Ad Quality On TV: Predicting Television Audience Retention

Yannet Interian, Sundar Dorai-Raj, Igor Naverniouk, P. J. Opalinski, Kaustuv and Dan Zigmond*
Google, Inc.

ABSTRACT

This paper explores the impact of television advertisements on audience retention using data collected from television set-top boxes (STBs)[1]. In particular, we discuss how the accuracy of the retention score, a measure of ad quality, is improved by using the recent "click history" of the STBs tuned to the ad. These retention scores are related to – and are a natural extension of – other measures of ad quality that have been used in online advertising since at least 2005 [2]. Like their online counterparts, TV retention scores could be used to determine if an ad should be eligible to enter the inventory auction and, if it is, how highly the ad should be ranked [1]. A retention score (RS) could also be used by the auction system for pricing, or by the advertiser to compare different creatives for the same product.

1. INTRODUCTION

Online advertisers frequently measure their success and return on investment by using measures such as click through rate (CTR) [6], conversion rate (CvR), and bounce rate (BR) [7]. Since extending our ad platform to television ads in 2007, Google has been exploring ways to design similar measures for TV.

Google aggregates data, collected and anonymized by DISH Network L.L.C., describing the precise second-by-second tuning behavior for millions of television set-top boxes, covering millions of US households, for several thousand TV ad airings every day[2]. From this raw material, we have developed several measures that can be used to gauge how appealing and relevant commercials appear to be to TV viewers. One such measure is the percentage initial audience retained (IAR): how much of the audience, tuned in to an ad when it began airing, remained tuned to the same channel when the ad completes.

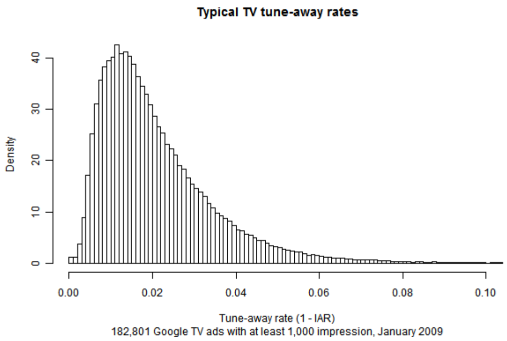

Figure 1: Density of tune away rate for TV ads, defined by the percentage of watchers who click away from an ad.

In many respects, IAR is the inverse of online measures like CTR. For online ads, passivity is negative; advertisers want users to click through. This is somewhat reversed in television advertising, in which the primary action a user can take is a negative one: to change the channel.

However, we see broad similarities in the propensity of users to take action in response to both types of advertising. Figure 1 shows tune-away rates (the additive inverse of IAR) for 182,801 TV ads aired in January 2009. This plots looks remarkably similar to a CTR to a comparable number of randomly selected paid search ads also run that month[3]. Although the actions being taken are quite different in the two media, the two measures show a comparable range and variance.

We believe measures of audience retention could have several important applications in TV. Advertisers could use retention scores (defined more formally in Section 2) to evaluate how campaigns are resonating with customers, for example, while networks and other programmers could use these same scores to inform ad placement and pricing. Like online ad quality scores, audience retention provides insight into users' advertising preferences and is useful whenever know-

[1] A preliminary version of some of these results was presented at the Advertising Research Foundation's Re:Think09 Conference.

[2] These anonymous set-top box data were provided to Google under a license by the DISH Network L.L.C.

[3] TV ads were restricted to those estimated to have at least 1,000 STBs tuned, which was approximately 83% of all ads aired that month. Extreme outliers with tune-away rates larger than 0.1 have been excluded. These totalled less than 0.4% of total TV ads. Impressions for TV ads are defined as the number of STBs that were tuned to the ad for at least 5 seconds.

ADKDD'09, June 28, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-671-7 ...$10.00.

ing such a preference would influence a business decision.

The only TV ad quality-type scores that we are aware of are computed by Nielsen Inc. using survey data in shows with large audiences, such as the Superbowl. For this year's Superbowl, Nielsen published a likeability score and a recall score for the top ads [4]. The scores are computed using 11,466 surveys, and they report on the 5 best liked ads and the 5 most recalled ads.

In this paper we define a more rigorous measure of audience retention for TV ads. The primary challenge in designing such a measure is that many factors appear to impact STB tuning during ads, making it difficult to isolate the effect of the specific ad itself on the probability that a STB will tune away. We propose two ways of modeling such a probability. We show that a model that takes into account the behavior of the STB before the ad better explains the data than a model that uses only information about the time and location (network) of the ad. To the best of our knowledge, we are the first to attempt to derive a measure of TV ad quality from large scale STB data.

2. RETENTION SCORE

In this section we introduce metrics for audience retention and demonstrate that such metrics can be used to rank ads based on their creative appeal.

2.1 Definition

We calculate per airing the fraction of initial audience retained (IAR) during a commercial. This is calculated by taking the number of TVs tuned to an ad when it began and then remained tuned throughout the ad airing as shown in equation (1). The intuition behind this measure is that when an ad does not appeal to a certain audience, viewers will vote against it by changing the channel. By including only those viewers who were present when the commercial started, we hope to exclude some who may be channel surfing. However, even these initial viewers may tune away for other reasons. For example, a viewer may be finished watching the current program on one channel and looking for something else to watch.

$$\text{IAR} = \frac{\text{Audience that Viewed Whole ad}}{\text{Audience at Beginning of the ad}} \tag{1}$$

We can interpret IAR as a probability of tuning out from an ad. Raw, per-airing IAR values are difficult to work with because they are affected by the network, day part, and day of the week, among other factors. In order to isolate these factors from the creative (ad), we define *Expected IAR* of an airing as

$$\widehat{\text{IAR}} = \text{E}(\text{IAR}|\widehat{\theta}), \tag{2}$$

where $\widehat{\theta}$ is a vector of features extracted from an airing, which exclude any features that identify the creative itself; for example, hour of the day and TV channel ID, but not campaign ID or customer ID. Then we define the *IAR residual* as in equation (3) to be a measure of the creative effect.

$$\text{IAR residual} = \text{IAR} - \widehat{\text{IAR}} \tag{3}$$

There are a number of ways to estimate (2); we discuss them in the next section. Using equation 3 we can define *underperforming airings* as the airings with IAR residual below the median. Now that we have a notion of underperforming airings we can formally define the *retention score* (RS) for each creative as one minus the fraction of airings that are underperforming.

$$\text{RS} = 1 - \frac{\text{Number of underperforming airings}}{\text{Total number of Airings}} \tag{4}$$

2.2 Retention Score And Viewer Satisfaction

In order to understand the meaning of retention scores, we conducted a simple survey of 78 Google employees. We asked each member of this admittedly unrepresentative sample to evaluate 20 television ads on a scale of 1 to 5, where 1 was "annoying" and 5 was "enjoyable." We chose these 20 test ads such that 10 of them were considered "bad" while the remaining 10 were considered "good," based on the categories shown in Table 1.

Ad Quality	Retention Score
"good"	> 0.75
"bad"	< 0.25

Table 1: Using retention scores to categorize ads as either "bad" or "good." These categories were matched empirically with a human evaluation survey.

Human evaluation	Mean RS
At least "somewhat engaging"	0.86
"Unremarkable"	0.62
At least "somewhat annoying"	0.30

Table 2: Correlating retention score rankings with human evaluations. Survey scores of 3.5 or above (or "somewhat engaging") received retention scores averaging 0.86, while survey scores of 2.5 or below (or "somewhat annoying") received retention scores averaging 0.30. These numbers match well with categories defined in Table 1.

Table 2 summarizes the results. Ads that scored at least "somewhat engaging" (i.e, mean survey score greater than 3.5) had an average retention score of 0.86 for all creatives. Ads that scored at the other end of the spectrum (mean less than 2.5) had an average retention score of 0.30. Ads with survey scores in between these two had an average retention score of 0.62. These results suggest our scoring algorithm and the categories defined in Table 1 correlate well with how a human being would rank an ad.

Figure 2 gives another view of this data. Here the 20 ads are ranked according to their human evaluation, with the highest-scoring ads on top. The bars are colored according to which set of 10 they belonged to, with gray ads coming from the group that outperformed the model and black ads coming from the group that underperformed. Although the correlation is far from perfect, we see fairly good separation of the "good" and "bad" ads, with the highest survey scores tending to go the ads with the best retention scores.

3. VIEWERS AND THEIR REMOTE CONTROLS

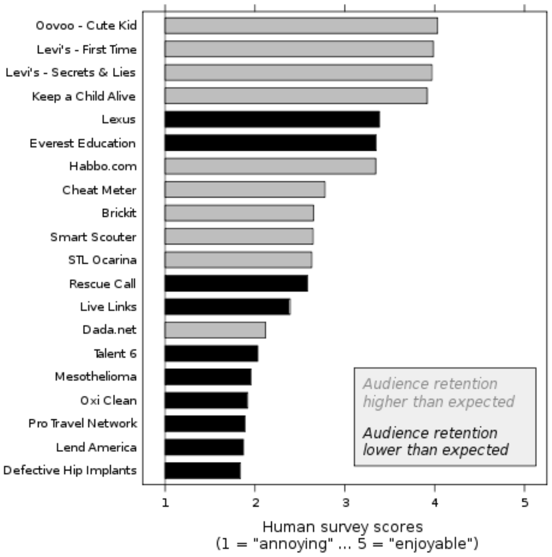

Figure 2: Correlating retention score rankings with human evaluations. Each bar represents the average score given by the human evaluator, with dark bars having lower than expected retention scores and light bars having higher than expected retentions scores.

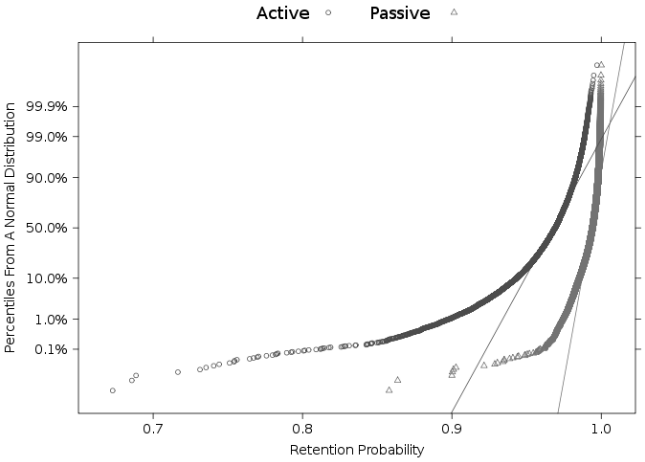

Figure 3: Distribution of Retention Probability per airing for Active and Passive STBs. The QQ plot shows "Active" viewers have a lower retention probability and a longer distributional tail.

In this section we show that STBs with active behavior before an ad are more likely to tune away from that ad. The analysis in this section serves as a motivation for the model features that we use later in the paper.

Understanding user behavior on the web helps improve quality of ads and search results. For example, Sculley et al. [7] looked at the amount of time people spend on the landing page of an ad to define the bounce rate. Agichtein et al [3] used click behavior to improve search results.

In this paper, we explore the effect of taking into account STB behavior in the time leading up to the airing of an ad. A fact that we have known for a long time is that most TV viewers are passive. Typically, only 1-3% of the viewers present at the beginning of an ad tune out before the ad ends. We also know that women and children are less likely to tune out. We suspect that people may be in different moods at different times – they are sometimes in a "clicky" mood, surfing for something to watch, or they could be in a "sticky" mood, watching passively.

Figure 3 divides users into two groups. The "Passive" set are STBs that have not tuned into a different channel for an hour before a given ad. In the "Active" set are STBs that have tuned out at least once. We define *retention probability* as the fraction of viewers at the beginning of the ad who viewed the whole ad. IAR, given by (1), is a particular type of retention probability and is defined at the level of an airing. Retention probability is a broader term which allows us to investigate different subpopulations of interest. Figure 3 shows for each airing the retention probability for "Active" and "Passive" viewers.

Now that we know that active and passive viewers have different behavior, the next question is to evaluate the variability in tune-out that we observe at the airing level. The difference in tune-out between passive and active STBs can be the difference between 20% and 80%. Figure 4 shows the distribution of those percentages in prime time and overnight. The distributions are quite different.

Figure 5 shows the retention probability of STB as a function of the number of tune out events in the hour before the ad. The different lines show probabilities for 7 of our top networks. The distributions are truncated to twenty events in the previous hour before the ad aired. Typical retention probabilities for an airing are in the range from 0.97 to 0.99 (see Figure 6, 90% of the airings are in the range $(0.969, 0.99)$). Note the strong downward trend in retention probabilities as STB activity increases.

4. RESULTS

In this section, we report experimental results showing that retention scores are able to measure some intrinsic property of the ads. We also show that the models we use to generate retention scores are stable and the retention scores themselves are stable. In the last subsection, we compare the two models used to compute retention scores.

To compute retention scores, we have to estimate the Expected IAR (as defined in equation (2)). We have experimented with various models for doing so. Here we discuss two of them – one that assumes that all the STB are the same and another that uses STB-dependent features like the number of tune out events in the hour before the ad.

Model 1 is a logistic regression model which estimates parameters for day part (categorized time of the day during which the ad appeared), week day (Mon-Fri) versus week end (Sat-Sun), network (TV channel) and creative length (typically, 30 seconds, but sometimes, 15, 45 or 60 seconds).

Model 2 is also a logistic regression model, but estimates parameters for *n_number_events* (number of tune-out events the the hour before the ad), *event_last_10_min* (a categorical variable with three values: whether there was a tune-out in the last 10 minutes before and ad, the last one minute or neither), network and creative length. The two additional tune-out events attempt to capture the behavior of the user

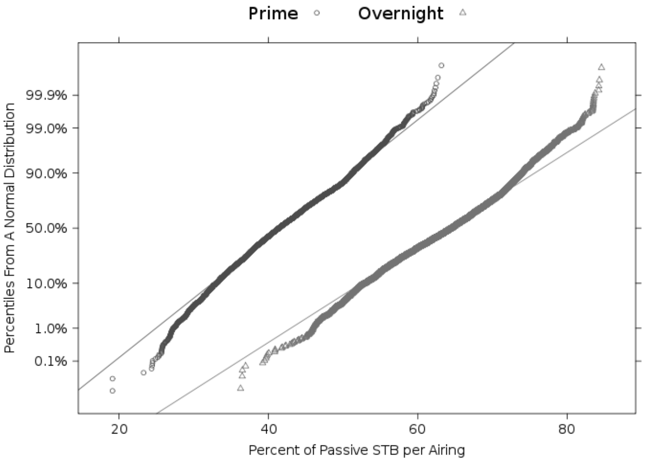

Figure 4: Distribution of Passive STB during Prime Time and Overnight. The mean percentage of "Passive" viewers for Overnight is greater.

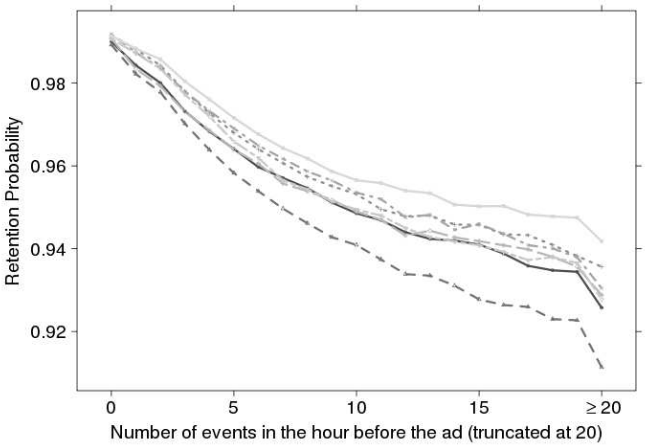

Figure 5: Retention Probability as a function of the number of events in the hour before the ad for 7 of the top networks. The distribution is truncated at 20 events in the previous hour. STBs with more than 20 events are averaged into the last group. Each line represents a different network.

prior to watching an ad.

4.1 Model Stability

In this section we investigate the stability of the retention score as computed by Model 1, which does not incorporate user behavior, and Model 2, which does. For both models we estimated the retention score for a sample ads in March 2009 and compared the scores to the same ads in April 2009. The ads selected for this test had to meet a minimum requirement of 20 airings for each month. The results of the comparison is shown in Figure 7. The variance of the difference between identical ads from the two months is reduced from 0.015 from Model 1 to 0.011 to Model 2, a 27% decrease.

4.2 Predictive Tests Of Retention Scores

If retention scores are measuring some intrinsic property of the ads, then it should be possible to predict future audience behavior based on them. To test this, we selected pairs of "good" and "bad" ads and then ran these back-to-back on seven different TV networks.

Our first experiment during several days in December 2008 and January 2009, for a total of 66 distinct airings[4]. For each pair of airings, the non-creative factors (e.g., time of day, day of week, network, etc.) were held essentially constant. We also alternated the order of the "good" and "bad" ads to neutralize any position bias. We expected ads with positive retention scores to retain more audience than ads with negative retention scores.

Figure 8 shows the results for these 66 airings. The y-axis gives the IAR for the "good" ad, while the X axis gives the IAR for the "bad" ad. Points above the diagonal line are those in which the "good" ad retained more audience. This was the case for all 66 airings, demonstrating that retention scores calculated from our model residuals are strong predictors of future ad performance. For ads selected randomly without any distinction of "good" or "bad," we would have expected the points in Figure 8 to fall equally above and below the diagonal line.

These tests provide strong evidence that our statistical models are able to isolate the impact of creatives on audience behavior, despite the significant noise introduced by non-creative factors.

We re-analyzed recent ad airings to see how well our retention scores for creatives were predicting future audience behavior (see Figure 9). We looked at every pair of ads aired in the same pod sometime in April and compared whether the ad with a higher retention score (based on model data from February and March) actually retained more audience then the other in that airing. We then plotted this probability against the difference in retention score to see how big a difference in retention score is needed between two ads to successfully predict their relative performance. We found that ads with a 0.6 (60%) difference is retention score generally performed in the expected order over 90%, while ads with a 0.1 (10%) performed in the expected order only slightly more than 50% (i.e., chance).

The squares in Figure 9 are from live experiments similar to the one described in Figure 8. Each experiment varied in creative pairs as well as how close they were in terms of retention score. The agreement between the experiments and the empirical curve suggests that virtually all differences in retention score are predictive, and that larger differences allow for stronger predictions.

4.3 Comparing Models

The main measure we care about is whether this new model could help us better predict the difference in between the good and bad creatives. As we mentioned before, we define a creative to be "good" and "bad" according to Table 1. We used one month of data on 25 major networks to construct the models. As shown in Table 3, with Model 2, we were able to label 35% of the creatives as either good or bad while with Model 1 we predict 24%. The first model

[4] The networks used were ABC Family, Bravo, Fine Living, Food Network, Home & Garden Television, The Learning Channel, and VH-1.

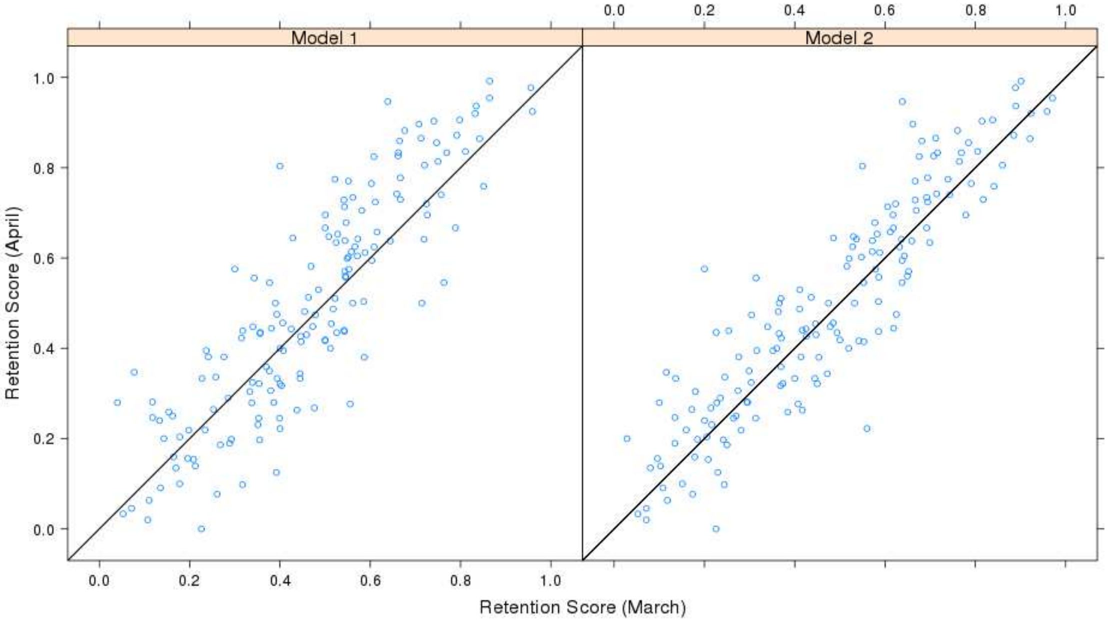

Figure 7: Model stability for Model 1 (left) and Model 2 (right). Each point on the plot represents the retention score for a single ad determined for April 2009 versus the same ad in March 2009. The results for Model 2, which includes user behavior, have less variability than Model 1, indicating that computations of retention scores with Model 2 are more stable.

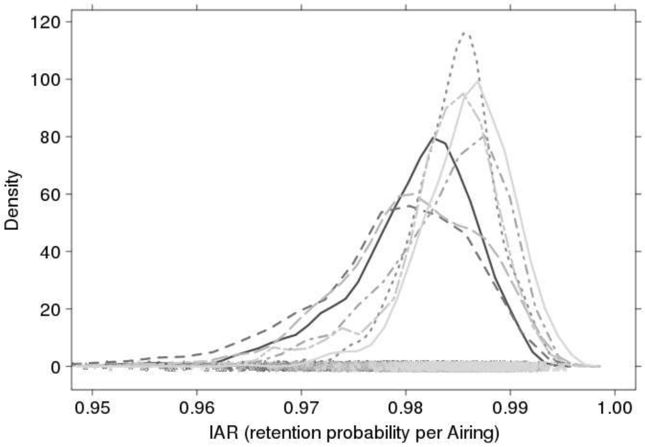

Figure 6: Distribution of IAR or Retention Probability per airing for 7 of the top networks. (0.5% of the airings have IAR below 0.95). Each line represents a different network.

predicted 10% of the creative to be good and 14% to be bad; the second model predicted 20% to be good and 15% to be bad. The last row in Table 3 shows percent of the creative that were predicted by both models. As the results show most of the creatives (> 80%) predicted by Model 1 were also predicted by Model 2.

Both Model 1 and Model 2 are fit using logistic regression, but are based on different features. Since the same airings are used in both models, we compare the two the fits using their estimated dispersion numbers. The dispersion number is given by

	%Good Creative	%Bad Creative
Model 1	14%	10%
Model 2	20%	15%
Both Models	12%	9%

Table 3: Comparing models in predictive power. Model 1 identifies 14% of the creatives as good and 10% as bad. The two models overlap significantly with both models labeling 12% of the same ads as good and 9% of the ads as bad. This table implies that Model 2 is able to make better distinctions between good and bad.

$$\sigma^2 = \frac{1}{N-p}\sum_{i=1}^{N}\frac{(y_i - n_i\widehat{y}_i)^2}{n_i\widehat{y}_i(1-\widehat{y}_i)}, \tag{5}$$

where N is the sample size, p is the number of parameters estimated in the model, y_i is the observed count, n_i is the number of binomial trials, and $\widehat{y}_i$ is the expected proportion from the logistic regression model [5].

The dispersion number can be thought of as a ratio of the variance between the observed binomial trials and the variance explained by the logisitic regression model. Hence, a value of one for (5) is optimal, while values greater than one (or overdispersion) suggest our data are noisier than expected under a binomial assumption. In the context of Model 1 and Model 2, a smaller dispersion number helps us identify which of the two models explains more of the

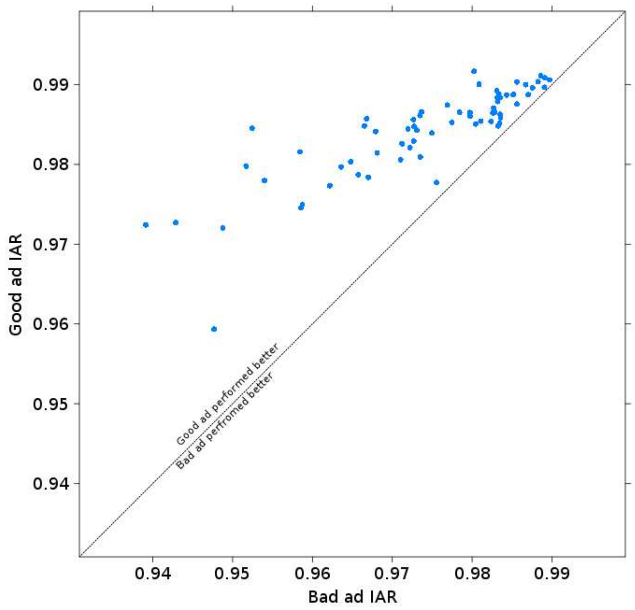

Figure 8: Results of an experiment which placed a "good" ad and "bad" ad side by side. Each point represents the IAR of the good ad (y-axis) vs. the IAR of the bad ad (x-axis). There were 66 airings from seven different networks.

variance. We get a dispersion of 83.7 for Model 1 and 3.4 for Model 2, which implies Model 2 accounts for 25 times more variability than Model 1.

5. CONCLUSIONS AND FUTURE WORK

Many factors influence the tuning behavior of TV audiences, making it difficult to understand the precise impact of a specific ad. However, by analyzing the tuning of millions of individuals across many thousands of ads, we can model these other factors and yield an estimate of the tuning attributable to a specific creative and confirm that creatives themselves do influence audience viewing behavior. This retention score – the deviation from the expected behavior – can be used to rank ads by their appeal, and perhaps relevance, to viewers, and could ultimately allow us to target advertising to a receptive audience much more precisely.

We hope these methods will inspire and encourage more relevant advertising on television. Advertisers can use retention scores to evaluate how campaigns are resonating with customers. Networks and other programmers can use these same scores to inform ad placement and pricing. Most importantly, viewers can continue voting their ad preferences with ordinary remote controls – and using these techniques, we can finally count their votes and use the results to create a more rewarding viewing experience.

We hope to extend our work presented here by adding additional features to the models. One such aspect relates to ad retention close to the beginning, middle, and end of the hour. We have empirical evidence showing viewers dropping off during these periods. Another concept is audience fatigue, which suggests that viewers' retention probability decreases as the frequency of an ad increases. This be especially true for "bad" ads.

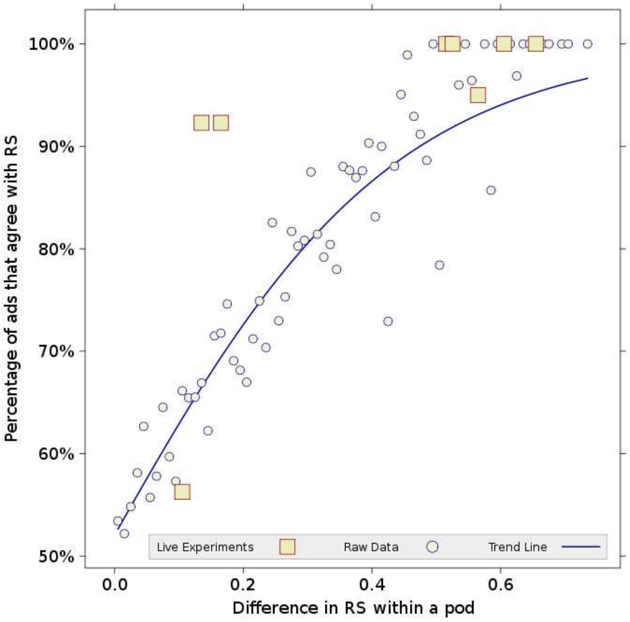

Figure 9: Comparison of ads within a pod by agreement in IAR and retension score. Each point (circle) represents the average number of times the IAR between two ads matched their respective retension scores. Live experiments (squares) run at the end of 2008 and beginning of 2009 match the curve but tend to be above our empirical results.

6. ACKNOWLEDGMENTS

We would like to thank Dish Network for providing the raw data that made this work possible, and in particular Steve Lanning, Vice President for Analytics, for his helpful feedback and support. We would like to also thank Ignacio Mosqueria for his suggestion of looking at the behavior of users locally around the ad. We are also grateful to Daryl Pregibon and Roberto Bayardo for their helpful comments on the paper. And finally, we would like to thank the ADKDD '09 conference chairs for accepting this paper and the reviewers, whose comments and suggestions are greatly appreciated.

7. REFERENCES

[1] How is my keyword's quality score used? `http://adwords.google.com/support/bin/answer.py?hl=en\&answer=49174`, 2009.

[2] Inside adwords: Quality score improvements. `http://adwords.blogspot.com/2008/08/quality-score-improvements.html`, 2009.

[3] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA, 2006. ACM.

[4] Nielsen Inc. Nielsen says bud light lime and godaddy.com are most-viewed ads during super bowl xliii. `http://www.nielsenmedia.com/nc/portal/site/Public/menuitem.55dc65b4a7d5adff3f65936147a062a0/?vgnextoid=ac08bca0e985f110VgnVCM100000ac0a260aRCR`, 2009.

[5] McCullagh P. and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1989.

[6] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 521–530, New York, NY, USA, 2007. ACM.

[7] D. Sculley, Robert Malkin, Sugato Basu, and Roberto J. Bayardo. Predicting bounce rates in sponsored search advertisements. In *KDD, to appear*, 2009.