Improving the Effectiveness of Time-Based Display Advertising

Daniel G. Goldstein
Yahoo! Research
New York, NY, USA
dgg@yahoo-inc.com

R. Preston McAfee
Yahoo! Research
Burbank, CA, USA
mcafee@yahoo-inc.com

Siddharth Suri
Yahoo! Research
New York, NY, USA
suri@yahoo-inc.com

ABSTRACT

Display advertisements are typically sold by the impression where one impression is simply one download of an ad. Through an online behavioral experiment, we previously showed that the longer an ad is in view, the more likely a user is to remember it and that there are diminishing returns to increased exposure time [4]. A time-based pricing scheme is more exact than an impression-based scheme on various memory metrics. Thus, the time-based scheme is more economically efficient and may become an industry standard. We answer an open question along this line of research: given a time-based pricing scheme, how should time slots for advertisements be divided? We provide evidence that ads can be scheduled in a way to lead to greater total recollection, which advertisers value, and increased revenue, which publishers value. We document two main findings. First, we show that displaying two shorter ads results in more total recollection than displaying one longer ad of twice the duration. Second, we show that this effect disappears as the length of the shorter ads increases. We also give a theoretical model can account for both phenomena.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Economics

General Terms

Economics, Experimentation

Keywords

display, advertising, memory, recall, recognition, exposure, time

1. INTRODUCTION

Display advertising is a $10 billion dollar per year industry [10] where advertisers pay publishers to place ads next to content on publisher websites. Display ads are often sold in these transactions on a per impression basis, where an impression is simply one download of an ad. Furthermore, agreements between publishers and advertisers are often made through guaranteed contracts. For example, a publisher might agree to deliver 10 million impressions to men age 50–70 on finance related pages, or 8 million impressions to people interested in sports. Display ads are also sold on exchanges like Yahoo!'s Right Media Exchange (RMX) or Google's DoubleClick exchange (GDC). In this context a publisher would auction the right to show a display ad targeted to a specific user as the page loads in real time. Whether display ads are sold via contract or via an exchange, they are mostly sold by the impression.

In addition to increasing short-term sales [6, 11], advertisers seek to increase their brand recognition and brand awareness with display ads [3]. As a result, advertisers measure the effectiveness of brand advertising using memory metrics [17]. Our previous work established a causal link between the amount of time an ad is in view and the probability that a user will remember it on recall and recognition tasks [4], which are proxies for ad effectiveness that have been in use for nearly a century [15]. In addition, these experiments showed that there are diminishing returns to increased exposure time. That is, there was a steep increase in the probability of remembering for exposure times up to roughly 40 seconds, followed by a less steep, although still increasing, effect of time beyond that. These results suggest that time of exposure, as opposed to the number of impressions delivered, may be a better standard for pricing ads since it more precisely influences the branding impact which display advertisers seek.

In this work we study how a publisher might sell time-based ads to increase the total effect on memory per unit of time. Since advertisers value ad recognition and ad recall, such a scheme would also increase the revenue of the publishers, thus it would be beneficial to both parties. The central question this work addresses is the following. If display ads are sold based on time, how should time slots be divided and scheduled? Should publishers show many different ads of short duration or display fewer ads of longer duration? Longer duration ads might increase the likelihood that a user sees them. Alternatively, swapping out shorter ads gives users more ads to notice. Thus it is not *a priori* clear which would result in more overall memory.

To answer this question we conduct an online behavioral experiment where users read an article and either one display ad is shown for t seconds which is then replaced by another display ad for t seconds, or one display ad is shown for $2t$ seconds. Two phenomena emerge. First, we find that two 10 second ads result in 39%–63% more total memory than one 20 second ad, depending on the specific memory metric used. Second, two 20 second ads result in roughly the same amount of total memory as one 40 second ad. We show that a standard model where users have the same, constant probability of noticing an ad during each time step does not capture these phenomena. Thus to reconcile these findings we propose a slightly more complex theoretical memory model which does.

1.1 Problem Definition

Assume that the total amount of time users spent on a publisher's website is t (measured in seconds). Now say that the publisher divides this time in to slots of s seconds each. Assume that the publisher gets paid c dollars for each person who remembers the ad. Thus the publisher would like to place ads in time slots so as to maximize the total amount of recollection in each time slot. Thus, the publisher would like to maximize

$$\sum_{i \in [t/s]} \sum_{\text{ad } a \text{ in slot } i} c \Pr(\text{user recollects ad } a). \quad (1)$$

Publishers have few levers to use to influence the above expression. First, they can change the duration s of each time slot. In the experiment that follows, we will try two different values of $s = 20, 40$ seconds. Second, the publisher could place one ad in the entire time slot or split each time slot, placing one ad in the first half of the time slot and another ad in the second half of the time slot. Thus we will compare the recollection rate of showing one ad for the whole time slot or two ads, one in each half of the time slot. At first glance this might seem unfair since it appears we are comparing one probability to the sum of two probabilities. Recall that the advertiser is getting paid proportional to the total amount of memory for the ads shown. Thus the publisher would like to maximize total recollection per time slot. Moreover, the units here are not in terms of probabilities, but in terms of dollars. So if we let p_1 denote Pr(remember an ad shown in first half of slot), p_2 denote Pr(remember an ad shown in second half of slot) and p_{12} denote Pr(remember an ad shown in both slots), then comparing $cp_1 + cp_2$ to cp_{12} is the same as comparing $p_1 + p_2$ to p_{12}.

It is not *a priori* clear that splitting a time slot will benefit the advertisers. It could be that for advertiser a, $p_1^a + p_2^a > p_{12}^a$, but for advertiser b, $p_1^b + p_2^b < p_{12}^b$. We will show empirically that advertisers are not worse off by splitting time slots.

2. RELATED WORK

Before we discuss how our work relates to the literature, we define and motivate the metrics we use to gauge the effectiveness of a display advertisement. *Unaided recall* is the proportion of site viewers who report remembering an advertiser with only a minimal prompt such as "Which advertisers, if any, did you remember being present on the website?" *Recognition* metrics use prompts. *Text recognition* uses the name of the advertiser as a prompt, e.g. "Did you see a Netflix ad on the previous page?". *Visual recognition* uses the actual image of the ad to verify memory. A vast literature studies the effectiveness of television advertising using these and related metrics [7]. Dreze and Hussheer conducted a study on banner advertisements where they advocate the use of these memory metrics in online advertising [3]. When referring to generally affecting the memory of an ad, whether it be recall or recognition, we will use the term "recollection." Surprisingly few studies, however, have considered the effectiveness of online advertising in improving recollection. We will describe the most relevant of these next.

In our previous work we showed that exposure time has a causal effect on memory for an ad [4], whereas prior work had established only a correlation (see [2] and commentary in [4]). Moreover, we showed that there are diminishing returns to this effect. The first seconds of exposure caused a steep increase in the memory for an ad, and further exposure time had a smaller, albeit still increasing impact on recollection. The conclusion of that work is that as exposure time exerts a causal influence on the probability of an display ad being remembered. The implication of this work is that given advertisers who value memory of their ads, time of exposure is a more exact measure and thus a more efficient basis for pricing: charging based on what advertisers value allows for price discrimination and efficient allocation of advertising slots. While that work laid the groundwork for this research, it gave no guidance to advertisers and publishers as to how display ads should be sold. This is the exact question this work seeks to address. More specifically, we seek to understand how to allocate time slots to influence overall memory for ads.

Sahni [13] conducted a field experiment on a widely used restaurant search website in India. The design of the experiment allowed for exogenous variation in the number and frequency of sponsored search ads users saw on the site over the course of two and a half months. The author did not study the exposure time of ads, but rather the amount of time between exposures. Braun and Moe looked at a similar phenomenon using a theoretical model [1]. The key result of Sahni's work is that increasing the time between exposures, up to two weeks, increases the probability of a purchasing event. Thus, this result may be viewed as a complement to ours. We investigate how the length and timing of exposures influences recollection, while Sahni shows how the amount of time between exposures influences purchasing.

3. METHODS

The experiments reported here were conducted on Amazon Mechanical Turk[1]. Mechanical Turk is an online labor market where requesters can post jobs and workers can choose which jobs to do for pay. After a worker submits a job, the requester can either accept or reject the work based on its quality. The fraction of jobs that a worker submits which are accepted is that worker's approval rating which functions as a reputation mechanism used to help ensure work quality. Mechanical Turk was originally built to accomplish tasks that are easy for humans but hard for machines like image recognition, audio transcription and adult content classification. Hence jobs on Mechanical Turk are called "Human Intelligence Tasks" or "HITs". There is a burgeoning literature in the academic community around using Mechanical Turk as a platform for online behavioral experiments [8, 14, 12]. In this setting, experimenters take on the role of requesters and post their experiment as a HIT and workers are the paid participants in the experiment. Recent studies show behavior observed in Mechanical Turk

[1] www.mturk.com

Figure 1: Screenshot of the article with the Netflix ad.

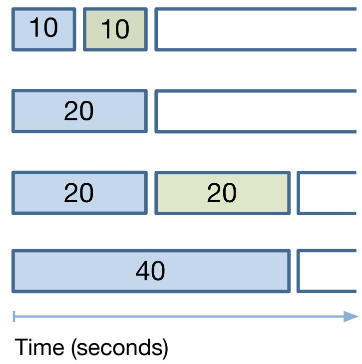


Figure 2: The four time treatments in which participants were randomly placed. Each colored rectangle represents an ad with the number of seconds it was in view. The white rectangles on the right side of the figure indicate the absence of an ad.

experiments matches behavior observed in university lab experiments extremely well [9, 5, 16].

We used the Mechanical Turk API to restrict our participant pool to workers in the United States to help ensure that they can read and understand English. We also restricted to those workers who have an approval rating of 90% or more. The Amazon API gives each worker account a unique, anonymous identifier. By storing these WorkerIDs we were able to ensure that a worker could only do the experiment one time. In all, we had 1,100 participants. The experiment ran over the course of two, roughly one week periods. Next we describe the format of the experiment and the various treatments to which the participants were randomly assigned.

3.1 Experimental Design

Participants were paid a \$0.50 flat rate for the HIT plus \$0.10 for each question answered. We chose not to pay based on the correctness of the answers to alleviate incentives for sharing answers between workers. The preview page of the HIT consisted of a brief consent form along with the instructions indicating that the HIT involved reading a web page and answering questions about it.

After reading and accepting the instructions participants were then shown an image of a webpage from an actual Yahoo! site. The article image consisted of text with images along with a display ad. See Figure 1 for a screenshot. Since 99% of screens on the Web can show an image of 600 pixels in height[2] we chose this to be the height of the article image to ensure that the article and display ad were always in view in their entirety and the user never had to scroll to see any part of them.

The goal of this research is to compare the effectiveness of two shorts ads to one longer one. Thus, for each memory metric we compared the sum of the metric over the two short ads to the metric measured on one long ad. The two short ads treatments necessarily involves two advertisements, and of course different advertisements may be differentially memorable. To hold all of this constant, the simplest test involves two orderings of the short ads. Denote the treatment that shows ad A followed by ad B as AB. Then the simplest test is to compare the effectiveness of AB for one user and BA for a second user, to AA for a third user and BB for a fourth. The AB/BA treatment shows each ad for the same amount of time and in the portions of an impression as the AA/BB treatment, and thus has dedicated the same amount of resources to each advertiser. If the total effectiveness of AB/BA exceeds AA/BB, then publishers should split time slots between two advertisers.

On the experimental webpage, all objects were static except for the display ads, which were changed in two different ways. In one class of treatments, an ad was displayed for t seconds, then replaced by another ad for t seconds, which was then replaced by whitespace. In a second class of treatments, an ad was displayed for $2t$ seconds and then replaced with whitespace. This design allowed us to compare the memory of two ads shown for t second each to the memory of one ad shown for $2t$ seconds, as described above. We used one pair of time treatments where $t = 10$ and another pair of time treatments where $t = 20$ resulting in four time treatments overall. Figure 2 gives a pictorial representation of our time treatments.

In pursuit of external validity we used four different ads as stimuli. These ads were actually used at some point, but were discontinued well before the time of the experiment. One ad treatment had a Netflix ad shown first and a Jeep ad shown second and the second ad treatment had the opposite order. The third ad treatment had an Avis ad shown first and an American Express ad shown second and the fourth ad treatment had the opposite order. In "single ad" time treatments, in which only one ad was shown, the second ads in the order described above were left out. See Figure 5 in Appendix A for pictures of the ads. In all, the four time treatments and the four ad treatments yielded a 4×4, between subjects design. Subjects were randomly placed into one of these 16 treatments at the point of accepting the HIT to avoid any confound between dropping out of the experiment and the treatment assigned.

After participants finished reading the article at their own pace they clicked a link and were taken to a page where the played a game for a fixed amount of time. We chose Tetris

[2]See `http://www.w3schools.com/browsers/browsers_display.asp`

for the game as it is widely known and generally considered fun. Also, it is a visual game consisting of primary shapes which would avoid ad-specific linguistic memory interference. The game, and the pieces in the game, were rendered in black and white to avoid interference with the colors in the ads. The game time was chosen such that, on average, the amount of time between the first ad disappearing and the following questionnaire is the same. This ensures that on average, everyone experiences roughly the same amount of time to forget the ad between being exposed to it and asked about it. After this amount time expired, users were automatically directed to a questionnaire.

To measure the effectiveness of these treatments, we adopt the standard industry metrics involving recall and recognition, splitting recognition into text recognition and visual recognition. Using standard industry metrics is important to permit a conclusion that shorter duration ads improve what matters to advertisers. We will primarily focus on influencing each of the three metrics, meaning that we examine the effectiveness of short duration ads at enhancing recall, and text or visual recognition.

Once faced with the questionnaire, participants were unable to press the "back" button on their browser to return to the article. Participants were asked two multiple choice reading comprehension questions about the article on the previous page, after which they were asked an unaided recall question: "Which advertisements, if any, did you see on the page during this HIT? Type the name of any advertisers here if you can remember seeing their ads on the last page, or indicate that you are unable to remember any." The next page then consisted of four separate recognition questions with textual cues of the form, "Did you see a ___ ad?" with Netflix, Jeep, Avis, and American Express being the advertisers filling in the blank. After answering these questions participants then went to a page which consisted of four separate recognition questions with pictorial cues of the form, "Did you see the following ad?" with a picture of the Netflix, Jeep, Avis, and American Express ads following each question. The ads were chosen such that each had a strong visual resemblance to another ad to approximate an upper bound on the false positive rate, exhibited in Figure 5 in Appendix A. The Avis lure is primarily red, much like the Netflix ad, and the American Express lure is primarily black, much like the Jeep ad. Thus when the Netflix ad was shown, the Avis ad acted as its "lure" ad and *vice versa*. Similarly when the Jeep ad was shown the American Express ad acted as its lure ad and *vice versa*. The lure ads were used to measure the false positive rate for remembering an ad.

The data from a participant were encoded as 12 binary responses. The first four responses coded mentions of the two target ads and the two lures from the unaided recall question. The next four binary responses coded the recognition questions with textual cues and the final four responses coded the recognition questions with visual cues.

4. RESULTS

As mentioned, the experiment had 1,100 participants. Of those we excluded 42 for incomplete responses. An additional 22 were excluded for missing both of the reading comprehension questions. By piloting the experiment we measured that roughly 80% of participants take more than 50 seconds to read the article. Participant who took fewer than 40 seconds to read the article would not have been exposed to the intended time treatment and were accordingly removed before analysis, resulting in 110 exclusions. In addition, we excluded participants who took over four minutes to read the article as they likely were interrupted during the experiment, yielding another 10 exclusions. The data of the remaining 916 participants makes up the set we analyze.

Across all conditions, false alarm rates (the rate of incorrectly indicating memory for one of the visually similar lure ads) were low and quite similar to those in [4]: 0% for recall, 6.6% for text recognition and 7.5% for visual recognition questions.

Table 1 addresses whether a greater total probability of remembering an ad is achieved with two ads of length t or one ad of length $2t$. For each of the three memory metrics, the sum of the metric for the two 10 second ads is significantly higher than the metric for the single 20 second ad. For example, on the visual recognition measure, the expectations of two ads sum to .578, while that of a single, 20 second ad is only .354, a 63% increase. Text recognition and unaided recall show a similar pattern with 53% and 39% increases respectively. The p-values listed are the probability of these data given the hypothesis that the sum of p_1 and p_2 is less than the single ad. Thus, when t is 10 seconds, the total amount of recollection in two shorter ads tends to win over a longer ad. The top 3 panels of Figure 3 are a graphical representation of these data. Here for all three metrics, the dotted green line indicating the sum of the memory measure for two 10 second ads is significantly higher than the memory measure for one 20 second ad.

However, when t is 20 seconds, a somewhat different picture emerges. The differences in Table 1 are in the same direction, but not statistically significant. This can also be seen graphically in the bottom three panels of Figure 3. Here the dotted green line indicating the sum of the memory rates of the two 20 second ads is higher, although not significantly higher than, the memory rates of one 40 second ad. One conclusion is thus that a pair of shorter ads lead to more total memory than one ad shown for twice as long, but only when the short ads are relatively short themselves (around 10 seconds).

We have shown that if A and B are advertisers, more total impact on memory is created when splitting an impression between two advertisers than giving each advertiser its own full slot. That is, in the terminology established earlier, the memory under AB + BA is greater than memory under AA + BB. However, it is in principle possible for this inequality to hold in general, but not for both advertisers A and B. For example, advertiser A could benefit greatly from the split impressions, while advertiser B suffers slightly. To check whether this occurs in practice, we take advantage that the experimental design uses four unique advertisers, each of which can be used as a test to see whether two short ads lead to more recall than one ad of twice the duration. The results are shown in Table 2. Again, a difference between the "10+10 vs. 20" and the "20+20 vs. 40" condition emerges. In the former case, in 11 of 12 tests, the sum of the shorter ads exceeds the longer ad, while in the latter case, this effect is diminished and the sum beats the whole in only 8 of 12 tests. P-values are reported in Table 2, though it should be noted that the sample sizes here are four times smaller, decreasing statistical power. Nonetheless, the consistent sign and at times sizeable absolute magnitude of the differences in the

Condition	Measure	p_1	p_2	Sum	Single	p-value
10+10 vs. 20	Visual Recognition	.372(.03)	.206(.03)	.578(.04)	.354(.03)	<.001
	Text Recognition	.215(.03)	.161(.02)	.377(.04)	.247(.03)	.002
	Recall	.117(.02)	.094(.02)	.211(.03)	.152(.02)	.06
20+20 vs. 40	Visual Recognition	.407(.03)	.165(.02)	.572(.04)	.519(.03)	.16
	Text Recognition	.280(.03)	.114(.02)	.394(.04)	.374(.03)	.34
	Recall	.208(.03)	.042(.01)	.250(.03)	.215(.03)	.20

Table 1: Comparison of the sums of the memory rates for two successive ads shown t seconds each compared to one ad shown for $2t$ seconds. p_1 and p_2 are the mean memory rates for the first and second ad from the two-ad condition. *Sum* is $p_1 + p_2$. *Single* is the mean of the single-ad condition. Standard errors are given in parentheses. The p-values are bootstrapped estimates of the probability of the data given the hypothesis that the sum is less than the single measure.

Condition	Advertiser	Measure	p_1	p_2	Sum	Single	p-value
10+10 vs. 20	Avis	Visual Recognition	.241(.06)	.207(.05)	.448(.08)	.286(.06)	.05
		Text Recognition	.155(.05)	.121(.04)	.276(.06)	.175(.05)	.10
		Recall	.086(.04)	.103(.04)	.190(.05)	.079(.03)	.04
	American Express	Visual Recognition	.466(.07)	.207(.05)	.672(.08)	.383(.06)	.002
		Text Recognition	.190(.05)	.190(.05)	.379(.07)	.250(.06)	.08
		Recall	.086(.04)	.035(.02)	.121(.04)	.133(.04)	.55
	Netflix	Visual Recognition	.436(.07)	.288(.06)	.725(.09)	.446(.07)	.006
		Text Recognition	.218(.06)	.288(.06)	.507(.08)	.375(.06)	.11
		Recall	.127(.04)	.231(.06)	.358(.07)	.25(.06)	.13
	Jeep	Visual Recognition	.346(.07)	.127(.04)	.473(.08)	.312(.06)	.05
		Text Recognition	.308(.06)	.055(.03)	.362(.07)	.203(.05)	.03
		Recall	.173(.05)	.018(.02)	.191(.06)	.156(.05)	.30
20+20 vs. 40	Avis	Visual Recognition	.417(.06)	.221(.05)	.637(.08)	.500(.07)	.09
		Text Recognition	.278(.05)	.118(.04)	.395(.07)	.269(.06)	.08
		Recall	.181(.05)	.044(.02)	.225(.05)	.115(.04)	.06
	American Express	Visual Recognition	.397(.06)	.181(.05)	.578(.08)	.422(.06)	.05
		Text Recognition	.265(.05)	.139(.04)	.404(.07)	.344(.06)	.25
		Recall	.162(.04)	.042(.02)	.203(.05)	.188(.05)	.40
	Netflix	Visual Recognition	.418(.07)	.146(.06)	.565(.09)	.605(.07)	.64
		Text Recognition	.345(.06)	.122(.05)	.467(.08)	.581(.08)	.85
		Recall	.291(.06)	.073(.04)	.364(.07)	.326(.07)	.35
	Jeep	Visual Recognition	.390(.08)	.091(.04)	.481(.09)	.582(.07)	.82
		Text Recognition	.220(.06)	.073(.04)	.292(.07)	.345(.06)	.71
		Recall	.220(.06)	.018(.02)	.238(.07)	.255(.06)	.57

Table 2: Advertiser-specific variant of Table 1.

"10+10 vs. 20" recognition conditions lend support to the conclusion that the benefits of shorter ads hold for all four advertisers studied. Figure 4 graphically shows the "10+20 vs. 20" visual recognition data. In interests of space, data for the other conditions are provided in Table 2.

4.1 Remembering at Least One Ad

Another way in which we could compare the memory for two short ads vs. one long ad is to compare the probability of remembering either ad in the two short ad treatment to the probability of remembering the ad in the longer treatment. To do this formally and equivalently, one must compare the probability of not remembering either ad in the short ad treatments to the probability of not remembering the ad in the long ad treatments. Table 3 shows that on the measure of not remembering any ads, one is more likely not to remember anything when shown one longer ad as compared to two shorter ads. For example, again on the visual recognition measure in the "10+10 vs. 20" condition, the probability of *not* recognizing is .646 when shown one ad, but drops to .520 when participants are shown two ads. These differences are significant at the .05 level for text and visual recognition, but not for recall, and the results are always in the predicted direction. Thus even in this measure, showing two 10 second ads is better than one 20 second ad, and this effect again disappears at longer time intervals.

4.2 Effect of the Second Ad on the First

Curiously, it appears as if recall for an ad can be improved by placing another ad after it, which can be seen by comparing the memory rates for the first 20 second ads in the two ad conditions (the values of p_1 in rows 4–6 in Table 1) to ads shown for 20 seconds in the one ad condition (the values of *Single* in rows 1–3 of the same Table). A similar comparison can be made with the same cells in Table 3, where the probabilities of *not* recognizing are greater when an ad is not followed by another than when it is. However, these effects are not statistically significant (Two-tailed p-values: Visual recognition: .27, Text Recognition: .48, Recall .14).

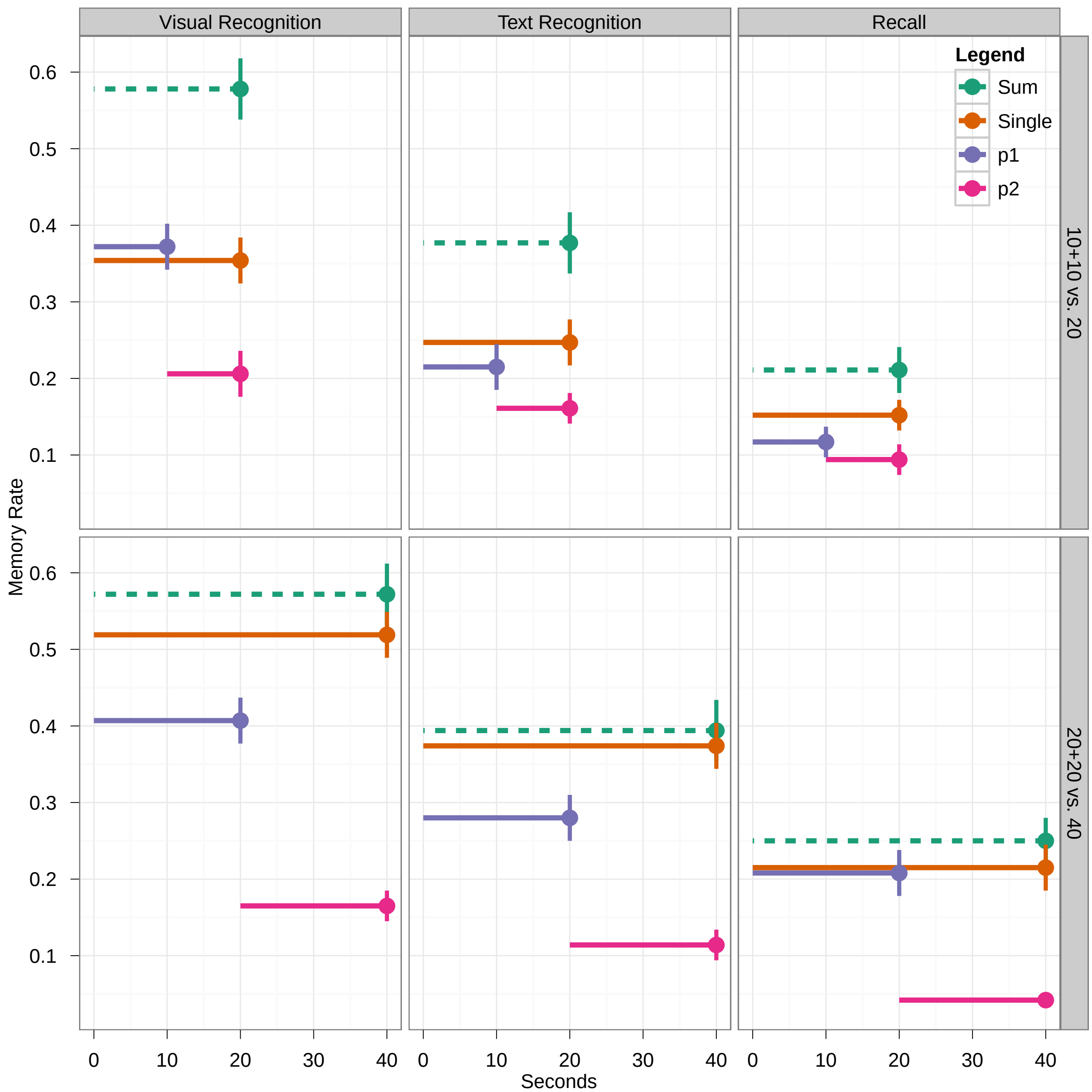


Figure 3: The x-coordinate of the leftmost endpoint of each horizontal line indicates the time at which the ad appeared (either at 0, 10, or 20 seconds). The length of each line indicates the duration of the ad. The y-axis indicates the probability of remembering the ad according to the 3 memory metrics. Vertical line segments are confidence intervals of one standard error. For each memory metric, the dotted green line shows sum of the metric for the two short ads. The top 3 panels compare two 10 second ads to one 20 second ad. The bottom 3 panels compare two 20 second ads to one 40 second ad.

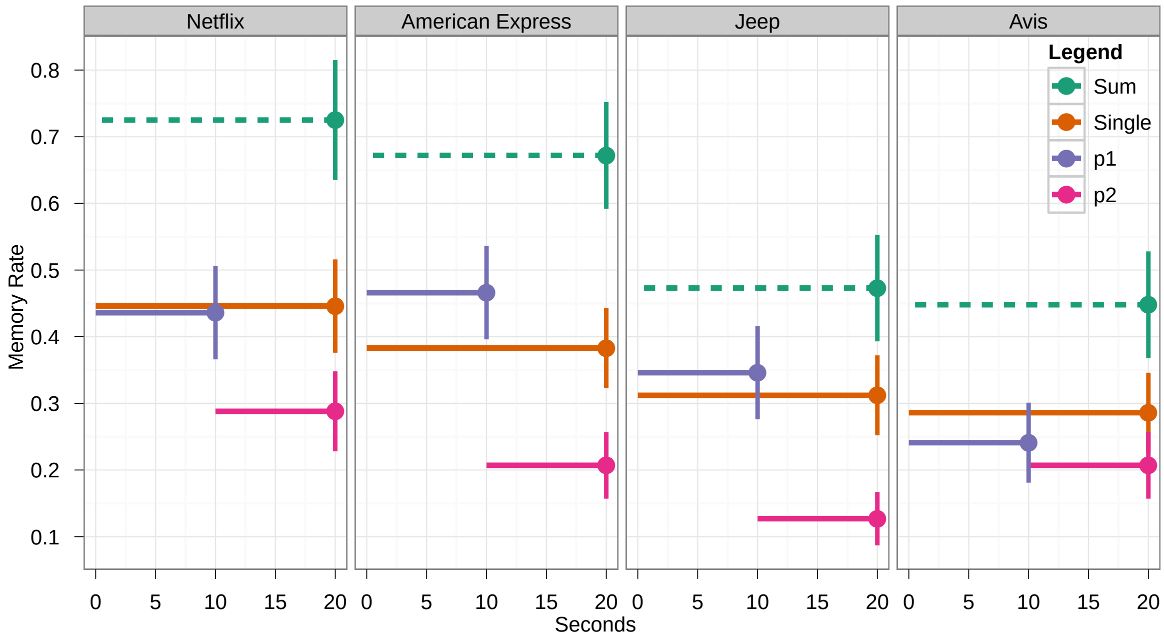


Figure 4: The visual recognition rate only for each of the four advertisers in the 10+10 vs. 20 second condition. The x-coordinate of the leftmost point of each horizontal line indicates the time at which the ad appeared. The length of the line indicates the duration the ad remained in view (i.e. the exposure time). The y-axis indicates the probability of remembering the ad in the visual recognition test. The vertical line segments are confidence intervals of one standard error. The dotted green line shows the sum of the metric for the two 10 second ads.

Condition	Measure	q_1	q_2	Cov	Product	Single	p-value
10+10 vs. 20	Visual Recognition	.628(.03)	.794(.03)	.022	.520(.03)	.646(.03)	.003
	Text Recognition	.785(.03)	.839(.02)	.006	.664(.03)	.753(.03)	.02
	Recall	.883(.02)	.906(.02)	.007	.807(.03)	.848(.02)	.12
20+20 vs. 40	Visual Recognition	.593(.03)	.835(.02)	.001	.496(.03)	.481(.03)	.62
	Text Recognition	.720(.03)	.886(.02)	.002	.640(.03)	.626(.03)	.62
	Recall	.792(.03)	.958(.01)	.008	.767(.03)	.785(.03)	.32

Table 3: Comparison of the products of the non-memory rates for two successive ads shown t seconds each compared to one ad shown for $2t$ seconds. q_1, q_2 and *Single* are the complements of p_1 and p_2 and *Single* in Table 1. Cov is the covariance of q_1 and q_2. *Product* is the probability of not remembering either of the two ads shown. Standard errors are given in parentheses. The p-values are bootstrapped estimates of the probability of the data given the hypothesis that the product is greater than the single measure.

5. MODEL

To understand this complex pattern of results, it is helpful to consider a simple model, which posits that the ad will be encoded to memory after it is attended to, and the probabilities of attending to an ad are independent and identically distributed per unit time. As a consequence, the probability of not remembering an ad is just the product of the probabilities of not attending to it during each second of exposure. In this simple model the probability of noticing the ad, y after T time steps would be

$$y = 1 - \prod_{t=1}^{T}(1-r) = 1-(1-r)^T, \tag{2}$$

where r is the recall rate for each time step. It is easy to see why this model would not explain our data well. It would predict that an ad shown from 0–10 seconds would have the same recall rate as one shown from 10–20 seconds. The data for p_1 and p_2 in Figure 3 shows that this is clearly not the case. Thus we augment the model so that the recall rate r decreases with time. Equation 2 also assumes that all ads have the same recollection rate. Figure 4 shows that this is not the case so we augment the model with and advertiser parameter. Also, our empirical work cannot distinguish between advertisement variation and user variation because we subjected individual users to at most two advertisements. But, Table 3 shows weak correlation of user recollection across ads, suggesting the advertisement attributes, rather that user characteristics, is the larger source of variation.

Again we discretize time in seconds, for example. The augmented model specifies that the probability that a user recalls an advertisement is

$$y = 1 - \prod_{t=1}^{T}(1-r_t)^{ax_t}, \tag{3}$$

where r_t is the receptivity at time t, x_t is a dummy variable that is 1 if the advertisement is visible at time t and 0 otherwise, and a is a parameter that varies with the advertiser, and measures the saliency or receptivity of the particular ad. A higher value of a indicates increased saliency. This formalizes what we call the *independent memory* model. It has the property that the probability of recollection is determined by independent "recollection events" which happen with probability r_t at time t. In addition, advertisement specific features are accommodated by a parameter a.

Note that the "unit of time" does not matter because the model aggregates by using a simple change of variables in Equation 3. For example, modeling the recollection rate using two ten second intervals is accommodated by using recollection events with probabilities R_1, R_2, in $y = 1 - \prod_{t=1}^{2}(1-R_t)^{ax_t}$, etc. where

$$R_1 = 1 - \prod_{t=1}^{10}(1-r_t),\ R_2 = 1 - \prod_{t=11}^{20}(1-r_t).$$

Equation 3 displays the decreasing returns found empirically by having the r_t decrease with t. It also captures gains from short time intervals. In fact, according to this model splitting a time unit always produces gains. Let y_t^a be the recall probability of advertisement a at time t, and let $t = 12$ indicate "after times 1 and 2". Then our experiment compares running ad a followed by b, plus ad b followed by a, to ad a followed by a and ad b followed by b. This produces a positive net gain:

$$\begin{aligned}
y_1^a &+ y_2^b + y_1^b + y_2^a - y_{12}^a - y_{12}^b \\
&= 1-(1-r_1)^a + 1-(1-r_2)^b + \\
&\quad 1-(1-r_1)^b + 1-(1-r_2)^a - \\
&\quad (1-(1-r_1)^a(1-r_2)^a) - (1-(1-r_1)^b(1-r_2)^b) \\
&= (1-(1-r_1)^a)(1-(1-r_2)^a) + \\
&\quad (1-(1-r_1)^b)(1-(1-r_2)^b) \qquad (4) \\
&> 0
\end{aligned}$$

An easy way to see this is the case where ad b is not memorable at all, that is $b = 0$. In this case using impressions produces a recall of $1-(1-r_1)^a(1-r_2)^a$. In contrast, splitting the time produces a recall of $1-(1-r_1)^a+1-(1-r_2)^a$, which is larger.

Figure 3 showed that the memory rates for the three metrics is fairly low in the second 20 second interval. So,

$$1 - \prod_{t=21}^{40}(1-r_t)^a$$

is a low number. Thus the terms of the form $1-(1-r_2)^a$ or $1-(1-r_2)^b$ in Equation 4 are small, and so the gains from splitting 40 second ads are relatively low.

6. CONCLUSION

Display ads are currently often sold by impression. In this pricing scheme a two second impression cost the same as a two minute impression. But the user who experienced the two minute impression is far more likely to remember the ad than the user who experienced a two second impression. If the total amount of time a publisher gets is divided into slots, we have shown that two, short ads increases memory per slot over a single, longer duration ad. In addition to increasing the total memory per slot, this scheme also increases the probability of remembering at least one ad per slot. We have also shown that advertisers are also better off in this scheme because their ads gain more memory. Since advertisers value recollection of their ads, they would be willing to pay more for such a scheme. Thus, the two, short ad scheme benefits both advertisers and publishers. This result strengthens the case for moving from an impression based pricing scheme to one based on exposure time.

Increasing the memory for advertisers will likely create additional benefits for internet publishers and advertisers at the expense of non-internet publishers such as television. In a time based pricing scheme advertisers would able to improve their metrics while buying the same amount of time on the internet thereby improving advertiser returns. This improvement will induce advertisers to substitute internet advertising for some other forms of advertising, bringing more money online, increasing online publisher revenues. Advertisers must be better off as a group, because the loss of demand for offline advertising will tend to lower the prices of offline advertising. Since, the performance of offline ads are the same and there is less money spent on offline advertising the value of offline advertising goes up. Furthermore, since online advertising must be competitive with offline advertising, online advertising performance per dollar must rise. Thus some of the value from impression splitting is captured by advertisers, and some is captured by publishers.

7. REFERENCES

[1] M. Braun and W. Moe. Online advertising campaigns: Modeling the effects of multiple ad creatives. Working Paper, July 2011.

[2] P. J. Danaher and G. W. Mullarkey. Factors affecting online advertising recall: A study of students. *Journal of Advertising Research*, 43:252–267, 2003.

[3] X. Drèze and F.-X. Hussherr. Internet advertising: Is anybody watching? *Journal of Interactive Marketing*, 17(4), 2003.

[4] D. G. Goldstein, R. P. McAfee, and S. Suri. The effects of exposure time on memory of display advertisements. In *Proceedings of the 12th ACM conference on Electronic commerce*, EC '11, pages 49–58, New York, NY, USA, 2011. ACM.

[5] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory. *Experimental Economics*, 2011. To apper.

[6] R. A. Lewis and D. H. Reiley. Does retail advertising work? measuring the effects of advertising on sales via a controlled experiment on Yahoo! Working Paper, June 2011.

[7] L. M. Lodish, M. Abraham, S. Kalmenson, J. Livelsberger, B. Lubetkin, B. Richardson, and M. E. Stevens. How tv advertising works: A meta-analysis of 389 real world split cable tv advertising experiments. *Journal of Marketing Research*, 32(2):125–139, 1995.

[8] W. A. Mason and D. J. Watts. Financial incentives and the performance of crowds. In P. Bennett, R. Chandrasekar, and L. von Ahn, editors, *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 77–85. ACM, 2009.

[9] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5:411–419, 2010.

[10] PricewaterhouseCoopers. IAB internet advertising revenue report: 2010 full year results, April 2011.

[11] J.-P. D. Puneet, K. Y. Goh, and P. K. Chintagunta. The effect of banner advertising on internet purchasing. *Journal Of Marketing Research*, 43(1):98–108, 2006.

[12] D. G. Rand. The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 2011. To appear.

[13] N. Sahni. Effect of temporal spacing between advertising exposures: Evidence from an online field experiment. Working Paper, September 2011.

[14] A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 275–284, New York, NY, USA, 2011. ACM.

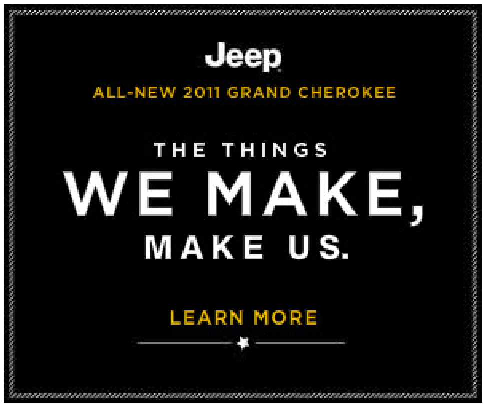
[15] D. Starch. *Principles of Advertising*. A. W. Shaw Company, Chicago, 1923.

[16] S. Suri and D. J. Watts. Cooperation and contagion in web-based, networked public goods experiments. *PLoS One*, 6(3), 2011.

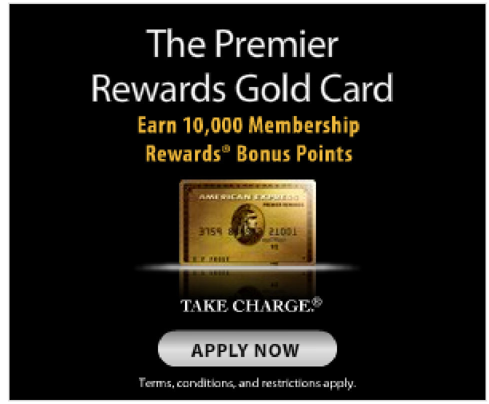
[17] W. Wells. Recognition, recall, and rating scales. *Journal of Advertising Research*, 40(6):14–20, 2000.

APPENDIX

A. ADVERTISEMENTS AND LURES



(a)



(b)

(c)

(d)

Figure 5: The ads used as targets and lures in the experiments. When the Jeep ad (5(a)) was the target, the American Express (5(b)) ad was the lure and *vice versa*. Similarly, when the Netflix ad (5(c)) was the target the Avis ad (5(d)) was the lure and *vice versa*.