant than the others. From a short-run revenue calculation perspective, the page view is the primary quantity of interest, as page views can be easily converted to a dollar figure based on the advertising monetization rate. But not all page views are created equal. For example, we have found that the number of sent mails (and resulting pageviews) is a more reliable predictor of future engagement than the pageviews resulting from simply reading mail or reloading one's inbox. The reason is likely that sending mail both leads to more mail in response and signals that the user is using the account as her primary email. We thus look at a variety of such metrics to measure engagement.

3.4 Quantifying Spam Exposure

Yet another critical point in our study is how to quantify the spam exposure of a user. Typically, the spam that a user has been exposed to lands in her inbox does so precisely because the filters have been unable to recognize it as spam. Consequently, this number is hard to measure for a user. We could rely on the "spam votes" of a user a proxy for this quantity, but it is well known that very few users give any votes. In fact, the average Yahoo! Mail user gives less than one vote in an entire year, whereas some users are extremely proactive in marking emails as spam. To complicate matters, even spammers and bot accounts give spam-votes, aiming to subvert the machine-learned filters by providing false examples.

The strategy available to us is to use the number of inbound emails classified by the Yahoo! filter as a measure of the spam targeted towards the user and infer "inbox exposure" from this classified spam. Of all delivered mail (not blocked before connection), more than half is classified as spam and sent to the spambox. The false negative rate relates the spambox quantity to implied inbox-exposure. For example, if the false negative rate is 0.10, then for every 9 messages in the spambox, we expect 1 piece of spam to slip into the inbox. For the empirical analysis, we estimate the false negative rate and use it to infer inbox-exposure, which we will use in all our analysis. Due to confidentiality concerns of Yahoo Inc., we cannot report the exact estimates of the false negative rate, but will describe the process through which we model and infer it.

Estimating the False Negative Rate: We estimate the false negative rate in two ways. First, we utilize daily usage logs of users over a 6 month period. Note that if the false negative rate were 0, then conditional on past behavior, daily spam box quantity should be unrelated to inbox quantity, because there is no slippage. In contrast if the rate is non-zero, increases in the spambox will be positively correlated with increases in the inbox. We estimate this relationship using a regression of inbox quantity on spambox quantity and lagged values of both quantities, all on the daily level. This gives us an estimate, lets call it FN.

To confirm this estimate, we examine how spambox levels correlate with "delete without reading" in the inbox. "Delete without reading" is a strong sign of spam, but many legitimate mails are deleted without reading as well. In fact 53% of all inbox messages are deleted in this fashion. If the false negative rate was 0, then there should not be a relationship between spambox and delete without reading, conditional on inbox volume (inbox volume and spambox volume could be related, so we control for this). We estimate the empirical relationship using a time-series regression and find that 1 message in the spam box leads to $.8FN$ deletes without reading. That is, very close to our initial estimate of the false negative rate using the other methodology and consistent with the idea that not all users simply delete spam, but most do. Given the mutual consistency of both approaches, we proceed with our estimate of the false negative rate in all analysis.

Maintained Assumptions on the False Negative Rate: The assumption of a constant false negative rate might seem too strong when we consider the fact that users have different propensities to sign-up for email mailing lists. In our analysis, however, the individual variations are less important for following reasons. First, we only use this estimate to normalize in the aggregate sense —

obtaining the aggregate inbox-spam in terms of the classified spam. Thus, in our case, all we require is that within a pair of users, there are no systematic differences in false negative rate; this is essentially assured by our bi-directional matching procedure. When examining the differential impact of large increases in exposure vs. small increases (non-linearities), the assumption requires that when a user experiences a large increase in spam, the classification rate stays the same. Indeed, given how machine classification benefits from large quantities, one might think that large quantities of spam are classified with less error. We will see that we actually find an increasing marginal impact of exposure, meaning that either this is not an issue, or the real pattern is even more convex.

The area that is most hampered by the constant false negative rate assumption is the analysis of user characteristics. For instance, if Yahoo! does a better job of classifying spam for older users, then we will overstate the inbox-exposure for these users. In the results section, we note these concerns where applicable.

3.5 Creating the Matching

In this section, we describe the method of nearest neighbor matching that we used. The basic framework is to match users who are very similar to each other in the matching period, and then analyze how their behaviors differ in subsequent time periods. We first discuss how to create the neighborhood set $N(i)$ for each user.

Using kNN for Matching: In order to define the matching, we use two criteria to define the neighborhoods $N(i)$ — a distance based threshold and a k-nearest neighbor based threshold. The distance between the vectors is measured in ℓ_2 norm. We have a distance threshold d that we use to filter our pairs that do not lie within d distance of each other. On top of this, we apply a k-nearest neighbor based threshold – each point i contains no more than k of its nearest neighbors in $N(i)$. This ensures that a dense region of the $\mathbf{x}$ manifold is not over-represented in our estimate.

Using Bi-directional Matching: To avoid bias, we only use bi-directional matches. What this means is that dyad i–j is only included in the analysis if i is j's nearest neighbor and j is also i's nearest neighbor. The nearest neighbor property is not generally bi-directional (i's nearest neighbor might be j, but there is a node closer to j, say r, that is further from i). The most important reason we include only bi-directional pairs is that it ensures that in the exposure period, the average difference within a pair of users is 0 for all attributes we match on, by construction, because the labeling of users within the pair is purely nominal. In our estimation, this means that we can reliably link differences in spam exposure within the pair to differences in engagement, knowing that there is no other reasons for a systematic difference.

An additional reason is that it naturally eliminates a known issue with matching or propensity score estimators that occurs when relatively few users are the "unexposed match" to relatively many exposed users. For instance, consider a job training analysis in which we predict the probability (propensity score) of receiving training. PSM matches a pair in which one person actually received training and one did not, but had similar predicted probabilities of receiving training. By construction, there are relatively few individuals who have a high predicted probability of receiving training but in reality do not receive it. This means that these people are the "controls" for a relatively large number of treated individuals, thus increasing the impact of their behavior on observed estimates. In our routine, we get around this problem by only using bi-directional matches. In our case, the problem that would arise is that some users in the less dense portion of the kNN graph match to users in a denser portion. These users in the less dense portion might be different in ways that induce bias (for instance if they are always slightly more engaged).

Using Locality Sensitive Hashing: Computing the matching efficiently for a large number of data points and a moderately large number of dimensions is a non-trivial task. In order to compute this, we utilize the locality sensitive hashing technique [2]. Essentially, the idea is to compute a hash function h such that the probability of two points falling into the same hash bucket is inversely proportional to the distance between them.

$$\Pr[h(i) \neq h(j)] \propto \|\mathbf{x}_i - \mathbf{x}_j\|$$

We first bucket all points using this hash function and then do an exhaustive search inside each bucket to find the k-nearest neighbors for each point that fall within the distance threshold. We tune our LSH construction such that with high probability we get all neighbors for all points within the distance threshold.

4 Empirical Results

In this section we present the results of our empirical application. We start by linearly modeling the short-run (1-month in the future) impact of spam exposure on the various metrics of webmail engagement. We then examine the effect more closely using a flexible non-linear model. Next, we examine how mail spam impacts non-mail usage of properties on the Yahoo! network of sites (contagion effects). We then proceed to estimate the medium-run (2–3 months) impact of spam exposure on future engagement. Finally we examine how user characteristics modulate the impact of spam.

4.1 Short-run Impact of Spam on Mail Engagement

Estimating Equation: In this subsection, we look at the impact of month 3 spam on month 4 engagement. Recall month 3 is our first post-match month, and thus the first time spam exposure will meaningfully vary within a pair of users. In our baseline specification, for each pair of users i, we estimate the following equation with robust ordinary least squares. Let y equal the engagement metric we are interested in (page views, sent mail, etc) and s the number of spam messages that reach an user's inbox. Let the months be denoted by $1, 2..$ etc. Let Δy_{it}, Δs_{it} denote the differences in the engagement and the exposure metric for the i^{th} user-pair for the t^{th} month. Recall that months 1 and 2 were used to find matching users (thus, the average $\Delta y_{it}, \Delta s_{it}$ values are essentially zero for $t = 1, 2$). We run the following regression to estimate the relation between $\Delta y_{i,4}$ and $\Delta s_{i,3}$.

$$\Delta y_{i,4} = \beta \Delta s_{i,3} + \rho y_{i,4} + \gamma_1 \Delta y_{i,3} + \gamma_2 \Delta s_{i,4} + \gamma_3 \Delta s_{i,4}^2 + \gamma_4 \Delta s_{i,4}^3 + \varepsilon_i$$

This specification controls for month 4 spam exposure using a cubic polynomial and includes a lagged value of the dependent variable, to control for the contemporaneous impact of spam last month and activity bias (see [25]). β is the quantity of interest, as it gives the first order impact of spam exposure on engagement 1 month in the future. Table 1 gives the estimates of β for the our key engagement metrics.

Absolute Impact: As the results in Table 1 show, across all metrics, the relationship between exposure and engagement is consistent with the hypothesis that spam exposure discourages usage. That spam has a negative impact is perhaps obvious; however Table 1 gives a quantitative estimates for all metrics, not just the sign of the effect. In Column (1), we see that the impact of one spam message in the inbox reduces mail page views next month by 0.472 pageviews. For a webmail provider, page views are the primary metric to gage the revenue impact, as they can be converted to dollars based on the ad revenue from each page view. The R-squared numbers show that these regressors account typically account for 10% of the variation in the dependent variable.

However page views do not tell the whole story, as other metrics, such as sent mail, are thought to be better long-term predictors of engagement. In column (2), we estimate that a spam message in the inbox reduces webmail time spent next month by 24 seconds. Column 3 shows that about 1/4 of the page view impact comes through reading fewer messages. Column (4) shows sent mail impact. Sent mail includes composed emails (written from scratch), replies and forwards. Overall, users send much less mail than they receive or read, as mass/automated emails are a large fraction of legitimate email traffic as well. The impact on sent mail is negative with most of the impact coming through composed messages. This makes sense from a disengagement/frustration perspective. One still replies to emails, but perhaps looks for other communication outlets to send new messages if the account is inundated with spam. In Column 8, we see that spam leads to fewer session logins as well.

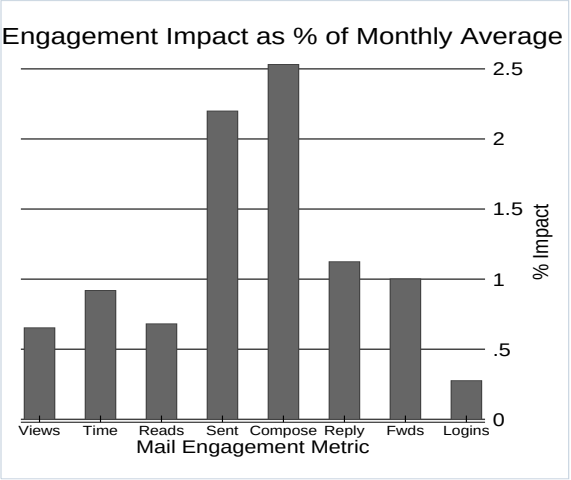

Figure 1: Differential impact of spam exposure magnitude on sent mail and mail page views.

Impact as Percentage of Baseline Usage: In Figure 1, we show the relative size of the impact on each of the engagement metrics. We create this by converting the impact of 1 spam message in the inbox last month, estimated in Table 1, to percentages as a function of the averages for each metric in the matching months. The largest percentage impact occurs for composed messages, consistent with the story that this sort voluntary user engagement is the most susceptible to a negative experience. The percentage impact on composed emails is more than twice as large as the impact on replies and forwards. Monthly "consumption" metrics, views, time spent and reads, show between a 0.5–1% decline as a result of a spam message in the inbox. Logins show the lowest relative impact — although users engage less heavily after spam exposure, in general they still login to the webmail client with close to the same frequency.

4.2 Differential Impact by Exposure Change Size

In the previous section we modeled the impact of spam exposure as a linear function. This was mainly to facilitate interpretation and comparisons across engagement

	Exposure Metric							
	(1) Page Views	(2) Time	(3) Reads	(4) Sent	(5) Composed	(6) Reply	(7) Fwd	(8) Login
Δs_{t-1} (β)	-0.472***	-24.20***	-0.108***	-0.0305***	-0.0251***	-0.00326***	-0.00104***	-0.0572***
	(0.0236)	(1.614)	(0.0250)	(0.00289)	(0.00234)	(0.000912)	(0.000228)	(0.010)
Δy_{t-1}	0.414***	0.483***	0.113***	0.402***	0.335***	0.509***	0.261***	0.74***
	(0.00703)	(0.0185)	(0.0263)	(0.0741)	0.0923)	(0.0341)	(0.0140)	(.0001)
R-squared	0.162	0.177	0.10	0.089	0.065	0.123	0.048	

Table 1: Impact of spam exposure on engagement 1-month in the future. Robust standard errors are in parentheses and *** means p-value < 0.01.

metrics. In this subsection we examine how the impact of the change in spam exposure depends on the *magnitude* of the change. To do so, we make use of the Frisch-Waugh theorem from linear regression [15]. We first regress the exposure metric on the control variables (the variables other than past spam difference) and then take the residual. We then regress the independent variable of interest, last month's spam exposure, on the control variables, and take the residual. The relationship between the residuals of the dependent variable (engagement metrics) and the residuals of the independent variable (last month spam exposure) gives the relationship between these two variables, net of the impact of the control variables.

Non-linear Impact on Sent Mail, Logins and Mail Page Views: In Figure 2 we plot the relationship using a local polynomial smoother (Epanechnikov kernel, bandwidth=10) for three key engagement metrics: sent mail (left axis), mail logins (left axis) and mail page views (right axis). All three metrics display the same pattern. The y-intercept at zero is almost exactly zero for all metrics, which is comforting, because it means that we (correctly) estimate that if a pair has no exposure difference, there is not an engagement difference. This can be seen as a confirmation of the validity of our matching procedure (we also do this via simulation runs in the following section). The slope close to zero is negative, but significantly less than the slope for large differences in exposure — relatively small changes in exposure tend to discourage engagement, but the impact is muted. For all metrics, at about 15 spam messages in the inbox in a one-month period, the negative impact shows a sharp increase (gets more negative). For sent emails and logins, this slope increase levels off near 25 spam messages, but for mail page views, the steep slope persists over all ranges of values for which we have sufficient data.

Key Takeaways: The differential impact in Figure 2 gives insight into how spam negatively impacts the user experience. Note that the x-axis in Figure 2 is the absolute difference in number of spam received by the two users in a pair over 1 month. Small changes in spam

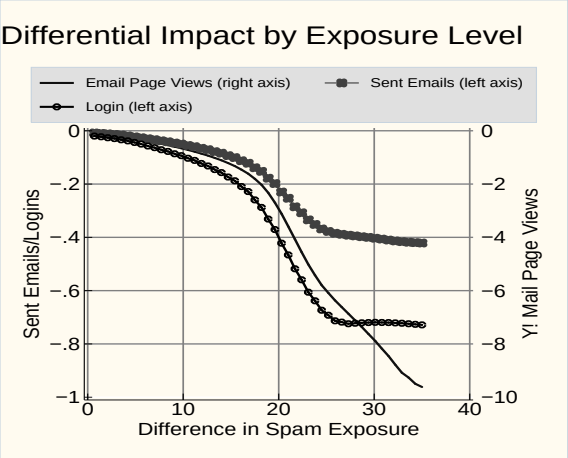

Figure 2: Differential impact of spam exposure magnitude on sent mail and mail page views.

exposure has a muted impact on the user, whereas large changes have a much more pronounced effect. When the increase in spam exposure reaches the level of once every other day, the marginal impact ticks up considerably. This disengagement is likely the result of a disruption of the user experience. Since small changes are less disruptive, the marginal effect is lower. One possible conclusion to draw from this nonlinear trend is the following: it is likely more worthwhile to make a relatively large investment for a big increase in filtration accuracy (and thus obtain a super-linear improvement in engagement), rather than pay a relatively modest sum for an incremental improvement.

4.3 Contagion effects

So far we have documented a negative impact of mail spam on many facets of webmail engagement provided a quantitative estimates the magnitudes. The next natural question is "Does exposure to online abuse in one domain carry over to engagement in a firm's other web properties?" These so-called "contagion effects" or

	Aggregate Effect		Controlling for Mail	
	(1) Non-mail Page Views	(2) Non-mail Time Spent	(3) Non-mail Page Views	(4) Non-mail Time Spent
Contagion effect, Δs_{t-1}	-0.064** (0.03)	-4.33*** (0.03)	-0.0176 (1.72)	-1.470 (1.71)
Δ Mail page views t			0.117*** (0.003)	
Δ Mail time spent t				0.136*** (0.006)
Δy_{t-1}	0.639*** (0.028)	0.640*** (0.027)	0.711*** (0.055)	0.703*** (0.056)
R-squared	0.253	0.214	0.265	0.226

Table 2: Contagion effects of mail spam on other network activities. p<0.01: ***, p<0.05: **.

"brand damage effects" are often used as justification for investment in anti-abuse technology. Our empirical framework allows us to examine this question by looking at engagement across the Yahoo! network of sites.

Contagion Estimates: In Table 2 we estimate the impact of Yahoo! Mail spam on page views and time spent occurring on other parts of Yahoo!. In columns (1) and (2), we do not control for the contemporaneous impact on mail activity – this is why there are empty spaces for these regressors. The estimated contagion effects in this case are negative and statistically significant coming in around 17% (13%) of the direct effect magnitude for time-spent (resp. pageviews), as given in Table 1. In evaluating the revenue impact of a proposed change in the spam filter, these spillover effects should indeed be taken into account. However, to qualify as a pure contagion effect, we would want to be sure they are not mechanically due to lower Yahoo! Mail engagement. The reason is that Yahoo! Mail uses various techniques to get the user to engage with the rest of the Yahoo! network. For example, news stories are shown in the "welcome screen" and there is a web search bar. In column (3) and (4), we control for contemporanous Yahoo! mail usage. Controlling for mail usage reduces the estimated impact of spam exposure by 80% – the remaining figures are no longer statistically significant. The conclusion is that while there measurable spillover effects, the direct cause seems to be lower mail engagement itself. Since mail use creates positive spillovers on the rest of the site, lowering mail engagement has a more than 1:1 effect on engagement. Once we control for this effect, nearly all of the supposed contagion effects go away.

Key Takeaways: Our conclusion is thus that while in the short term there are economically meaningful spillovers of mail spam on the non-mail network activity, the spillovers do not seem to be driven purely by contagion or brand-damage reasons. Rather, they seems to be more mechanically linked to the decreased mail engagement. This is not to say that contagion effects to do not exist, just that in this case they are swamped by the direct negative impact. Our careful analysis allows us to separate these subtle differences.

4.4 Medium-run impact

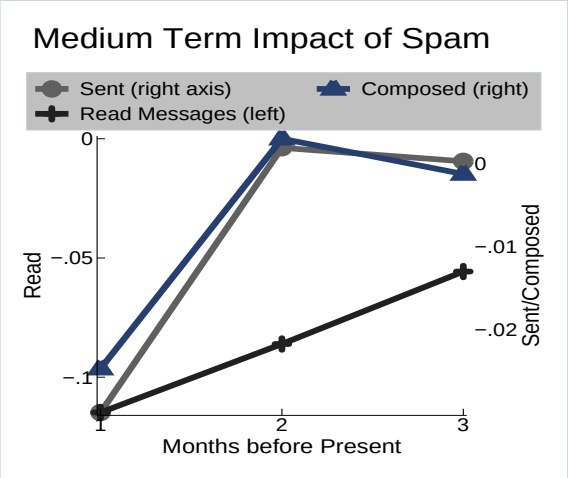

Figure 3: Direct impact of spam on future behavior 1–3 months post-exposure.

In this subsection we examine the impact of spam exposure on engagement up to 3 months in the future. In Figure 3 we plot the impact coefficient of spam exposure on sent mail, composed messages and read messages for the range of 1 to 3 months in the past. The estimates use the same specification as Table 1. The regressions control for any short-run impacts that have already occurred. For instance, in estimating the 3-month impact (impact of spam 3 months ago), we control for the immediate change in behavior this had (the short-run effect) by including lagged dependent variables in the regression. What this means is we are estimating the direct impact. For example, if the 2-month effect is estimated to be zero, say, this does not mean the effect goes away, it only means that there is no *additional* effect as compared to the 1-month impact.

Engagement Estimates: Examining Figure 3 a few trends are immediately clear. The first is that the effect decays over time. For sent mail and composed mail, the negative impact occurs entirely in the first month following exposure. Recall that percentage-wise, these two metrics saw the largest short-run declines. Evidently part of the reason for this is that the total impact is felt in the first month following exposure. The graph also confirms the analysis of the previous section that the impact on sent mail occurs primarily through composed messages, not replies or forwards. For reading messages, the decline is less steep as there is still significant impact 3-months out. We thus conclude that while spam can have a

	(1) Page Views	(2) Sent Mail	(3) Reads
1{Male}=1	-0.0037	-0.00015	-0.0065***
	(0.0029)	(0.0003)	(0.0020)
1{New user}=1	-0.0107	-9.63e-06	-0.0014
	(0.0076)	(0.0006)	(0.0053)
1{Light}=1	0.0036	-0.0006**	0.0011
	(0.0029)	(0.0003)	(0.0020)
1{Heavy}=1	-0.0027	-0.0005	-0.0013
	(0.0038)	(0.0003)	(0.0027)
1{User <30}=1	-0.00194	0.00123***	0.0106***
	(0.0030)	(0.0003)	(0.0020)
1{User >50}=1	-0.0090*	0.0009*	-0.0043
	(0.0051)	(0.0005)	(0.0035)
1{High baseline	-0.0568	-0.0007	0.0605**
exposure}=1	(0.0410)	(0.00369)	(0.0286)
R-squared	0.162	0.089	0.010

Table 3: Differential impact of spam exposure by user characteristics. p<0.01: ***, p<0.05: **, p<0.1: *.

direct impact on behavior up to 3-months down the road, this is not the case for "volitional" categories in which the initial impact is large, such as sent/composed mail.

4.5 Breakdown by user characteristics

In this subsection we augment the regression specification used in Table 1 by interacting dummy variables for user characteristics with spam exposure. The interaction terms give the differential impact of spam based on the characteristic in question. The results are summarized in Table 3. All of the characteristics except gender and user age (self-reported age of the user) were used in matching. For the two measures that were not used in matching, the indicator variable only equals 1 if both users fall under the designation. For example, the variable 1{User <30} is defined as 1 if both users are under the age of 30. High baseline exposure is defined as being in the top 1/3 of spam exposure in the matching months. Light users are those that had page views in the bottom third during the matching months, heavy is top third. All other variables are self-explanatory.

Sent Mail and Page Views: We see that for sent mail and page views, user characteristics do not appear to predict the response to spam. However, the fact that heavy users do not show a higher *absolute* impact of spam exposure, means that percentage-wise, light users are the most adversely affected. Spam exposure is likely an important feature in retention, as it is known that decreased usage among light users is an important predictor of quitting.

Reading Messages: For reading messages, we find that the impact is significantly larger for males (more negative) and smaller for young (in calendar age) users. Users with higher baseline spam exposure respond slightly less to changes in spam exposure, however as we noted, this analysis is tenuous because we assume that spam classification accuracy is not a function of past exposure, when in reality it might be, due to user votes, for instance.

Takeaways: Overall we do not see major difference in the impact of spam based on user characteristics. The most notable result is that the percentage impact is highest for light users.

5 Related Work

There are two broad classes of existing works related to our research. On the methodology side, our work is related to the traditional causality methods literature. On the application side, our work is related to those quantifying the impact of spam. While we cannot cover every work here, we will mention some key works from each side in order to put our paper in context.

Estimating Causality: The study of causality has been an active area for many years. In particular, our work is developed within the framework of causal models developed by Rubin in early 1970s [36]. Our method of matching users by covariates or features is based on the theory developed in [36, 37]. The major steps that distinguish us from this work are the combined use of the matching and the regression to adapt this technique to the continuous setting, the use of criterion such as nearest neighbor matching, bi-directional matching, and locality sensitive hashing to speed up the computation. The propensity score matching method (PSM) uses the propensity score (predicted probability of exposure) to match users instead of actual covariates, and was first proposed in [35] and many follow-up works, nicely surveyed in [8], have proposed different refinements under the framework of the PSM. Besides PSM, other alternative ways to do such matching such as inverse propensity weighting [19, 20] and doubly robust estimation [18, 26] are also popular. As we mentioned earlier, all these works usually require that treatment and untreated/unexposed (control) groups be clearly identified. Thus, it is not directly applicable in our spam study as discussed earlier in Section 2.

Causal effects have been studied in many application scenarios, especially on the Web [9, 38]. For example, [9] applied several PSM to study the effect of online ads. To the best of our knowledge, there is no previous study on the causality effect of email spam on user engagement.

Impact of Email Spam: As discussed before, email spam has become a critical problem, being also related to

various online nefarious activities [28] such as phishing, scamming and spreading malware. Our paper is related to recent works that try to quantify the impact of spam from the economic side. For example, [22] conducted a study to quantify the conversion rate of the spam in order to understand how much spammers earned off bulk email distribution. The focus was thus the economics of the spam campaigns, rather than the user level metrics. [42] studied how much inconvenience of users is caused by the spam mails, by measuring the user's "willingness to pay" to remain unaffected by spam. [7] studied the cost of spam and the cost saved by spam filtering. The goal of all these papers is to quantify the cost from an organization's point of view, and their main metric is amount of working time spent in dealing with spam. Our aim was instead to measure the effect on the user engagement metrics from the economic perspective of the email service provider. Since the email service provider is the the key entity that invests in anti-spam technology, we feel this is a useful perspective to adopt.

Studying the impact of spam on users is part of a broader trend trying to characterize the economic incentives each of the stakeholders has in combating spam. Understanding the underground economy is the counterpart of what we are doing here. As mentioned before, researchers have concentrated on individual parts of this economy—the supply chain [22, 23], the labor market [31, 30] and malware distribution [6]. We consider our work as complementary to this thread, shedding light onto the ESP-centric part of the economic cycle.

6 Discussion and Summary

In this paper we described a large scale matching method, along with the corresponding regression method, in order to infer causal effects from observational data, specifically applicable in the case when the exposure variable is continuous. In situations where exposure is not a decision of the user but is correlated with engagement metrics, observational methods run into the correlation without causation problem. The gold standard to measure causality of course is a randomized experiment, but they are often too risky from a revenue or brand management perspective (the negative impact might outweigh the knowledge gains), unethical (involve exposing users to bad outcomes) or not ideal because the underlying behavior requires large changes in the independent variable of interest to measure a behavioral response. Mail spam runs afoul of all these requirements of A/B testing and is inherently interesting to study, given how pervasive it is in email-based communication.

We provide quantitative estimates that show that the impact of spam in the inbox can have serious revenue implications and can contribute to a large percentage drop in user engagement. The effect is largest for more voluntary user activities such as sending and especially composing emails. The function mapping spam changes to engagement appears to be convex, with the marginal impact increasing with the size of the exposure change. We carefully looked for contagion effects and found that while there are meaningful spillovers (reduced engagement across the Yahoo! site) the spillovers can be mechanically linked to decreased webmail activity so are thus not pure "brand-loss" effects, even though they are still relevant in evaluating the revenue impact. User characteristics are not particularly informative in predicting the response to spam; the most notable result is that light users are equally affected in absolute terms by a piece of spam in the inbox, meaning that percentage-wise the impact is far greater for these users.

Our result shows why it is important to quantitatively estimate a behavior even when the sign of the impact is "obvious." Merely documenting that mail spam has negative impact on engagement would not be particularly informative, but pinning the magnitude of the impact and the channels through which it operates can help the firm make investment decisions in filtration technology and optimize the user-interface to mitigate the effects. We believe the method can be fruitfully applied to other forms of abuse, such as abusive user-generated content, and other online experiences, such as pop-up ads.

References

[1] E. Allman. The economics of spam. *Queue*, 1(9):80, 2003.

[2] A. Andoni. Nearest neighbor search: the old, the new, and the impossible, 2009.

[3] P. Barford and V. Yegneswaran. An inside look at botnets. *Malware Detection*, pages 171–191, 2007.

[4] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paaß, and S. Strobel. New filtering approaches for phishing email. *J. Comput. Secur.*, 18(1):7–35, 2010.

[5] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and B. Zupan. Spam filtering using statistical data compression models. *Journal of Machine Learning Research*, 7:2673–2698, 2006.

[6] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: The commoditization of malware distribution. In *Proceedings of the 20th USENIX Security Symposium*, 2011.

[7] M. Caliendo, M. Clement, D. Papies, and S. Scheel-Kopeinig. The cost impact of spam filters: Measuring the effect of information system technologies in organizations. *IZA Discussion Paper No. 3755*, 2008.

[8] M. Caliendo and S. Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

[9] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *16th ACM SIGKDD*, pages 7–16. ACM, 2010.

[10] T. Claburn. Spam made up 94% of all e-mail in december. Technical report, Information Week, `http://www.informationweek.com/news/internet/showArticle.jhtml?articleID=197001430`, 2007.

[11] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, 1979.

[12] G. V. Cormack. Email spam filtering: A systematic review. In *Foundations and Trends in Information Retrieval*, 2008.

[13] A. Dasgupta, M. Gurevich, and K. Punera. Enhanced email spam filtering through combining similarity graphs. In *WSDM*, pages 785–794, 2011.

[14] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *CHI*, pages 581–590, New York, NY, USA, 2006. ACM.

[15] D. Fiebig and R. Bartels. The frisch-waugh theorem and generalized least squares. *Econometric Reviews*, 15(4):431–443, 1996.

[16] J. Goodman, G. V. Cormack, and D. Heckerman. Spam and the ongoing battle for the inbox. *Commun. ACM*, 50(2):24–33, 2007.

[17] C. Herley and D. Florêncio. A profitless endeavor: phishing as tragedy of the commons. In *Proceedings of the 2008 workshop on New security paradigms*, NSPW '08, pages 59–70, New York, NY, USA, 2008. ACM.

[18] K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 07 2003.

[19] D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[20] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting Sample Selection Bias by Unlabeled Data. In *Advances in Neural Information Processing Systems 19*, pages 601–608, 2007.

[21] C. Ivey. http://www.shoestringmillionaire.com/the-asymmetrical-economy-of-spam/, 2011.

[22] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: an empirical analysis of spam marketing conversion. In *Proceedings of the 15th ACM conference on Computer and communications security*, pages 3–14, 2008.

[23] C. Kanich, N. Weaver, D. McCoy, T. Halvorson, C. Kreibich, K. Levchenko, V. Paxson, G. Voelker, and S. Savage. Show me the money: characterizing spam-advertised revenue. In *Proceedings of the 20th USENIX Security Symposium*, pages 8–12, 2011.

[24] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *KDD*, 18(1):140–181, 2009.

[25] R. Lewis, J. Rao, and D. Reiley. Here, there, and everywhere: correlated online behaviors can lead to overestimates of the effects of advertising. In *WWW 2011*, pages 157–166. ACM, 2011.

[26] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[27] S. Malinin. Spammers earn millions and cause damages of billions. http://english.pravda.ru/russia/economics/15-09-2005/8908-spam-0/, 2005.

[28] T. Moore, R. Clayton, and R. Anderson. The economics of online crime. *Journal of Economic Perspectives*, 23(3):3–20, 2009.

[29] S. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2007.

[30] M. Motoyama, K. Levchenko, C. Kanich, D. McCoy, G. M. Voelker, and S. Savage. Re: Captchas – understanding captcha-solving from an economic context. In *Proceedings of the USENIX Security Symposium*, August 2010.

[31] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. Voelker. Dirty jobs: The role of freelance labor in web service abuse. In *Proceedings of the 20th USENIX Security Symposium*, 2011.

[32] Y. Namestnikov. The economics of botnets. *Analysis on Viruslist. com, Kapersky Lab*, 2009.

[33] E. Park. Update on global spam volume. http://www.symantec.com/connect/blogs/update-global-spam-volume.

[34] J. M. Rao and D. H. Reiley. The economics of spam. *Journal of Economic Perspectives*, Forthcoming Summer, 2012.

[35] P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41, 1983.

[36] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[37] D. B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal Of Educational Statistics*, 2(1):1–26, 1977.

[38] D. B. Rubin and R. P. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, 21(2):206–222, 2006.

[39] D. Sculley and G. M. Wachman. Relaxed online svms for spam filtering. In *SIGIR*, pages 415–422, New York, NY, USA, 2007. ACM.

[40] A. K. Seewald. An evaluation of naive bayes variants in content-based learning for spam filtering. *Intelligent Data Analysis*, 11(5):497–524, 2007.

[41] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns. In *Proceedings of the 4th USENIX conference on Large-scale exploits and emergent threats*, LEET'11, pages 4–4, Berkeley, CA, USA, 2011. USENIX Association.

[42] S.-H. Yoo, C.-O. Shin, and S.-J. Kwak. Inconvenience cost of spam mail: a contingent valuation study. *Applied Economics Letters*, 13(14):933–936, November 2006.

7 Appendix: Comparison to Propensity Score Matching via Simulations

There has also been much research into developing techniques, e.g., covariate matching, bias reduction, propensity score matching (PSM) [35, 20, 9], etc, which have shown promising results in removing this bias in observational studies. In this section, we outline the basic framework of propensity score matching and then discuss why the basic framework is unsuitable for us. We then compare our proposed method, nearest neighbor matching, with two variants of propensity score matching model based on a simulation data set with ground truth. Although our use of nearest neighbor matching method was prompted by concerns e.g. continuous exposure variable that make the naive PSM inapplicable, nevertheless we want to test whether there exist variants of PSM that are more adapted for our purposes. In order to do such a test, we needed to simulate the actual ground truth measure so that we can compare the effects unearthed by each method to the ground truth. In what follows, we first give an outline of PSM and then describe a variant we develop, stratified-PSM, that we compare with the nearest neighbor matching technique that we use. We then describe how we created the simulation dataset and compared the different algorithms.

7.1 Propensity Score Matching

In this section, we first briefly explain the PSM method of estimating effects before describing the modifications. In the classical PSM model, we have clearly defined treated and untreated (unexposed) groups—denote them by U_1 and U_0 respectively. The goal is to study the effect or outcome y on the treated users. For each user u, we use $y_u(s=1)$ or $y_u(s=0)$ to represent the effect on user u depending on whether the user is treated or remains untreated. Thus, we are interested in measuring the effect of treatment as $\Delta y = E[y_u(s=1) - y_u(s=0)|u \in U_1]$. However, a single user u can either be in the treated or the untreated group, but not both. A naive estimator of the above effect would thus be $\Delta y = E[y_u(s=1)|u \in U_1] - E[y_u(s=0)|u \in U_0]$ – this faces the problem of selection bias, since the populations in U_1 and U_0 are different, and have different properties which can be correlated with outcome y. The basic idea in PSM to overcome this bias is to select one or more users in the control group for each treated user, based on some pre-exposure features $\mathbf{x}_u$. Under the condition of unfoundedness,

$$Pr(y_u(s=0)|\mathbf{x}_u, u \in U_0) = Pr(y_u(s=0)|\mathbf{x}_u, u \in U_1),$$

we have the following estimator

$$\Delta y = E[y_u(s=1)|u \in U_1] - E_{\mathbf{z} \in U_1}[y_u(s=0)|u \in U_0, \mathbf{x}_u = \mathbf{z}],$$

where $\mathbf{z} \in U_1$ means $\mathbf{z}$ is a feature vector of a treated user. To avoid matching on the whole feature vector $\mathbf{x}_u$, we can match on the one-dimensional propensity score $p(\mathbf{x}_u)$ which is the probability that a user with vector $\mathbf{x}_u$ belongs to the treatment group. Then we have

$$\Delta y = E[y_u(s=1)|u \in U_1] - E_{v \in p(U_1)}[y_u(s=0)|u \in U_0, p(\mathbf{x}_u) = v],$$

where $v \in p(U_1)$ means that v is a propensity score of a treated user.

7.2 Unsuitability of PSM

As described above, the main aim in PSM is to try to learn a consistent estimator of $p(\mathbf{x})$, the probability the user has been exposed to a certain amount of spam, based on the all the feature we have constructed. In our case, we proceed differently due to a couple of reasons as pointed out – the basic underpinning of propensity score matching methods is being able to model the probability that a particular user falls into the treatment group. If the exposure variable is continuous, this assumption, and hence the modeling falls apart. We instead have to have a variant where we would have to create separate models for each value of the exposure. Secondly, the primary reason for propensity score matching is because matching users becomes difficult if the activity vector is high dimensional and the number of users is small – this is not the case for us: we have tens of features and we have over a million users; and we are able to find close matches. Lastly, being able to create a model that is a consistent estimator of $p(\mathbf{x})$ is very important, else we could be subject to un-intended biases that arise from this modeling.

In the presence of these issues, the commonly used ways of applying propensity score matching (PSM) does not apply to us. In the next subsection we describe a variant of PSM, where we stratify the dataset into multiple exposure levels and solve a PSM for each level.

PAIR	PSM1	PSM2	PSM2-W
1.579	3.376	4.578	5.878

Table 4: The L1 difference from the ground-truth. The smaller the value the better.

7.3 Variants of PSM for continuous exposure

In our problem, we care about the effect on engagement difference Δy if the spam fraction increases by Δs. To adapt PSM in our setting, we start out by first grouping users by discretizing their spam fraction values. Given a set of user U and their spam fraction range $[a,b]$, we have the following two ways of grouping users:

- **Equal-depth grouping.** In this method, we order all the users based on their spam fraction values increasingly. We then split the order list equally into m segments. In this method, each group has the same number of users.

- **Equal-width grouping.** In this method, we cut the spam fraction $[a,b]$ equally into m segments, each with a width of $(b-a)/m$. Users are grouped accordingly. In this method, each group can have different number of users.

Given a grouping method, for each pair of user segments, we use the segment with the lower spam fraction as the treated group and the one with the higher spam fraction as the control group – we compute Δs, the difference of the spam fraction between these two groups, as the difference of the average over the users in the two groups. We can then use a PSM model to compute the effect Δy. At the end, we will have a set of $(\Delta s, \Delta y)$ pairs.

To get the estimation function between the effect difference and spam fraction difference, we use the local regression method [11] to fit a curve on the set of $(\Delta s, \Delta y)$ pairs. We use PSM1 to denote Equal-depth grouping and PSM2 to denote Equal-width grouping. Please note that we have the same number of users for each $(\Delta s, \Delta y)$ in PSM1 but we have different numbers of users for PSM2. Thus for PSM2, we have a weighed version PSM2-W by weighing each point proportional the number of users in the treatment group before fitting the curve.

7.4 Simulation Results

To test the validity of our method by comparing it against ground truth, as well as to compare different variants of PSM with our method, we generate a simulation data with ground truth by the following procedure: we subsample 50K users from the mail-spam data that described in Section 4. For each user, we only kept 8 matching features – the mail pageviews, the incoming mail, incoming spam, the outgoing mails for two months. The spam-fraction in exposure month is the exposure variable, and the mail page-views in the post-exposure month is the effect variable. Because we want to generate the ground truth effect as close to the real effect as possible, we then learnt a gradient boosted decision tree model that tries to fit the effect variable in terms of the matching features and the exposure variable. This model that we learnt of user behavior was then used to create the new values of the effect variable for each user – as the user-set was sub-sampled, we strengthened the impact of exposure on the mail-pageviews by adding in another component to the model – this was a log-normally distributed random variable whose expectation depends on the logarithm of the difference of the spam exposure of this user from the mean spam exposure of all users: this changed each predicted effect value by around 10%. This aggregated model was then used to generate the new data, and also to create the ground truth curve for each value of Δs by predicting the new effect and then averaging over all user with the same matching features.

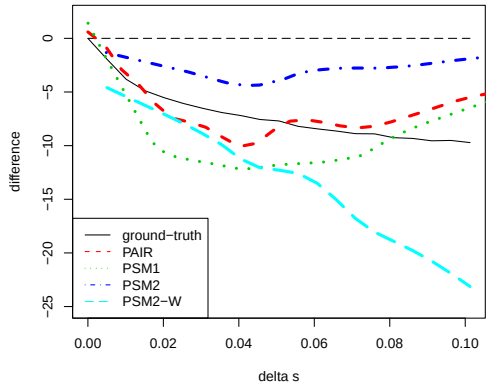

Figure 4: Comparison of our method PAIR and the variants of PSM methods.

We show the comparison results in Figure 4. For PSM methods, we set the number of user groups $m = 20$. (We tried different values for m and found the results are not very sensitive.) For our method, we obtain 1.17M pairs after our nearest neighbor matching and filtering steps. Each pair gives us a $(\Delta y, \Delta s)$ point and we use the same local regression method [11] to get a fitted curve. In Figure 4, we show the ground truth curve for $\Delta y(\Delta s)$, as well as the estimated curves for every method. Each of the estimates does capture the negative correlation between Δs and Δy. But, the estimates produced by the PSM methods are certainly worse than the one created by the nearest neighbor matching method. This is measured quantitatively by the L1 difference between the each estimated curve with the ground truth one – which we compute using 20 sampled points of

$\Delta s = \{0.005, 0.01, \cdots, 0.095, 0.1\}$. The L1 differences are shown in Table 4. One of the reasons of PSM performing worse is that when Δs becomes large, the resulting buckets have small number of users, and hence the variance is high. This simulation provides evidence that the matching method provides a reasonable set of estimates to ground truth, and that it performs better than some obvious variants of PSM, when dealing with continuous treatment values.