Real-Time Evaluation of Email Campaign Performance

July 2007

André Bonfrer*
Xavier Drèze

* Corresponding author. André Bonfrer is Assistant Professor of Marketing at the Singapore Management University, Lee Kong Chian School of Business, 50 Stamford Road #05-01, Singapore 178899, andrebonfrer@smu.edu.sg, Tel +65-6828-0398, Fax +65-6828-0777. Xavier Drèze is Assistant Professor of Marketing at the Wharton School of the University of Pennsylvania, 3730 Walnut Street, Philadelphia, PA 19104-6340, 215-898-1223, Fax 215-898-2534, xdreze@wharton.upenn.edu. This research was funded in part by the Wharton-SMU research center, Singapore Management University and in part by a WeBI - Mac Center Grant.
Real-Time Evaluation of Email Campaign Performance

We develop a testing methodology that can be used to predict the performance of email marketing campaigns in real time. We propose a split-hazard model that makes use of a time transformation (a concept we call virtual time) to allow for the estimation of straightforward parametric hazard functions and generate early predictions of an individual campaign’s performance (as measured by open and click propensities). We apply this pre-testing methodology to 25 email campaigns and find that the method is able to produce in an hour and fifteen minutes estimates that are more accurate and more reliable than what the traditional method (doubling time) can produce after 14 hours. Other benefits of our method are that we make testing independent of the time of day and we produce meaningful confidence intervals. Thus, our methodology can be used not only for testing purposes, but also for live monitoring. The testing procedure is coupled with a formal decision theoretic framework to generate a sequential testing procedure useful for the real time evaluation of campaigns.

Keywords: Database marketing, email, pre-testing, advertising campaigns.
Introduction

Email can be a powerful vehicle for marketing communications. Many marketers favor this new medium because it provides them with a cheaper and faster way to reach their customers. Further, the online environment allows marketers to measure consumers’ actions more accurately (Weible and Wallace 2001). This is a boon for marketing scientists in their desire to increase the effectiveness of marketing efforts and measure the ROI of marketing expenditures.

Although email response rates started out high (especially when compared with those reported for online and offline advertising), they declined over time and are now below 2.5% (DMA 2005). Finding ways to raise these response rates is critical for email marketers. A useful tool to achieve this is an effective email testing methodology. Identifying potential strengths and weaknesses of the content (the email creative) and the target population before the email is sent out at full scale can help marketers improve the response rates for their campaigns.

As a motivating example for the problem we are interested in, consider the case of a product manager at one of the major movie studios. With two or three new DVDs coming out every week, studios often rely on email marketing to generate interest for upcoming titles. This online promotion is particularly important for smaller titles (e.g., Transporter 2) that will not benefit from mass advertising and will not be heavily pushed by Amazon or Blockbuster. For such titles, the product manager would typically ask her creative agency to come up with a few concepts, and pick one of the designs to send out. She would also use a series of criteria (e.g., movie genre, gender) to select a subset of her database as a target for the email. Given the number of new titles to promote every week, she would work on a short production cycle and send emails without formally testing the quality of the design or the target sample (this is different from large releases such as King Kong which are planned months in advance and receive broad advertising and channel support). The success of our manager’s emails might be much improved if she could test multiple creatives and target selection in a fast (she has short lead times) and inexpensive (the titles do not warrant large expenses) way. There are no models in the extant marketing science literature that can be directly applied to provide such a test.

Many companies using email for marketing communications face such real-time constraints. For example, we monitored a random selection of 196 companies over a period of six months and found that about a third of these newsletters sent email to its customers once a
week or more (see Appendix B for details). These companies would clearly benefit from tools which allow them to test the performance within tight time constraints.

The importance of testing elements of the marketing mix is not new to marketing scientists. For example, in new product development (and distribution) the ASSESSOR model (Silk and Urban 1978; Urban and Katz 1983) has been used for decades to forecast the success of new products based on laboratory based test marketing. Methods have also been developed to perform multiple parallel testing as predicated by the new product development literature (Dahan and Mendelson 2001; Smith and Reinertsen 1995). A novel approach used by Moe and Fader (2002) utilizes advance purchase orders made via the CDNOW website to generate early indicators of new product sales for music CDs. In advertising, the efficacy of an advertising campaign is assessed using a battery of tests designed to identify the best creative to use (e.g., the most persuasive or the most memorable) using selected members of the target audience. Field experiments with split-cable television technology have also been used to study the impact of advertising on brand sales (Lodish et al 1995; Blair 1988).

In direct marketing, modeling techniques have been developed to help marketers select the right customer group for a given content. Bult and Wansbeek (1995) build a regression model to predict each customer’s likelihood of responding to a direct marketing communication, and then select which customers should be contacted by explicitly maximizing the expected profits generated by each communication. Using a dynamic programming approach rather than using regression, Bitran and Mondschein (1996) take the profit maximization objective a step further by incorporating inventory policies (inventory and out of stock costs) into the decision. Gönül and Shi (1998) extend this dynamic programming approach by allowing customers to optimize their own purchases behavior over multiple periods (i.e., both the firm and the customer are forward looking). Gönül, Kim, and Shi (2000) recognize that customers can order from old catalogs and that one can still garner new sales from old customers who were not sent a new catalog. Thus, they propose a hazard function model of purchases where customers are sent a catalog only if the expected profits with the additional mailing exceeds the profits without the mailing. Elsner, Krafft, and Huchzermeier (2004) use Dynamic Multilevel Modeling (DMLM) to optimize customer segmentation and communication frequency simultaneously. Gönül and Ter Hofstede (2006) find the optimal mailing policy for a catalog marketer given individual level predictions of customers’ order incidence and volume.
In practice, few direct marketing campaigns are rolled-out untested. The traditional approach used to predict the ultimate response rate to a direct marketing offer is the *doubling method* or half life analysis (Nash 2000, Rosenwald 2006, Hughes 2006). In the doubling method, the marketer uses historical data to compute the time it takes for half of the responses to be received (the doubling time). Then, when performing a new test, he waits the doubling time, and multiplies by two the number of responses received at that time to estimate the ultimate response rate. The waiting period depends on the medium used. For first class mailing, firms will wait a minimum of two weeks; for third class mailing, they will wait four weeks. Once the tests results are in the decision makers make a go/no-go decision based on profitability. Morwitz and Schmittlein (1998) show that when making such projections, managers typically do not sufficiently regress test response rates to the mean.

Testing is also popular in other Internet marketing applications. Online advertisers track banner ad performance in real time to identify the appeal (click-through) of various advertising creatives. Click-stream models can be implemented to test the appeal of content by measuring the click-through rates or website stickiness (Bucklin and Sismeiro 2003). Eye tracking technology may be used to identify where (and if) a customer is viewing the advertising message embedded on a webpage (Drèze and Hussherr 2003).

While some of these testing methodologies might be adapted to the context of email marketing, some unique features of email present several new modeling challenges. First, many firms have implemented tracking technologies for email campaigns that monitors whether and when a customer responds to an email. Given the goal of real-time testing, it is essential that we make full use of this continuous time data.

Second, in contrast to a typical clickstream setting, email communications are initiated by the firm rather than the customer. This adds a layer of complexity in that, while the delivery of an email is often close to being instantaneous, there is an observable delay between the time the email is sent out and the time it is opened. The opening of the message will depend on how often customers check their email. Thus, although marketers have direct control over when emails are sent out, they have little control over whether and when a customer responds to the email. This is different from traditional clickstream models where the user requests the content, and we can assume that the content is being processed immediately.
A third difference in email marketing involves the lead time for generating both the creative and the execution of the campaign. While email delivery is not free nor instantaneous, even a large email campaign can be sent out at relatively low cost and delivered in a matter of hours, consequently campaigns are short lived and often run with short lead times and consequently with compressed deadlines on a weekly (e.g., American Airlines, Travelocity, The Tire Rack), bi-weekly (e.g., Longs Drugstore), or even daily basis (e.g., Sun Microsystems). These short lead times place significant constraints on testing.

For these reasons, effective email marketing communication requires a testing methodology that is able to generate actionable results in as short a time as possible. The goal of such a testing procedure is to generate predictions of open incidence and click-through rates of any email campaign as quickly and accurately as possible. Our paper describes the development and test of such an email pre-testing model. We begin by developing a split-hazard model of open behavior. The split-hazard component models the incidence of open (versus not open). We use a Log-Logistic hazard function to predict the distribution of open times. Click behavior is then modeled using both a censored split hazard model and a simpler Binomial model. To help produce stable estimates even when data is sparse (a common occurrence when trying to test campaigns in a short amount of time), we use Bayesian shrinkage estimation with correlated priors between open and clicks. Shrinkage leverages information contained in past campaigns and weights this past information with response data observed in the focal campaign. The performance of the model in predicting response rates for email campaigns is compared with the doubling method used by direct marketing practitioners.

In our application of the models to actual data, we find it necessary to account for intraday variations in customer responses (e.g., to account for fewer emails opened at night). Consequently, we develop a concept of virtual time that allows us to produce a model that fits the data well while keeping the specification simple. Virtual time involves adjusting the speed of time through the day to adapt to the marketers’ and customers’ availability. For instance, compare an email sent at noon to an email sent at two in the morning. While both emails might ultimately be opened, it is likely that the delay between send and open will be smaller for the email sent in the middle of the day than for the email sent in the middle of the night simply because the average recipient is more likely to be awake and checking his email during the day than during the night. To reflect that recipients’ availability fluctuates during the day, we speed
up and slow down the speed at which time elapses through the day. An hour of real time at noon (when customers are available) could be expanded to two hour of virtual time to reflect that a lot can happen during this interval while an hour at two in the morning (when most customers are asleep) could be compressed to half an hour to reflect that little would happen in the middle of the night. Using virtual time allows us to keep the model specification simple. This makes shrinkage straightforward and allows for an easy interpretation of the model parameters.

The testing procedure is applied to data obtained from a large entertainment company. The analysis shows that, using our approach, we can reduce testing time from 14 hours to less than ninety minutes without any decrease in testing reliability. It also highlights the pitfalls inherent with compressing testing time. Indeed, the more compressed the testing time, the more sensitive the quality of the results are to the specifications of the hazard function and to intra-day seasonality.

Our model offers a number of substantial practical benefits. (1) The fast evaluation of a campaign allows for early warnings about the probable success or failure of the campaign. This can lead to timely go/no-go decisions for either campaigns or creative. (2) The model provides diagnostic information that can be integrated in a formal decision process to improve the results of an under-performing campaign or discard any campaign that does not perform above some threshold level of response. (3) The testing procedure coupled with such a decision process can generate higher average (cross-campaign) response rates. (4) Only a small sample is required for testing. The small sample size makes it easy to test the effectiveness of multiple advertising copies or to test the reaction of different target groups of customers. (5) Our process incorporates historical data and thus will lead to better decisions as more campaigns are tested. Indeed, as the number of campaigns grows, the email marketer learns more about the distribution of response rates for campaigns sent. This leads to more accurate forecasts.

1. Research Setting and Data Description

We calibrate and test our models using a database of twenty-five email campaigns sent as part of the online newsletter of an entertainment company. Most of the emails are in the form of promotions aimed at inducing customers to purchase a movie title online or offline, or to click on links to access further content (different campaigns have different purposes). Each campaign has a subject line displaying the purpose of the promotion (e.g., “Spiderman III now out on DVD!”). The main body of the email is only visible after the recipient has opened the email. Within the
body of the email, recipients are able to click on various links to learn more about the promotion or go to a movie specific website. It is important to note that clicks can only occur if a recipient opens the email.

Summary statistics for the email campaigns are reported in Table 1. The campaigns vary in size from around 5,000 to 85,000 emails sent. Our database consists of 617,037 emails sent, of which 111,419 were opened, and 9,663 of those emails were clicked on at least once. Across campaigns, the open rate (number of opens/ number sent) is thus 18.1% and the click-through rate (number of clicks/number of opens) is 8.7%. The unconditional click rate (number of clicks/number sent) is about 1.6%.

There is a wide variation in both open and click-through rates across campaigns (see Figure 1 for a scatter-plot of campaigns’ open and click rates). Clearly, the more successful campaigns are the ones that have both high open rates and high click-through rates (e.g., campaign A in Figure 1). Some campaigns (e.g., B) yield low open rates, but high click rates. This is probably an indication that the execution is good, but either they are targeted at the wrong audience, or the subject line used in the email failed to attract interest. The firm might improve the campaign by testing different target groups and or different subject lines. Other campaigns (e.g., C) enjoy high open rates but have low click-through rates. This is probably an indication that these campaigns have broad appeal but are poorly executed. The firm might be able to move these campaigns to the upper right quadrant by using better creative. Of course, there are also campaigns (e.g., D) that under perform on both opens and clicks; improving these campaigns would require improving both the targeting and the content. If this cannot be done, it may be best to drop the campaign altogether.

Given the variance in response rate shown in Figure 1, it is clear that it is difficult to predict the success of a campaign ex ante. The goal of our model is thus to predict out-of-sample open and click rates quickly and with small samples. Providing a forecast in a timely manner allows the firm to adjust the creative or targeting of the campaign when needed, or make go/no-go decisions.

Figures 2a and 2b present histograms of the time (in hours) for customers to open the email since it was sent, and the time (in minutes) it takes the customer to click on the email after it is opened. Given our objective of reducing the time allocated to testing, several features of our data are highly pertinent to model construction:
1. Opens usually occur within 24 hours of sending; clicks occur within a minute of being opened.

2. It takes a few hours for response rate to build up. There is a relatively low level of email activity immediately after a campaign is launched, followed by a build-up peaking around two hours.

3. The histogram of the delay between send and open (Figure 2a) reveals a distinct multi-modal pattern underlying particularly during the first 24 hours after an email is sent. This pattern is also visible on individual campaign histograms.

The first feature requires that a rapid testing model of open and click rate work well with censored data. Indeed, by shortening the testing time, we reduce the amount of uncensored data available to us. The second feature suggests that we must be careful about our assumptions regarding the data generation process when building our model. This is particularly important in our case as we are trying to make predictions about the entire distribution of email activity based on only a few hours of activity.

The multimodal pattern found in the time until opening is troublesome as it does not appear to conform to any standard distribution and might be difficult to capture with a simple model. To understand what may be driving this multi-modal pattern, we plot the distribution of opens throughout the day (see Figure 3a). This graph shows considerable variation in activity through the day. There are fewer emails opened late at night and early in the morning than during the day. We refer to this pattern as intraday seasonality. We show in the next section how this seasonality may be the cause of the multimodal feature of Figure 2a.

Given these results, we develop a model that has the following features: To generate estimates within the first few hours after sending, the model works with censored data and only a small amount of data is needed for estimation. The model also takes into account intraday seasonality to allow a parsimonious parametric approach to model the number of email messages opened.

2. Model Setup

We develop a rapid testing methodology for a specific application: the testing of online email campaigns. Rapid testing provides early feedback on whether a campaign is likely to be successful or not. In the spirit of traditional testing models, it is important that our methodology consumes as few resources as possible. Ideally, the model would also be parsimonious (i.e.,
have few parameters). It would estimate quickly such that a test can be implemented in real time
and would allow for the monitoring of an email campaign as it is being sent. Indeed, an overly
complex or over-parameterized model that takes hours to generate predictions would defeat the
purpose of rapid testing.

We first describe in more detail how the model accommodates intra-day seasonality. Next, we develop a split hazard model of open and click probabilities that takes into account the possibility that some emails are never opened or clicked on (given open). We then derive the shrinkage estimators for the open and click models and state the likelihood function used in the estimation.

2.1. From physical time to virtual time
A traditional approach to handling seasonality, such as that displayed in Figures 2a and 3a, is to introduce time-varying covariates in the model. There are two main problems with this approach. First, the covariates are often ad-hoc (e.g., hourly dummies). Second, they often make the other parameters less interpretable (e.g., a low open rate during peak hour could be larger than a high open rate during off-peak hours). To alleviate these concerns, we build on the approach proposed by Radas and Shugan (1998). Radas and Shugan (hereafter RS) de-seasonalized a sales process by changing the speed at which time flows. They showed that by speeding up time during high seasons, and slowing down time during low seasons, one can create a new (virtual) time series that is devoid of seasonality. The benefits of this approach, assuming that one has the right seasonality pattern, is that one can use straightforward models in virtual time and easily interpret the meaning of the parameters of these models.

The effectiveness of the RS approach hinges on having a good handle on the seasonal pattern present in the data. In their application (the movie industry) they produce seasonal adjustments by combining past sales data with industry knowledge (e.g., presence of major holidays with high movie demand). A shortcoming of this approach is that some of the seasonality may be endogenous to the firms’ decisions. For instance, if movie studios believe that Thanksgiving weekend is a ‘big’ weekend, they may choose to release their movies during that weekend rather than during off-peak week-ends (Ainslie, Drèze, and Zufryden 2005). Thus, part of the seasonality observed during Thanksgiving will be due to the fact that more consumers have the time and desire to see movies on that weekend (consumer induced seasonality) and part of the seasonality will be due to the fact that more movies are available (firm induced
seasonality). If one uses past data as a base for seasonal adjustment without considering the
decisions of the firm, one can potentially overcorrect and attribute all the seasonal effects to
consumer demand while it is in fact also partly due to firm supply.

In our case, we are confronted by both consumer- and firm-induced seasonality. For
instance, the average consumer is much less likely to open emails at four in the morning than at
four in the afternoon. Similarly, firms do not work 24 hours a day. If we look at when the firm
sends its email (Figure 3b), we observe little (but some) activity during the night, then a peak at
eight in the morning, a peak at noon, and a lot of activity in the afternoon. It is likely that these
peaks are responsible for some of the increase in activity we see in Figure 3a at similar times.

To separate consumer induced seasonality from firm induced seasonality, we benefit
from two features of our modeling environment not present in RS. First, we have continuous
time individual level data. While RS had to work with aggregate weekly measures, we know the
exact time each email is sent and opened. Second, while a movie can open on the same day
throughout the country, emails cannot all be sent at the same time. Emails are sent sequentially;
for example, a million-email campaign can take up to 20 hours to send. Thus, we can simulate an
environment that is devoid of firm based seasonality by re-sampling our data such that the
number of emails sent at any point in time is constant through the day (i.e., Figure 3b for such a
firm would be flat).

To resample the data, we proceed in three steps. First, for each minute of the day, we
collect all emails that were sent during that minute. Second, we randomly select with
replacement 100 emails from each minute of the day (144,000 draws). Third, we order the open
times of these 144,000 emails from 0:00:00 to 23:59:59 and associate with each actual open time
a virtual time equal to its rank divided by 144,000. The relationship between real and virtual time
based on their cumulative density functions is shown in Figure 4. This represents the passing of
time as seen by consumers independent of the actions of the firm.

We can use the relationship depicted in Figure 4 to compute the elapsed virtual time
between any two events. For instance, if an email were sent at midnight and opened at two in the
morning, we would compute the elapsed virtual time between send an open  by taking the
difference between the virtual equivalent of two a.m. (i.e., 00:29:44 virtual) and midnight (i.e.,
00:00:00 virtual) to come up with 29 minutes and 44 seconds. Similarly, if the email had been
sent at noon and opened at two p.m., then the elapsed virtual time would be 11:05:10 – 09:08:30 = 1 hour 56 minutes and 40 seconds.

Applying the virtual time transformation to the elapsed time between send time and open time for all emails in our dataset has a visible impact on the frequency distribution. The underlying seasonal pattern and the multimodality that were apparent in Figure 2a all but disappears.

Our approach does not account for another form of endogeneity that could arise if the firm were to strategically send good and bad email campaigns at different times of the day such that the two types of campaigns would not overlap. Fortunately, this was not company policy in our application—email campaigns were sent out whenever they were ready.

2.2. A split-hazard model of open and click time

The time it takes for customers to open an email from the time it is sent, or the time it takes to click on an email from the time the customer opens it are both modeled using a standard duration model (e.g., Moe and Fader 2002, Jain and Vilcassim 1991). Since both actions can be modeled using a similar specification, we discuss them inter-changeably. Starting with opens, we account for the fact that in an accelerated test, a failure to open an email is indicative of one of two things. Either recipients are not interested in the email, or they have not had a chance to see it yet (i.e., the data is censored). Of course, the shorter the amount of time allocated to a test, the higher the likelihood that a non-response is indicative of censoring rather than lack of interest. Thus, we model the open probability and the open time simultaneously in a right-censored split hazard model (similar to Kamakura, Kossar, and Wedel 2004, and Sinha and Chandrashekaran 1992).

The probability that a customer will open or click an email varies from campaign to campaign, and is denoted with $\delta^e_k$, where $e$ is a superscript identifying different campaigns, and the subscript $k$ denoting an open ($k=o$) or click ($k=c$). The likelihood function is constructed as follows. We start with a basic censored hazard rate model of the open or click time distribution:

\[
L^e_k = \prod_{i=1}^{N^e} f(t^e_{ik} | \Theta^e_k)^{R^e_i} S(T^e | \Theta^e_k)^{1-R^e_i},
\]

where:

- $e$ is a subscript that identifies a specific email campaign,
- $k$ is a superscript that identifies the model used ($k \in \{o = \text{open}, c = \text{click}\}$),
- $i$ is an index of recipients,
\[ N^e \] is the number of recipients for email \( e \),
\[ R_{ik}^e \] is 1 if recipient \( i \) opened/clicked email \( e \) before the censoring point \( T^e \),
\[ T^e \] is the censoring point of campaign \( e \),
\[ t_{ik}^e \] is the elapsed time between send and open (open and click) in the event that
the recipients opened (clicked) the email,
\[ f(t \mid \Theta) \] is the pdf for time \( t \), given a set of parameters \( \Theta \),
\[ S(t \mid \Theta) \] is the corresponding survival function at time \( t \),
\[ \Theta \] is a set of parameters for the pdf and survival functions.

The likelihood function in (1) needs to be adjusted to account for the fact that some recipients
will never open or click the email. Let \( \delta_k^e \) denote the probability that email \( e \) will be opened or
clicked then the likelihood component for action \( k \) at censor time \( T^e \) is:

\[
L^e_k(t_{ik}^e, T^e, R_{ik}^e \mid \Theta_k^e) = \prod_{i=1}^{N^e} \left[ \delta_k^e f(t_{ik}^e \mid \Theta_k^e) \right]^{R_{ik}^e} \left[ \delta_k^e S(T^e \mid \Theta_k^e) + (1 - \delta_k^e) \right]^{(1-R_{ik}^e)}
\]

(2)

The estimation of \( \delta_k^e \) and \( \Theta_k^e \) for any parametric hazard function can be performed by
maximizing this general likelihood function.

We tested a variety of candidate hazard rate specifications (including the Exponential,
Weibull, Log-Normal and the more general Box-Cox specification) to model the open and click
processes (see Appendix A). Overall, the tests revealed that the Log-Logistic distribution is best
suited for our application. The probability density function and the survivor function for the Log-
Logistic are (for details about the Log-Logistic and other distributions mentioned in this paper,
see see Kalbfleisch and Prentice, 1985):

\[
f(t \mid \alpha, \lambda) = \frac{\lambda \alpha (\lambda t)^{\alpha-1}}{(1 + (\lambda t)^\alpha)^2},
\]

\[
S(t \mid \alpha, \lambda) = \frac{1}{1 + (\lambda t)^\alpha}
\]

(3)

where \( \lambda > 0 \) is a location parameter and \( \alpha > 0 \) is a shape parameter. Consistent with previous
notation, we refer to the shape and location parameters for any given campaign \( e \) and email
response action \((k \in \{o, c\})\) as \(\alpha_k^e\), and \(\lambda_k^e\), respectively. Depending on the value of \(\alpha\), the Log-Logistic hazard is either monotonically decreasing \((\alpha \leq 1)\) or inverted U-shape \((\alpha > 1)\) with a
turning point at \(t = \frac{(\alpha - 1)^{1/\alpha}}{\lambda}\).

2.3. Shrinkage estimators

As in most practical applications, we benefit from having data available from past campaigns. This information can be used to improve the performance of our model. Specifically, we use the parameters estimated from past campaigns to build a prior on the open and click hazard functions, as well as the split hazard component. This is especially useful at the beginning of a campaign when data is sparse.

Since we use the Log-Logistic hazard function in our application, each of the split-hazard models has three parameters \((\alpha, \lambda, \delta)\). When building priors for these parameters, it is reasonable to assume that the open and click rates for a campaign are correlated. Indeed, broad appeal campaigns should yield both high open and high click rates while unappealing campaigns would exhibit both low open and low click rates.

To accommodate the possibility of correlated open and click rates, we use a bivariate Beta as prior distribution for the \((\delta_1, \delta_2)\). As in Danaher and Hardie (2005) and Schweidel, Bradlow and Fader (2007), we implement the bivariate distribution proposed by Lee (1996):

\[
(\delta_1, \delta_2) \sim g(x_1, x_2) = f(x_1 | a_1, b_1) f(x_2 | a_2, b_2) \times [1 + \omega (x_1 - \mu_1)(x_2 - \mu_2)]
\]

where the \(f(x_1 | a_1, b_1)\) and \(f(x_2 | a_2, b_2)\) functions represent univariate Beta densities for \(x_1\) and \(x_2\) given Beta parameters \((a_i, b_i)\) and \(\mu_i\) is the mean of the univariate Beta for \(x_i\) (i.e.,

\[
\mu_i = a_i / (a_i + b_i), \quad \omega \quad \text{is a function of the correlation between} \quad x_1 \quad \text{and} \quad x_2
\]

\((\omega = \text{Corr}(x_1, x_2) / (\sigma_1, \sigma_2)), \quad \text{and where} \quad \sigma_i^2 = a_i b_i / [(a_i + b_i)^2 (a_i + b_i + 1)] \) is the variance of the univariate Beta. The value for \(\omega\) can be thought of as the unnormalized correlation between \(f(x_1 | a_1, b_1)\) and \(f(x_2 | a_2, b_2)\).

There is no reason to believe that the hazard rate parameters for clicks and opens will be correlated as one process (opens) depends on the availability of recipients to open emails while the other process (clicks) is conditional on the recipients having opened the email (and thus be available) and depends on the amount of processing needed to understand the email and react to
it. It is possible, however, that the shape and location parameters of each hazard rate are correlated. Thus, we use a bivariate Log-Normal distribution as prior on \( \alpha_k \) and \( \lambda_k \):

\[
(a_k, \lambda_k) \sim \text{Log-Normal}(\mu_k, \Sigma_k)
\]  

(4)

For a given campaign, we use the method of moments to estimate the parameters \((a_o, b_o, a_c, b_c, \text{Corr(Open,Click)}, \mu_o, \Sigma_o, \mu_c, \Sigma_c)\) based on the parameters \((\alpha_o, \lambda_o, \delta_o, \alpha_c, \lambda_c, \delta_c)\) obtained from all other campaigns. The correlation, \( \rho^{LN}_{\alpha,\lambda} \) between parameters \( \alpha_k \) and \( \lambda_k \) of the Log-Normal distribution is adjusted for possible small sample bias using the correction factor described in Johnson and Kotz (1972, page 20):

\[
\rho^{LN}_{\alpha,\lambda} = \frac{\exp(\rho^{N}_{\alpha,\lambda} \sigma_{\alpha} \sigma_{\lambda}) - 1}{\sqrt{\exp(\sigma^2_{\alpha}) - 1} \{\exp(\sigma^2_{\lambda}) - 1\}}
\]  

(5)

Note that this correlation coefficient is independent of the means of the parameters, depending only on the standard deviations of the parameters (respectively, \( \sigma_{\alpha}, \sigma_{\lambda} \)).

As an indicator of the importance of accounting for the correlation between the various parameter, in our application the estimated correlation between the open and click rate across all campaigns is 0.433 and the correlation between \( \alpha \) and \( \lambda \) is -0.372 and 0.870 for the open and click hazard rate respectively.

The combination of the likelihood functions given in (2) with the Log-Logistic hazard specification in (3) and the bivariate Beta and Log-Normal priors, the following likelihood function is to be estimated for a given campaign (omitting the \( e \) superscript for clarity):

\[
L(t_o, R_o, t_c, R_c | \Theta, T) = \text{LN}(\lambda_o, \alpha_o | \mu_o, \Sigma_o) \text{LN}(\lambda_c, \alpha_c | \mu_c, \Sigma_c) \times
\]

\[
\text{Beta}(\delta_o | a_o, b_o) \text{Beta}(\delta_c | a_c, b_c) \left[ 1 + w \left( \delta_o - \frac{a_o}{a_o + b_o} \right) \left( \delta_c - \frac{a_c}{a_c + b_c} \right) \right] \times
\]

\[
\prod_{i=1}^{N} \left[ \delta_o \frac{\lambda_o \alpha_o \left( \lambda_o t_{io} \right)^{\alpha_o - 1}}{(1 + \left( \lambda_o t_{io} \right)^{\alpha_o})^2} \left[ 1 - \delta_o \left( 1 - \frac{1}{1 + \left( \lambda_o T \right)^{\alpha_o}} \right) \right] \right]^{\text{R}_o} \times
\]

\[
\prod_{i=1}^{N} \left[ \delta_c \frac{\lambda_c \alpha_c \left( \lambda_c (t_{ic} - t_{iw}) \right)^{\alpha_c - 1}}{(1 + \left( \lambda_c (t_{ic} - t_{iw}) \right)^{\alpha_c})^2} \left[ 1 - \delta_c \left( 1 - \frac{1}{1 + \left( \lambda_c (T - t_{iw}) \right)^{\alpha_c}} \right) \right] \right]^{\text{R}_c} \]

(6)
2.4. An Alternative Approach to Estimating Click Rates

Although theoretically sound, using a split-hazard model to estimate the parameters of the click times (conditional on an open) might be overly complex. Indeed, since most consumers click on an email within seconds of opening it, it is likely that few click observations are right-censored. If the vast majority of clicks are uncensored, we can test a simpler model for the estimation of click rates that does not take censoring of clicks into account. Our hope is that this more parsimonious model will perform better at the beginning of a test, when few data points are available.

Formally, we assume that clicks follow a binomial process, and use the same bivariate Beta prior for \((\delta_o, \delta_c)\) as we did in the full model. The likelihood function for this simplified model is:

\[
L_g \left( t_o, R_o, t_c, R_c \mid \Theta, T \right) = \text{LN}(\lambda_o, \alpha_o, \mu_o, \Sigma_o) \text{LN}(\lambda_c, \alpha_c, \mu_c, \Sigma_c) \times \\
\text{Beta}(\delta_o \mid a_o, b_o) \text{Beta}(\delta_c \mid a_c, b_c) \left[ 1 + w \left( \delta_o - \frac{a_o}{a_o + b_o} \right) \left( \delta_c - \frac{a_c}{a_c + b_c} \right) \right] \times \\
\prod_{i=1}^{N} \left[ \frac{\lambda_o \sigma_o \left( \lambda_o T \right) a_o^{-1}}{1 + \left( \lambda_o T \right)^{a_o^{-1}} \left( 1 - \delta_o \left( 1 - \frac{1}{1 + \left( \lambda_o T \right)^{a_o^{-1}}} \right) \right)} \right]^R_o \left[ \delta_o^{-1} \left( 1 - \delta_o \left( 1 - \frac{1}{1 + \left( \lambda_o T \right)^{a_o^{-1}}} \right) \right) \right]^{1-R_o} \left[ \delta_c^{-1} \right]^{(1-R_o)^{-R_o}}
\]

2.5. Comparisons with Benchmarks: The Doubling Method

Before discussing an application of our model, we would like to draw a comparison with existing approaches to predicting the success rate of direct marketing campaigns. The most common model used by practitioners is the doubling method (Nash 2000, it is sometimes referred to as the half-life analysis as in Hughes 2006). This method involves first looking at the responses of past direct marketing campaigns and computing the amount of time it takes for 50% of the responses to be received (the doubling point). The analyst then uses the heuristic that for any future campaigns, the predicted total number of responses is equal to double the number of responses observed at the doubling point. In our case, the doubling point is at 14 hours (see Table 1).

The doubling method is a powerful and simple heuristic. It makes three implicit assumptions. First, it assumes that not everybody will respond. Second, it assumes that it takes time for people to respond. Third, it assumes that the timing of the responses is independent from the rate of response and constant across campaigns. As a non-parametric method, it does not make any assumption about the underlying response process, nor does it provide any ways to
test whether the current campaign conforms to the data collected from previous campaigns or runs faster or slower than expected. Hence, it does not provide any ways to evaluate whether a current test should be run for a longer period or could be finished early; an important piece of information our model provides.

In essence, the doubling method aggregates time into two bins; each containing half of the responses. This aggregation loses vital timing information that could be used to better model the response process. If one wanted to speed-up the doubling method, any other quantile could be used to perform a test. For example, the 25th percentile would give a “quadrupling” method, the 33rd percentile would give a “trebling” method and so on.

Another benchmark examined in this paper is to wait for a predetermined time period and count the number of opens and clicks that are observed at that time. The campaign would be deemed a success if these numbers exceed a certain threshold. While this benchmark does not predict open and click rates, it can be used as a decision rule. We discuss this and other decision rules in more detail in section 3.2.

3. Application of the model to email campaign pre-testing

We now apply our models to the data from the email campaigns described earlier and evaluate the relative predictive validity of various models. We compare the predictions of our models with those of the doubling method. We then use our model in a campaign selection simulation and benchmark it against other possible heuristics.

3.1. Simulation and validation

The main purpose of the simulation and validation stages is to validate the models proposed in the paper by studying the accuracy of the predictions they make out-of-sample. We also want to find out which of the models has the best predictive performance and whether we can generate estimates that are useful for decision making within a short amount of time (say hours) such that testing is feasible for campaign planning. When doing so, we compare the models based on real time (no time transformation) and virtual time, for both the full model and the simplified click formulation. In summary, we fit and validate the following four models:

1) Full model in real time (using Equation 6 and real time)
2) Full model in virtual time (using Equation 6 and transformed time)
3) Simplified model in real time (using Equation 7 and real time)
4) Simplified model in virtual time (using Equation 7 and transformed time)
In each simulation, we adopt the perspective of a marketer who wishes to pretest his campaigns before committing to the final send. To this end, we look at each campaign assuming that the remaining \((E - I)\) campaigns have already been completed. Prior to the test, we know nothing about a focal campaign except the number of emails that need to be sent out. However, we can use all the information collected through the other campaigns to build our priors.

We set the sample test size at 2,000 emails (the sample size for the test was varied between 1,000 and 2,000 in increments of 200 emails but did yield any substantive difference in findings). We simulated different test lengths, in 30 minute increments, ranging from 30 minutes to 8 hours. For each test length, any email that had been opened prior to the simulated end of test was used in the non-censored component of the log-likelihood. All other observations are coded as censored (e.g., if an open were recorded 45 minutes after an email was sent, this observation would be coded as censored when we simulate a 30 minute test, and as an open for a 60 minute test and any subsequent ones). Based on this set of censored and uncensored observations, the parameters of the full and the simplified models are estimated using priors based on all other campaigns.

As a practical matter, the distributions of open and click times exhibit long tails, such that some responses continue to come in long after a campaign has run its course. Historical data reveals that 99% of all email responses are observed within 3 weeks of sending out the email communication. Typically, the company conducts post-campaign debriefing 2-3 weeks after the emails are sent out. Thus, the cut-off date is set at 3 weeks (504 hours) and the number of opens and clicks observed at that time is used as the true value the models need to predict. Our forecast of the number of opens and clicks at 3 weeks is constructed using the parameter estimates for each of the censored samples. The cumulative distribution of opens and clicks at 504 hours is calculated using:

1) \(\widehat{\text{Open}}_{504} = \delta_o^e \times F(504 \text{ hours} | \alpha_o^e, \lambda_o^e) \times (\text{Number Sent for campaign } e)\), and

2) \(\widehat{\text{Click}}_{504} = \delta_c^e \times F(504 \text{ hours} | \alpha_c^e, \lambda_c^e) \times (\text{Estimated opens for campaign } e)\).

**Predictive Performance of the Models**

The full results for each campaign and for each censoring point consists of a set of parameters \((\delta_o^e, \alpha_o^e, \lambda_o^e)\) for opens and \((\delta_c^e, \alpha_c^e, \lambda_c^e)\) for clicks. Since for each model we have 16 test times (between 30 minutes and 8 hours) and 25 campaigns, this means our analysis generates a total of
4x16x30 = 1,920 estimates. Given this large number of estimates it is difficult to present all the results in one table. It is also not that meaningful to present any sufficient statistic of these estimates since they are censored at a different point in time.

To assess the performance of the models in the pre-campaign tests, we calculate the Mean Absolute Deviation (MAD) for the predicted number of clicks and opens at each censoring point. The MAD expresses the absolute prediction error in terms of the number of clicks and number of opens, averaged across campaigns and is calculated by comparing the predicted opens and clicks with the true opens and clicks. This is done for each of the four models estimated, for each of the test lengths (from 30 minutes to 8 hours in half hour increment) and for each of the campaigns in the simulation.

The doubling time MAD for opens is 917. For all the open models tested, the MAD declines as testing time increases. Based on the MAD prediction error, the virtual time models outperform the corresponding real time models. The virtual time models achieve a prediction error that is equal or better that of the doubling method in as little as three hours (as opposed to 14 hours for the doubling method). This represents a reduction in test time of close to 80%. For real time, it takes slightly over four hours before the tested model outperforms the doubling time model.

Although, in the interest of space, we do not present the standard errors for the various campaigns’ parameter estimates, we find that using virtual time also produces tighter and more stable estimates (i.e., smaller confidence intervals) than the model using real time data. An additional advantage of the virtual time models is that processing time is much reduced due to improved convergence of the maximum likelihood procedure. Thus the virtual time model is well suited to test the performance of a campaign in compressed time applications.

The prediction errors for clicks exhibit a similar pattern. The simplified model slightly outperforms the full model in the early stages of testing (up to about 5 hours for the real time model, and 1.5 hours for the virtual time model). The simplified model in virtual time achieves a predictive performance similar to the doubling method in only two and a half hours—a 82% reduction in testing time.

3.2. A Campaign Selection Decision Rule
The results reported in the preceding section demonstrate that our models can produce predictions of the open and click rates faster and more accurately than the doubling method.
While the results do show a clear improvement in speed and accuracy, a natural question that arises is: are the benefits of the new models substantial from a managerial standpoint? To investigate this in the context of our application, we develop a formal decision rule based on the testing methodology developed in section 2. We take the perspective of a campaign manager who wishes to improve the ROI of his marketing actions by eliminating under-performing campaigns. One possible reason for not wanting to send a low yield campaign is that it costs more to send than its expected returns. Such underperforming campaigns also represent an opportunity cost in that they tie up resources that could be used to send more profitable campaigns. Furthermore, sending undesirable material could lead to higher customer attrition (Hanson, 2000).

The average click-through rate (CTR) across our 25 campaigns is 2% (weighting each campaign equally rather than by their size). Hence, let us imagine that our manager wishes to eliminate any campaigns with a CTR lower than 2%. Clearly, rejecting any campaigns with response rates lower than the historical average will have the consequence of improving response rates across future campaigns. If the manager wishes to further improve response rates, it is also possible to set this threshold higher.

To make a go/no-go decision on campaigns, the manager could consider one of the following five decision rules:

**Rule 1: Doubling Method**
Under the doubling method, the decision rule is quite simple. For a given campaign, send out the test sample, wait 14 hours and observed the CTR. If the CTR greater than 1% (1/2 of 2%) then run the full campaign, otherwise cancel the campaign.

**Rule 2: The Horse Race**
As mentioned in section 2.5, if the manager needs only to make a go/no-go decision rather than predict the actual outcome of the campaign, all he has to do is to compare the results of the current campaign to the other campaigns at some point during the testing phase. If these results are better than average, go ahead with the full campaign, else, cancel it. In essence, the manager is entering the campaign in a horse race, pitching the current campaign against previous ones.

This approach is simple and parameter free. It has the advantage of being faster than the doubling method and is simpler than our proposed model. It also can give useful diagnostics on

---

1 We thank an anonymous reviewer for proposing this alternative methodology.
whether the campaign failed because of the content (via clicks) or because of poor execution (via opens). If all one cares about is picking the right campaign, it may be sufficient to use this approach. We implemented this approach by using the same two and a half hour cutting point as we use in rule 3.

**Rule 3: The Proposed Model**
Under decision rule 3, the manager will run the test for two and a half hours. He will then fit our model and use its output to predict the long run CTR. If the predicted result is larger than 2% then he proceeds with the campaign, otherwise he cancels it.

**Rule 4: Sequential Statistical Test under Proposed Model**
Rule 3 makes a go/no-go decision based on the CTR predicted by our model. In doing so, it ignores the uncertainty around this estimate. A more astute manager would consider this uncertainty when making his go/no-go decision. He can do so in multiple ways. One approach is to continue the test until there is enough data to conduct a statistical test that would reject the null hypothesis that the CTR is equal to 2% (we call this rule 4). Another approach would be to incorporate a loss function into the decision (we call this rule 5).

Under rule 4, the decision process adopted by the manager is as follows:

1. Send 2,000 test emails.
2. Wait $M$ minutes.
3. Fit our proposed model.
4. Predict the end of campaign CTR.
5. Use standard error around the prediction to test the hypothesis that CTR is greater than 2%.
6. If the test fails to reject the null hypothesis, go back to step 2.
7. If the test shows that CTR is greater than 2% then proceed with the campaign, otherwise cancel it.

This method has the advantage of making the most of the data available to the analyst. It considers not only the point estimate CTR, but also the uncertainty around the estimate. Finally, it adapts to each campaign by cutting short the test if the results are clearly significant, or running the test longer if more precision is needed.

In our test, we set the significance level for the statistical test at $p = .05$ (two-tailed) and we set $M$ at 30 minutes to limit the number of models we need to fit (25 campaigns in half hour
Rule 5: Asymmetrical Loss Function

A second way to incorporate the uncertainty around the point estimate into the manager’s decision is to consider the fact that in many cases, under-estimation errors have a different impact on a business’ operations than over-estimation errors. In the case of email marketing, over- and under-estimation errors could be viewed from two perspectives. First, consider the low variable costs of sending email. In the short term, it is much more costly to not send a good campaign (loss opportunity) than to mistakenly send a bad one (cost of sending the campaign). Thus, under-estimation errors should be penalized more heavily than over-estimation errors and a manager should err on the side of sending too many campaigns rather than too few.

A second view is that a legitimate email firm operates on a permission basis. It cannot send emails to an individual who unsubscribe from the mailing list. Thus, a campaign manager might be concerned that under-performing emails may lead the recipients to re-evaluate the benefits they receive from the firm (maybe associating the emails with SPAM) and potentially defect. Such defection is costly as it deprives the firm from future revenues. In such case, the argument could be made that the relatively low cost of sending an email is more than matched by the high cost of defection and thus campaign managers should penalize over-estimation (leading to sending poor performing campaigns) more than under-estimation (leading to not sending well performing campaigns).

Given these two perspectives, the differential treatment of over- and under-estimation errors in the decision process can be handled using an asymmetrical loss function. Varian (1975) proposes the use of the LINEX loss function in such a case. This function is approximately linear on one side of zero (error) and approximately linear on the other side, depending on the value of a parameter $a$. The sign and magnitude of $a$ represent the direction and degree of symmetry (when $a > 0$, overestimation is more serious than underestimation, and vice-versa.) When $a$ is close to 0, the LINEX loss is almost symmetric and approximately equal to the squared error loss. Although other asymmetric loss functions exist, the LINEX loss function is one of the more commonly used in statistical applications (Calabria and Pulcini 1998; Pandey 1997; Zellner 1986).
If we call $\Delta$ the difference between the estimated and the true value of the parameter of interest then the LINEX loss function is expressed as:
\[
    l(\Delta) \propto \exp(a\Delta) - a\Delta - 1; \quad a \neq 0.
\]
As shown in Soliman (2002), the corresponding Bayesian estimator that minimizes the posterior loss is:
\[
    u^* = -\frac{1}{a} \log \left( E_u [\exp(-a.u)] \right).
\]
In our context, $u$ is the predicted CTR. To compute $E_u [\exp(-a.u)]$ we approximate the integral over $u$ of $f(\bar{u}) \exp(-a.\bar{u})$. To accomplish this, we make 10,000 draws from the posterior distribution of the estimated parameters (i.e., $\hat{\delta}_o, \hat{\delta}_c, \hat{\alpha}_o, \hat{\lambda}_o$) taking correlations among parameters into consideration. The CTR ($\bar{u}$) resulting from such parameter draws can then be computed, which can then be used to compute the mean of $\exp(-a.\bar{u})$. We adopt the perspective of a manager who is afraid of missing a good campaign so that $a < 0$; for exposition purposes, we used $a = -10$. Other values could be used depending on the risk aversion of the managers.

**Comparison of Decision Rules**

The results of decision rules 1 to 5 are reported in Table 2. Seven out of the 25 campaigns have a CTR greater than 2%. The average CTR for these seven campaigns is 4.45%. We use the doubling method (Rule 1) as the method to beat. In our case, the doubling method selects 11 campaigns, the seven correct campaigns plus four underperforming campaigns, yielding a CTR of 3.34%.

The horse race (Rule 2) does not perform as well as the doubling method and leads to a decrease in response rate of 20% (2.66% vs. 3.34% CTR); it is not very discriminate and retains far too many campaigns. As expected, our model (Rule 3) performs better than the doubling method. Indeed, the average campaign response rate can be increased by 28% (4.29% CTR vs. 3.34%) by using our model and waiting for 2.5 hours of virtual time rather than using the doubling method and having to wait for 14 hours. One can gain a further 15% (4.77% vs. 3.63%, for a total improvement of 43%) by using the sequential testing rule (Rule 4) rather than using a fixed time stopping rule. Rule 4 actually over-performs by selecting only six out of the seven correct campaigns and not having any false positives.

It should be noted that, when using a statistical test to decide when to stop testing, the testing time varies widely across campaigns. On average, the adaptive testing rule requires 2
hours 57 minutes. The fastest campaign reaches statistical significance in as little as an hour. The slowest campaign needs six hours.

Using the LINEX loss function (Rule 5), with the value of $a=-10$ (i.e., being afraid of missing a good campaign) leads to one more campaign being selected (as compared to the results of Rule 3). This is to be expected as the fear of missing a good campaign would lead to a more lenient test. The net result is a slight decrease in response rate as the campaign turns out to be an underperformer.

In short, our tests demonstrate that our method performs better than a simple comparison of campaign performance after a fixed amount of time (be it the doubling method rule or the horse race). Further, by taking advantage of the fact that our method produces not only a point estimate, but a distribution around the point estimate, one can further improve decision making either through a sequential testing method or the use of a loss function that reflects manager beliefs.

3.3. When Is the Best Time To Test?

Our comparison of the predictive ability for the split hazard rate model suggests that, on average, we can learn as much in two and a half hours as we can learn from the doubling method in 14 hours. However, it is important to remember that these two and a half hours are measured in virtual time. In real time, the test will require more or less time depending on the time of day that is conducted. Figure 5 shows how long two and half virtual hours correspond to in real time, depending on when the test starts. There appears to be a “sweet spot” in the afternoon, between 1pm and 7pm where a two and half virtual hour test can be carried out in less than two actual hours (the shortest it could take would be one hour and 15 minutes by starting at 5:34pm). Starting after 7pm will impose delays as the test is unlikely to be finished before people go to bed; if the test is started at 10:12pm it will take almost 6 and a half hours to complete.

Given the speed at which the tests can be carried, one can imagine that the firm could adopt a sequential refinement approach where it first tests a series of concepts at 9am. The results would come in by noon. Between noon and 1pm, it refines the winning concept by producing different variation of it. At 1pm, it tests the refined concepts, with the results coming in by 3pm. It can then take the winning concepts of this second phase and produce a final set of concepts that are tested at 4pm with the results out by 6pm. At that time it can select the final creative and send it out. This series of concept creation/testing/selection can be carried out within
a day and should produce much better results than sending out untested creatives. A key advantage of our methodology is the ability to develop multiple results in quick succession, a feature that is crucial for a sequential testing procedure.

3.4. Accounting for different types of campaigns

Through the use of informative priors, our methodology builds on the information collected across all past campaigns. The more similar these campaigns are to the focal campaign, the better the methodology will perform. To the extent that additional (i.e., movie level) campaign data are available, it may be possible to improve the results by using different priors and different virtual time for different campaigns. Of course, when only a subset of campaigns is used to construct the prior, a trade-off must be made. On the one hand, more informative priors can be generated from using only campaigns that are similar to the focal campaign. On the other hand, there will be fewer campaigns to draw upon and thus the prior might be more diffuse.

To investigate the possibility of improving the performance of the model by building priors out of more similar campaigns, the email campaigns were first divided into three groups based on the genre of the movie they are promoting: Action Movies (13 campaigns), Romance Movies (eight campaigns), and Others (four campaigns - most of the “Others” category are special interest emails promoting more than one title each). We then reran the whole analysis for each three sub-groups of campaigns using the simplified model. That is, we re-ran the simulated test for each campaign using only the similar campaigns as prior, and using virtual time based on only the campaigns in the same sub-group. Figure 6a shows that the speeds of time for each of the three subgroups are very similar but show some discrepancies. For instance, the ‘Others’ group exhibits slower virtual time until about 11 a.m., then speeds up and overtakes the other two groups until midnight. The ‘Romance’ group shows the opposite pattern.

The average parameter estimates for the different campaign genres are reported in Table 3. These results reveal that the response parameters differ significantly across movie genres. The ‘Romance’ movie campaigns produce much lower open rates than ‘Action’ and ‘Others’ campaigns (15.93% vs. 20.38% and 20.98% respectively). Romance campaigns also have lower click rates on average (8.13%), leading to the lowest overall click-through rate (1.34%). In contrast, ‘Other’ campaigns have much higher click rates (15.91%), leading to the highest overall click-through rates (3.80%).
Table 3 also shows the average shape and location parameters for the Log-Logistic hazard rate of each group. To help with the interpretation of these parameters, we show the resulting hazard rates in Figure 6b. This graph shows that the ‘Others’ category has a much flatter hazard rate. It peaks about one hour and 45 minutes after the email has been sent and after two days is only reduced by half. The ‘Action’ genre, in contrast, peaks in about 40 minutes and loses about 80% in two days. At that point, the hazard rate is actually lower than for ‘Others.’

This shows that the ‘Others’ category is much more appealing to consumers and has more staying power. This might be due to the fact that they promote more than one movie and thus may have a wider appeal. ‘Romance’ campaigns, in comparison, are much less appealing. They garner lower open rates, lower click rates and have a much steeper hazard function.

These differences across types of campaigns are quite large and suggest that using the priors based on similar campaigns rather than the general priors would lead to faster and more accurate tests. However, in our simulated tests the sub-group models do not perform as well on average as the pooled model. This result can be explained by the reduced ability for shrinkage to work given the smaller number of campaigns used to build the prior. In the ‘Others’ category for instance, there are only 4 email campaigns. This means that when testing a campaign, only three other campaigns are used as priors. This reveals itself to be too few campaigns to produce helpful priors. The sub-group tests suggest that it is better to use priors based on the full set of 24 other campaigns even if the majority of these campaigns are relative to somewhat different types of offer. It is likely that the reason for this is that, even though these other campaigns promote different types of movies, they still promote movies and are still aimed at the same basic type of consumers and thus contain relevant information.

The lack of positive results in this sub-group analysis does not invalidate the premise. We do find significant differences in behavior across the three types of campaigns. Unfortunately, there are not enough campaigns for each campaign genre to make the sub-group analysis work. We have no doubt that, as the firm collects more data on subsequent campaigns, the sub-group analysis would be worthwhile. It is easy to see how it would be optimal for the firm to start by using overall priors and then switch to more targeted priors as it builds a larger portfolio of past campaigns.
4. Discussion and Conclusion

The value of information increases with its timeliness. Knowing quickly whether a campaign is going to be successful provides the opportunity to correct potential problems before it is too late or even stop the campaign before it is completed. It is therefore imperative to develop methods that improve both the accuracy and the speed with which campaign testing can be done. In this article, we study a modeling procedure that can be implemented for the fast evaluation of email campaign performance.

The performance of an email campaign is defined by its open and click rates. The methodology we propose predicts these rates quickly based on estimates produced with small samples of the main campaign. Reducing the sample size and testing period to a minimum produces multiple modeling challenges. Indeed, we propose to send 2,000 emails, and wait less than two hours to produce estimates of how the campaign will perform after three weeks. In two hours, fewer than a hundred opens and fewer than ten clicks are typically observed. The key to successful prediction of the ultimate results of an email campaign based on so few data points lies in using the information to its fullest potential.

There are three elements that make our methodology successful: (1) using the appropriate model specification, (2) transforming time to handle intra-day seasonality, and (3) using informative priors. Each of these three elements provides its own unique contribution to the overall fit and predictive performance of the model.

The appropriate hazard function is critical because our compressed-time tests produce observations that are heavily right censored. Thus, we are often fitting a whole distribution based only on its first quartile (or even less). A misspecification of the hazard function could cause severe errors in prediction. In other words, the value of the responses of the first few people to respond to the email campaign is an important indicator of the success of the overall campaign. We find that the best fitting parametric model for open times is the Log-Logistic which can accommodate a non-monotonic hazard rate for the time to open and click.

For modeling click-through rates, we compare the same Log-Logistic hazard rate model with a Binomial process. We find that the straight binomial process is a good descriptor of the phenomenon given that consumers respond quickly after opening an email. Thus, we find that the click-through rate (the total number of clicks for a campaign, unconditional on open) is best
predicted using a combination of a Binomial model for the clicks, and a Log-Logistic split hazard model for the opens.

We apply our split-hazard model to a virtual time environment. The virtual time transformation removes intra-day seasonality and makes our testing procedure invariant to the time of day at which it is performed. This is a key factor in the robustness of our model in that it allows us to bypass the need to handle seasonality directly in the model and allows for a straightforward specification with few parameters. By limiting the numbers of parameters we must estimate to three for the open model and one for the click model, we make the best use of our limited data and we produce parameters that are directly interpretable (the click and open rates or estimate directly without the need for transformation).

Another benefit of our time transformation is that by making each campaign independent of the time of day, it is possible to compare results across campaigns, and to easily build informative priors for each of the parameters. This yields a procedure that produces meaningful estimates and confidence intervals with a minimum amount of data. It also allows a firm to conduct tests serially. That is, the firm could choose to modify either a campaign’s creative or target population as the result of a test, then retest the campaign and compare the new results to the first ones.

Overall, by putting these three elements together, the model is capable of running a test in one hour and 15 minutes that produces similar results to a traditional test in 14 hours (a 91% decrease in testing time). In addition, our methodology produces meaningful confidence intervals. The implication of this is that the firm can monitor the accuracy of its test and decide to run the test for a longer or shorter period of time depending on how well it performs. The model can be estimated within a matter of seconds, and could therefore be used in real time (e.g. either every minute or even whenever a new observation is received). Thus, our methodology can be used not only for testing purposes, but also for live monitoring. This finding is particularly important when one considers that a million-email campaign can take up to 20 hours to send. Using our model one could monitor the campaign as it is being sent, and implement a “stopping rule.” The rule would allow a manager to make a decision to terminate a campaign that is underperforming, or even to change the creative for the remaining recipients. If done right, this could significantly improve average response rates by limiting the detrimental impact of poor performing campaigns.
Our analysis of sub-group priors (i.e., building priors based on a sub-group of like campaigns rather than all other campaigns) failed to improve the testing methodology. Nevertheless, it showed promise in that there are large differences in customer responses across groups. For instance ‘Romance’ type movies yield a much lower response rate than more generic emails that promote multiple movies (1.34% CTR vs. 3.80%). Unfortunately, we do not have a large enough portfolio of campaigns to build informative priors for each sub-group of campaigns. Consequently, we observe a decrease in predictive performance for the sub-group analysis.

Our model represents a first step towards better online marketing testing. As such, there several avenues we see for further research in this area. For instance, due to the lack of individual level information in our dataset, we could not include covariates in our models. It is likely that adding such information, when available, would improve fit and predictive power of the model. Further, if the dataset contained many data points per recipient (we have an average of two emails sent per name) it would be possible to incorporate unobserved heterogeneity.

Another issue is the link between click and purchase behavior. The assumption is that click behavior is a measure of interest and is highly correlated to purchase. As many campaign managers are evaluated based on clicks, we feel our analysis is appropriate. However, in future applications and with better data, it should be possible to link click and purchase behavior, and thus optimize purchases rather than clicks.

In conclusion, our model has demonstrated strong performance advantages over extant methods used in practice. The methodology leverages features unique to email as a form of direct marketing. These features include compressed time intervals available for testing, near real-time response measurement and individual recipient tracking. Beyond email marketing, emerging technologies in marketing which also share these characteristics (e.g. mobile phone communications) could benefit from the testing methodology developed in this paper.
References


Figure 1: Open and click rates for the campaigns in our dataset. The black vertical and horizontal lines represent the median response rates.
Figure 2a and 2b. Histograms of elapsed time between sent and open (2a) and between open and click (2b) events, across all campaigns.
Figure 3a and 3b: Distribution of email opens (3a) and sends (3b) through the day (Pacific Standard Time).
Figure 4: Within day cumulative density functions of real and virtual time. Real time is distributed uniformly throughout the day resulting in a 45 degree line.
Figure 5: Test length as a function of time of day. This graph plots translates the 2.5 hour virtual waiting time into real time based on the time of day a test is performed.
Figure 6a and 6b: a) Speed of Time for Three Different Types of Email Campaigns

b) Email campaign open hazard rates by genre
Table 1: Campaign Summary Statistics. All campaign statistics are reported based on real time.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Emails</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sent/Campaigns</td>
<td>24,681</td>
<td>5,171</td>
<td>84,465</td>
<td>20,393</td>
</tr>
<tr>
<td>Opens/Campaigns</td>
<td>4,457</td>
<td>600</td>
<td>13,116</td>
<td>3,877</td>
</tr>
<tr>
<td>Clicked/Campaigns</td>
<td>387</td>
<td>53</td>
<td>1,444</td>
<td>360</td>
</tr>
<tr>
<td>Open rate</td>
<td>0.181</td>
<td>0.081</td>
<td>0.351</td>
<td>0.073</td>
</tr>
<tr>
<td>Click-through</td>
<td>open</td>
<td>0.096</td>
<td>0.035</td>
<td>0.303</td>
</tr>
<tr>
<td>Click-through rate</td>
<td>0.016</td>
<td>0.005</td>
<td>0.105</td>
<td>0.021</td>
</tr>
<tr>
<td>Doubling Time (hours)</td>
<td>13.76</td>
<td>4</td>
<td>29</td>
<td>5.57</td>
</tr>
<tr>
<td>First Open (minutes)</td>
<td>6.41</td>
<td>0.07</td>
<td>14.68</td>
<td>2.08</td>
</tr>
<tr>
<td>First Click (seconds)</td>
<td>2.08</td>
<td>1.0</td>
<td>6.0</td>
<td>1.656</td>
</tr>
<tr>
<td>Number of campaigns</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Results from the Decision Rule Simulation

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Rule 1</th>
<th>Rule 2</th>
<th>Rule 3</th>
<th>Rule 4</th>
<th>Rule 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Doubling Method</td>
<td>Horse Race</td>
<td>Proposed model</td>
<td>Adaptive model</td>
<td>LINEX a=-10</td>
</tr>
<tr>
<td>Number of Campaigns Selected</td>
<td>7</td>
<td>11</td>
<td>14</td>
<td>7</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>True Positives</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>False Positive</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>False Negative</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Average Testing Time (Hours)</td>
<td>14:00</td>
<td>2:30</td>
<td>2:30</td>
<td>2:30</td>
<td>2:30</td>
<td>2:30</td>
</tr>
<tr>
<td>Minimum Testing Time</td>
<td>14:00</td>
<td>2:30</td>
<td>2:30</td>
<td>1:00</td>
<td>2:30</td>
<td>2:30</td>
</tr>
<tr>
<td>Maximum Testing Time</td>
<td>14:00</td>
<td>2:30</td>
<td>2:30</td>
<td>6:00</td>
<td>2:30</td>
<td>2:30</td>
</tr>
<tr>
<td>Click-through Rate</td>
<td>4.45%</td>
<td>3.34%</td>
<td>2.66%</td>
<td>4.29%</td>
<td>4.77%</td>
<td>3.37%</td>
</tr>
<tr>
<td>Improvement over No Rule</td>
<td>123%</td>
<td>67%</td>
<td>33%</td>
<td>114%</td>
<td>139%</td>
<td>69%</td>
</tr>
<tr>
<td>Improvement over DM</td>
<td>33%</td>
<td>0%</td>
<td>-20%</td>
<td>28%</td>
<td>43%</td>
<td>1%</td>
</tr>
</tbody>
</table>

Table 3: Average Parameter Estimates for the Three Different Types of Campaigns

<table>
<thead>
<tr>
<th>Genre:</th>
<th>N</th>
<th>δ₀</th>
<th>α₀</th>
<th>λ₀</th>
<th>δᶜ</th>
<th>CTR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>13</td>
<td>20.38%</td>
<td>1.06</td>
<td>0.0018</td>
<td>8.58%</td>
<td>1.88%</td>
</tr>
<tr>
<td>Romance</td>
<td>8</td>
<td>15.93%</td>
<td>1.15</td>
<td>0.0020</td>
<td>8.13%</td>
<td>1.34%</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>20.98%</td>
<td>1.17</td>
<td>0.0012</td>
<td>15.91%</td>
<td>3.80%</td>
</tr>
<tr>
<td>All campaigns</td>
<td>25</td>
<td>19.1%</td>
<td>1.07</td>
<td>0.0019</td>
<td>0.0847</td>
<td>1.62%</td>
</tr>
</tbody>
</table>
Appendix A: Selecting the hazard function

Researchers in marketing have used various specifications for parametric hazard function when doing survival analysis (see Jain and Villcassim 1991, Sawhney and Eliashberg 1996; Chintagunta and Haldar 1998; Dréze and Zufryden 1998). Following their work, we considered the following four specifications: Exponential, Weibull, Log-Normal, and Log-Logistic.

We estimated a campaign level hazard rate model for each distribution using the complete set of opens and clicks available for each campaign (i.e., this is a straight hazard model that is neither split nor censored). We report the fit statistics for all four specifications in Table A1 (open model) and Table A2 (click model). The analysis suggests that the Log-Logistic distribution fits the data best overall for both open and click. The Log-Normal is a close second, but has the drawback of not having a closed form expression for its survivor function. It is important to note that the Exponential distribution performs relatively poorly, emphasizing the need for a non-constant hazard rate that allows for a delay between reception and open of an email, or between open and click (i.e., allows for enough time for consumers to process the message). The relatively poor fit of the Weibull distribution (which allows for a ramping up period) further shows that one also needs to accommodate for a decrease in the hazard rate after enough time has passed. Making the right assumptions regarding the change in hazard rate over time is thus crucial. This is especially true since much of the data available during the test will come from the first few hours of the test, representing the increasing part of the Log-Logistic hazard function. Estimating this based on a Weibull or Exponential hazard function would clearly misspecify the model.

Another approach to identifying the proper hazard function specification is to fit a Box-Cox model. The specification of the Box-Cox hazard function is such that it nests many of the traditional hazard rate specification (Jain and Villcassim 1991). Thus, one can fit a Box-Cox model and, assuming the true data generating process is one of the models nested in the Box-Cox specification, one can identify the proper hazard function by looking at the estimate parameters.

We fit the four parameter version of the Box-Cox described in (Jain and Villcassim 1991) to our campaigns. The parameter estimates failed to identify any specific distribution as possible candidate (the model rejected the Exponential, Weibull, Gompertz, and Erlang-2). Unfortunately, the Log-Logistic is not one of the distributions that are nested within the Box-Cox specification (Chintagunta and Prasad 1998). Hence, the Box-Cox approach cannot validate our choice of the Log-Logistic as base hazard specification; it can only validate our rejection of the other function.

We could have based our model on the Box-Cox hazard rate rather then the Log-Logistic. There are 3 reasons why we did not. First, the Box-Cox is a proportional hazard model. Hence, one cannot directly compare the likelihood function to the Log-Logistic likelihood for model selection purposes (in fact the Box-Cox produced positive log-likelihood for many campaigns). Second, the Box-Cox has no closed formed solution for its density and survivor functions. This means that these two functions must be numerically integrated during the estimation. This considerably slows down estimation. The base hazard model (without censoring or split hazard) took up to 7 hours to estimate as opposed to less then a minute for the Log-Logistic. This defeats the whole purpose of developing a fast testing methodology. Third, the specification developed by Jain and Villcassim (1991) is a four parameter function. We prefer our more parsimonious two parameter model as it puts less burden on the data.
Table A1: Fit statistics for the virtual open time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Exponential</th>
<th>Weibull</th>
<th>Log-Normal</th>
<th>Log-Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>BIC</td>
<td>AIC</td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>-3150</td>
<td>6303</td>
<td>6302</td>
<td>-2702</td>
</tr>
<tr>
<td>2</td>
<td>-5016</td>
<td>10035</td>
<td>10034</td>
<td>-4142</td>
</tr>
<tr>
<td>3</td>
<td>-3699</td>
<td>7401</td>
<td>7400</td>
<td>-3100</td>
</tr>
<tr>
<td>4</td>
<td>-32450</td>
<td>64904</td>
<td>64902</td>
<td>-26424</td>
</tr>
<tr>
<td>5</td>
<td>-2869</td>
<td>5741</td>
<td>5740</td>
<td>-2457</td>
</tr>
<tr>
<td>6</td>
<td>-7794</td>
<td>15591</td>
<td>15590</td>
<td>-6418</td>
</tr>
<tr>
<td>7</td>
<td>-13780</td>
<td>27564</td>
<td>27562</td>
<td>-12145</td>
</tr>
<tr>
<td>8</td>
<td>-21308</td>
<td>42620</td>
<td>42618</td>
<td>-18126</td>
</tr>
<tr>
<td>9</td>
<td>-4089</td>
<td>8181</td>
<td>8180</td>
<td>-3328</td>
</tr>
<tr>
<td>10</td>
<td>-18481</td>
<td>36966</td>
<td>36964</td>
<td>-15523</td>
</tr>
<tr>
<td>11</td>
<td>-4056</td>
<td>8115</td>
<td>8114</td>
<td>-3576</td>
</tr>
<tr>
<td>12</td>
<td>-11185</td>
<td>22374</td>
<td>22372</td>
<td>-9793</td>
</tr>
<tr>
<td>13</td>
<td>-4938</td>
<td>9879</td>
<td>9878</td>
<td>-3998</td>
</tr>
<tr>
<td>14</td>
<td>-8402</td>
<td>16808</td>
<td>16806</td>
<td>-7126</td>
</tr>
<tr>
<td>15</td>
<td>-5842</td>
<td>11687</td>
<td>11686</td>
<td>-5223</td>
</tr>
<tr>
<td>16</td>
<td>-29143</td>
<td>58290</td>
<td>58288</td>
<td>-25437</td>
</tr>
<tr>
<td>17</td>
<td>-4269</td>
<td>8541</td>
<td>8540</td>
<td>-3701</td>
</tr>
<tr>
<td>18</td>
<td>-1516</td>
<td>3035</td>
<td>3034</td>
<td>-1244</td>
</tr>
<tr>
<td>19</td>
<td>-6747</td>
<td>13497</td>
<td>13496</td>
<td>-5651</td>
</tr>
<tr>
<td>20</td>
<td>-26580</td>
<td>53164</td>
<td>53162</td>
<td>-23524</td>
</tr>
<tr>
<td>21</td>
<td>-5038</td>
<td>10079</td>
<td>10078</td>
<td>-4061</td>
</tr>
<tr>
<td>22</td>
<td>-5689</td>
<td>11381</td>
<td>11380</td>
<td>-4997</td>
</tr>
<tr>
<td>23</td>
<td>-17916</td>
<td>35836</td>
<td>35834</td>
<td>-16204</td>
</tr>
<tr>
<td>24</td>
<td>-8086</td>
<td>16176</td>
<td>16174</td>
<td>-6735</td>
</tr>
<tr>
<td>25</td>
<td>-7140</td>
<td>14283</td>
<td>14282</td>
<td>-5711</td>
</tr>
</tbody>
</table>
Table A2: Fit statistics for the virtual click time model. For the different specifications, the columns report the Log-Likelihood value (LL), the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). Shaded cells show the lowest AIC and BIC for a specific campaign.

<table>
<thead>
<tr>
<th>Campaign</th>
<th>Exponential</th>
<th>Weibull</th>
<th>Log-Normal</th>
<th>Log-Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>BIC</td>
<td>AIC</td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>-117</td>
<td>236</td>
<td>236</td>
<td>-115</td>
</tr>
<tr>
<td>2</td>
<td>-217</td>
<td>436</td>
<td>436</td>
<td>-207</td>
</tr>
<tr>
<td>3</td>
<td>-216</td>
<td>434</td>
<td>434</td>
<td>-215</td>
</tr>
<tr>
<td>4</td>
<td>-2567</td>
<td>5137</td>
<td>5136</td>
<td>-2506</td>
</tr>
<tr>
<td>5</td>
<td>-144</td>
<td>290</td>
<td>290</td>
<td>-142</td>
</tr>
<tr>
<td>6</td>
<td>-711</td>
<td>1425</td>
<td>1424</td>
<td>-697</td>
</tr>
<tr>
<td>7</td>
<td>-938</td>
<td>1879</td>
<td>1878</td>
<td>-908</td>
</tr>
<tr>
<td>8</td>
<td>-1972</td>
<td>3947</td>
<td>3946</td>
<td>-1921</td>
</tr>
<tr>
<td>9</td>
<td>-337</td>
<td>676</td>
<td>676</td>
<td>-330</td>
</tr>
<tr>
<td>10</td>
<td>-546</td>
<td>1095</td>
<td>1094</td>
<td>-546</td>
</tr>
<tr>
<td>11</td>
<td>-844</td>
<td>1691</td>
<td>1690</td>
<td>-842</td>
</tr>
<tr>
<td>12</td>
<td>-396</td>
<td>794</td>
<td>794</td>
<td>-386</td>
</tr>
<tr>
<td>13</td>
<td>-517</td>
<td>1036</td>
<td>1036</td>
<td>-506</td>
</tr>
<tr>
<td>14</td>
<td>-341</td>
<td>684</td>
<td>684</td>
<td>-340</td>
</tr>
<tr>
<td>15</td>
<td>-876</td>
<td>1755</td>
<td>1754</td>
<td>-876</td>
</tr>
<tr>
<td>16</td>
<td>-1923</td>
<td>3849</td>
<td>3848</td>
<td>-1887</td>
</tr>
<tr>
<td>17</td>
<td>-510</td>
<td>1022</td>
<td>1022</td>
<td>-486</td>
</tr>
<tr>
<td>18</td>
<td>-101</td>
<td>204</td>
<td>204</td>
<td>-96</td>
</tr>
<tr>
<td>19</td>
<td>-232</td>
<td>466</td>
<td>466</td>
<td>-221</td>
</tr>
<tr>
<td>20</td>
<td>-851</td>
<td>1705</td>
<td>1704</td>
<td>-840</td>
</tr>
<tr>
<td>21</td>
<td>-122</td>
<td>246</td>
<td>246</td>
<td>-118</td>
</tr>
<tr>
<td>22</td>
<td>-465</td>
<td>932</td>
<td>932</td>
<td>-450</td>
</tr>
<tr>
<td>23</td>
<td>-938</td>
<td>1879</td>
<td>1878</td>
<td>-930</td>
</tr>
<tr>
<td>24</td>
<td>-556</td>
<td>1114</td>
<td>1114</td>
<td>-537</td>
</tr>
<tr>
<td>25</td>
<td>-201</td>
<td>404</td>
<td>404</td>
<td>-199</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

42
Appendix B: Study of email newsletter frequency

To understand the extent to which companies are likely to benefit from real-time testing, we set out to get an indication of the number of companies who are engaged in frequently sending emails (e.g. once a week or more) to their permission-based lists. We subscribed a group of pseudo-recipients to 196 newsletters and monitored their activity for six months. These newsletters belonged to a wide range of companies (Airlines, Apparel, News, Electronics…). Since some companies may customize their newsletters we wanted to ensure there was some heterogeneity in the “recipients” we used. Thus whenever the registration process asked for gender information we registered different male and female accounts; when age information was asked, we registered both a 25 and a 45 year old. In addition, for each newsletter-age-gender combination, we registered three different recipients: one for which the incoming emails would be left untouched; one for which the emails would be opened; and one for which the emails would be opened and, if they contained links, those links would be clicked on to show interest in the content. Hence, for a web site that asks for both gender and age information, 12 recipient accounts would be opened (2x2x3). If the web site did not ask for any demographic information, only three accounts would be opened. Of the 196 newsletters, 16% asked for either gender or age information (split evenly between the two), another 16% asked for both types of information. A total of 107 (55%) of the newsletters actually contacted our recipients with enough frequency to be considered as actively engaging in Permission-Based Marketing (we set the cut-off at three or more emails per account during the six-month period) for a total of 12,946 emails received. The level of customization among the active newsletters was low (see Table A3). Four percent of the newsletters customized based on gender information (10% of the active newsletters that asked for gender information), five percent customized based on age (22% of the active newsletters that asked for age information), and five percent customized based on consumer actions (i.e., open or click). Finally, 61% of the newsletters used a fixed contact periodicity; weekly newsletters (63%) being the preferred contact interval (see Table A4).

Table A3: Description of Newsletters

<table>
<thead>
<tr>
<th>Gender Information</th>
<th>Subscribed</th>
<th>Received more than 3 emails from company</th>
<th>Received customized content</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>196</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td>Gender info</td>
<td>48</td>
<td>24%</td>
<td>41</td>
</tr>
<tr>
<td>Age info</td>
<td>47</td>
<td>24%</td>
<td>23</td>
</tr>
</tbody>
</table>

Table A4: Contact Frequency of Subscribed Newsletters

<table>
<thead>
<tr>
<th>Periodicity</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td>Bi-Weekly</td>
<td>3</td>
<td>5%</td>
</tr>
<tr>
<td>Weekly</td>
<td>41</td>
<td>63%</td>
</tr>
<tr>
<td>Bi-Monthly</td>
<td>14</td>
<td>22%</td>
</tr>
<tr>
<td>Monthly</td>
<td>6</td>
<td>9%</td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>