Finding the Sweet Spot: Ad Scheduling on Streaming Media

Prashant Rajaram[1]
Puneet Manchanda
Eric Schwartz

January 2019

This Version: December 2019

[1] Rajaram (*prajaram@umich.edu*) is a doctoral candidate, Manchanda (*pmanchan@umich.edu*) is Isadore and Leon Winkelman Professor and Professor of Marketing and Schwartz (*ericmsch@umich.edu)* is Assistant Professor of Marketing, all at the Stephen M. Ross School of Business, University of Michigan, Ann Arbor, MI 48109, USA. The authors would like to thank Mimansa Bairathi, Min Kim, Adam Smith, the Marketing faculty and doctoral students at the Ross School of Business, (seminar) participants at University of British Columbia, University College London, University of Rochester, University of Zurich, the 2019 Conference on AI, Machine Learning & Business Analytics, the 2019 Marketing Science Conference, the 2019 Haring Symposium, the 2018 MIDAS Symposium and the 2018 AAAI Workshop for their valuable comments and feedback.

Electronic copy available at: https://ssrn.com/abstract=3510785
Electronic copy available at: https://ssrn.com/abstract=3498039

Abstract

A majority of US households view content on online video streaming services, consuming on demand. Not surprisingly, ad spending on such services is growing rapidly. We develop a three-stage approach to deliver an optimal ad schedule that balances the interests of the viewer (content consumption) with that of the streaming platform (ad exposure). In the first stage, we use theoretical findings to develop two parsimonious metrics – Bingeability and Ad Tolerance – to capture the interplay between content consumption and ad exposure. Bingeability represents the number of completely viewed unique episodes of a show while Ad Tolerance represents the willingness of a viewer to continue watching after ad exposure. The second stage uses detailed data on viewing activity and ad delivery to predict these metrics for a viewing session using causal machine learning methods. This is achieved via tree-based algorithms combined with instrumental variables to accommodate the non-randomness in ad delivery. In the third stage, we use the predicted metrics as inputs to a novel constrained optimization procedure that provides the optimal ad schedule.

Keywords: *Advertising Scheduling, Streaming Media, Binge-Watching, Machine Learning, Optimization*

Electronic copy available at: https://ssrn.com/abstract=3510785
Electronic copy available at: https://ssrn.com/abstract=3498039

INTRODUCTION

Streaming video content is becoming increasingly popular. 55% of US households subscribed to at least one video streaming service in 2018, up from 10% in 2009 (Deloitte, 2018). In contrast to linear TV, on-demand streaming services give viewers agency, allowing them to consume content in a self-directed manner. As a result, viewers consume media content in a "non-linear" manner by not adhering to any set temporal schedules. For example, a common behavior viewers exhibit in such settings is that of rapid consumption of multiple episodes of a TV show, usually referred to as "binge-watching" (Cakebread, 2017; Oxford Dictionary 2018). The presence of consumer "eyeballs" on streaming media represents an attractive opportunity for advertisers, especially as these services allow for ad personalization due to the availability of rich data. As a result, advertising spending on streaming media services is expected to grow to $20 billion in 2020 from $4.7 billion in 2017 (eMarketer, 2018).[2] However, streaming media represents new challenges, especially as interruptions to the viewing experience via advertising detract from the viewers' feeling of being in control and can lead to decreased content consumption (Schweidel & Moe, 2016). In addition, platforms that provide these services need to balance the viewers' control of the consumption experience while delivering advertising commensurate with advertiser objectives. Advertiser objectives entail delivering a fixed number of ad exposures over a set of TV shows or movies within a given time frame (Johnson, 2019).[3] In general, there is little work that focuses on the interplay of (consumer directed) content consumption and ad exposures. While extant research in marketing has developed recommendations for ad scheduling, e.g., Dubé et al. (2005), the viewer does not have significant control in the settings considered. In addition, the focus of past ad scheduling work has been on several ad-related outcomes but not on studying *content consumption*.[4] There is limited research that has focused on content consumption patterns in settings where viewers have control, e.g., Schweidel and Moe (2016), which does not address the ad scheduling issue.

In this paper, we propose a comprehensive approach that best combines the interests of the viewer and that of the free ad-supported platform. Specifically, we use actual viewership data from a streaming media platform to propose ad schedules that maximize advertising exposure without compromising the content consumption experience for individual viewers. In order to do this, we need to surmount a few

[2] Streaming media providers monetize their services through three distinct mechanisms (including offering combinations of these): subscriptions, advertising and product sales (e.g., sale of a movie). It is hard to assess which is the dominant mechanism. However, the number of ad-supported platforms (with or without a free service) is growing rapidly with providers such as Hulu, CBS, Dailymotion, Ora TV, YouTube, Sony (Crackle), The Roku Channel, TubiTV, Popcornflix, Amazon (IMDb TV) and NBC (Armental, 2019; Patel, 2018; Sherman, 2019). There is also industry research suggesting that consumers prefer a platform's lower cost ad-supported streaming service to its premium no-ad version, when both options are offered (Liyakasa, 2018; Sommerlad, 2018). In this paper, we focus on free streaming services with an ad supported mechanism.

[3] Our focus is on the platform's ad scheduling problem. We do not know how the advertiser arrives at exposure targets (quantity, ad location within show, customer segment etc.) specified to the platform. We also do not have access to all the downstream data e.g., browsing, purchasing etc.

[4] Recent work on ad scheduling has focused on maximizing ad-related outcomes such as profits from sales (Dubé et al., 2005), campaign reach (Danaher et al., 2010), purchase (Sahni, 2015), site visits (Chae et al., 2018) or ad viewing completion rates (Krishnan & Sitaraman, 2013).

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

challenges. First, the control that viewers have can manifest itself in multiple and diverse behaviors, both in relationship to content consumption and the reaction to advertising. However, there is little standardization on how consumer behavior on streaming media can be captured and described. Second, there is plethora of content on streaming media platforms, varying in terms of genre, show type and show duration (episode length, number of episodes per season and number of seasons). It becomes very important therefore to capture the impact of these variables and their interactions in a tractable manner. Third, in real settings, platforms do not deliver advertising randomly. Thus, any approach that is proposed needs to address the non-random delivery of such advertising. Finally, in order for ad scheduling recommendations to have practical value, simplicity and speed are very important.

We address these challenges using a three-stage approach (Figure 1). Given the lack of standardization around the measurement of content consumption and ad exposure in streaming media settings, we begin by using theory from consumer psychology to develop summary measures or metrics that capture viewer's control over the consumption experience in streaming media settings. These metrics are deterministic transforms of the data primitives (minutes watched, ads see, etc.) that are available to streaming media platforms. The two aspects of viewer behavior that we are interested in are non-linear content consumption and the response to advertising exposure. In order to do this, we first need to specify a temporal unit of consumption for a given viewer. We denote this unit as a viewer-session (in future, we use the term "session" to denote this unit) which is defined as a period of time spent by a viewer watching one TV show separated by 60 minutes or more of inactivity as in Schweidel and Moe (2016).

"Insert Figure 1 about here"

The first metric, which we label "Bingeability," is based on the theory of "flow" (Ghani & Deshpande, 1994; Schweidel & Moe, 2016) as well as industry norms and captures the extent of viewer immersion in the content. In essence, this metric is based on a stylized count of complete and unique episodes of a TV show watched in a session. The second metric, which we label "Ad Tolerance," is based on the theory of hedonic adaptation (Frederick & Loewenstein, 1999; Nelson et al., 2009) and captures the viewer's reaction to advertising. Specifically, the metric captures the willingness of a viewer to watch ads and to watch content after being exposed to ads in a session. We explain the theoretical motivation, construction and validation for both metrics in more detail in the section "Stage I".

In the second stage, we construct a model to predict the value of the above metrics for a session using an extensive set of current and historic descriptors, both specific to the content and to the viewer. We use the process of "feature generation" to generate the entire set of descriptors (cf. Yoganarasimhan

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

(2019)). In order to deal with the large number of descriptors (in the thousands), we use a tree-based machine learning method (Extreme Gradient Boosting or XGBoost) that is known to capture non-linear relationships well (Chen & Guestrin, 2016; Rafieian & Yoganarasimhan, 2019). As noted above, the viewer behavior captured in our data is a function of the delivered advertising schedule. We therefore control for the non-randomness in ad delivery to a focal viewer using "instrumental variables" based on ad delivery patterns to other viewers. After predicting these metrics for each session in holdout samples, we use feature importance analyses and partial dependence plots to shed light on the importance and nature of the non-linear relationship with various feature sets (Friedman, 2002) – this allows us to go beyond a purely black-box approach.

In the third stage, we develop our ad scheduling recommendation. We begin by passing the predictions obtained from the previous stage through an "Ad Decision Tree" that helps identify sessions where ad exposure enhances, or at least does not detract from, content consumption. For these sessions, we apply a novel constrained optimization procedure built around our predictions to provide an "optimal" advertising schedule for the platform that maximizes ad exposure, subject to (predicted) Ad Tolerance.

We calibrate our approach on a novel data set that captures the viewing behavior of individuals on Hulu (when it had only a free ad-supported streaming service). We find that our proposed metrics, Bingeability and Ad Tolerance, perform well in terms of capturing viewer behavior with respect to content consumption and ad exposure. We also find strong evidence of state-dependence for these two metrics. In other words, TV shows that have a high Bingeability for a viewer in the past (week) result in a high Bingeability for the current session. Similarly, past Ad Tolerance is predictive of current Ad Tolerance. We also find that variations in ad spacing and ad exposure have a non-linear effect on content consumption. Based on these findings, our optimization module provides individual session level recommendations vis-à-vis pod (a block of ads) frequency and spacing. For example, we suggest that, on average, Hulu should decrease pod frequency (increase pod spacing) when a viewer is expected to have lower Ad Tolerance or higher Bingeability, holding the other constant. The optimization module can in general be used as a decision support system by the platform. Specifically, the platform can define critical thresholds of predicted Bingeability to decide to show ads and obtain the recommended ad delivery schedules to explore the inherent tradeoffs between content consumption and ad exposure for its viewers. We find that under the optimized ad schedule, the decision to show ads in *all* future sessions for existing viewers, i.e., when *predicted Bingeability is greater than 0*, benefits the platform *and* the viewer the most with content consumption increasing by 5% and ad exposure increasing by 71% (on average).

In sum, our paper makes four main contributions. First, it is one of the early papers that examines viewer behavior spanning content consumption *and* ad exposure in streaming media environments. Sec-

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

ond, using a combination of metrics, data and optimization, the paper makes explicit the tradeoffs between ad delivery and content consumption, thus balancing the interests of both parties. Third, it illustrates how the use of instrumental variables and partial dependence analyses help to address concerns around the purely predictive and black-box nature of machine learning methods. Finally, it provides a scalable and interpretable approach to ad scheduling at the individual session level.

DATA

Our data come from the streaming platform Hulu, spanning the period Feb 28, 2009 to June 29, 2009. At this time, the platform only offered a free ad-supported streaming service.[5] We have data on the viewing behavior of a random sample of over 10,000 accounts for this period. Each account could potentially be shared by household members or friends, but as all accounts were free, we do not expect account sharing to be prevalent. Hence, we assume that each account represents a unique viewer. In addition, during this period, viewers could only access Hulu via a browser as the mobile and tablet app was not launched until 2011 (Ogasawara, 2011). Thus, we are able to capture all Hulu viewing behavior for an account.

We restrict our data to TV show viewing behavior and not movie watching behavior for two reasons. First, TV show viewing behavior has more potential to build engagement with the platform because content length of a TV show (for all episodes) is typically longer than that of a movie. Second, given multiple episodes, TV shows lend themselves more to non-linear consumption (Deloitte, 2018). TV shows make up 55.5% of total titles[6] in the dataset, with the remaining being movies. Among the viewers who watch TV shows, we further select only those viewers who visit the platform at least twice to watch TV shows during our sample period to ensure that we can include viewer fixed effects in our model. Screening on this leaves us with a sample of 6,228 viewers who watch 568 TV shows spanning 18 genres.

Sessions

A 'session' (or sitting) is defined as time spent by a viewer watching show content or ads from exactly one TV show separated by 60 minutes or more of inactivity (Schweidel & Moe, 2016).[7] A session can be split into the following parts:

[5] Hulu offered an additional subscription plan with limited ads in 2010, and an additional premium plan with no ads in 2015 and phased out its free plan in 2016 (Ramachandran & Seetharaman 2016). However, as noted earlier, multiple streaming services such as YouTube, Dailymotion, Ora TV, The Roku Channel, TubiTV, Crackle, Popcornflix and IMDb TV continue to offer free ad-supported streaming plans.

[6] A title is classified as a movie if there is only one video (episode) for that title *and* the duration of this video is greater than 60 minutes. For the few cases where a TV show and a movie share a name (this typically occurs when one is a spin-off of the other), we classify the movie as a TV show. Note that our results are invariant to the inclusion or exclusion of these movies.

[7] As noted earlier, we need to define a viewing session in order to summarize/predict viewer behavior and decide on ad delivery. Note that our approach is general as it can be applied to *any* time separation used in the definition of a session. If no time unit is defined, then a continuous time model of content consumption and ad delivery needs to be specified along with continuous time ad scheduling recommendations. We believe that such a model is likely to be intractable, if not infeasible.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

$$(1)\quad \overbrace{Session\ Time}^{\text{Measured}} = \overbrace{Content\ Time + Ad\ Time + Filler\ Content\ Time}^{\text{Measured}} + \overbrace{Pauses\ - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

where *Session Time* represents the calendar time spent in the session, *Content Time* is time spent viewing show content (including minutes of content skipped in fast-forwards but excluding minutes of content seen again in rewinds), *Ad Time* is time of ad exposure, and *Filler Content Time* is time spent viewing filler content which are interjected between the main episodes. We classify all episodes less than 15 minutes e.g., short videos such as interviews, recaps, previews, trailers etc., as filler content. It is important to note that ads cannot be fast-forwarded, rewound or skipped unlike show content or filler content. All the previously mentioned variables are measured in our panel data. In addition, there are unmeasured variables that complete the above equation—*Pauses* is the time spent in a break, *Fast Forward* is the duration of content fast-forwarded, and *Rewind* is the duration of content rewound. A statistical summary of the sessions is shown in Table 1. The 2.5th to 97.5th percentile of the time spent in a session ranges from 1.82 minutes to 236.51 minutes (about 4 hours) with a median time spent of 42.70 minutes.

"Insert Table 1 about here"

In Table 2, we show a representative example of typical viewing behavior in a session. More examples are detailed in Web Appendix A. In the first row of the example, 'light gray shaded boxes' denote *Ad Time*, 'white shaded boxes' denote *Content Time* and the 'dark gray shaded box' denotes *Filler Content Time*. In the second row of each example, the 'white shaded dashed line boxes' denote *Session Time*, and the 'black shaded boxes' indicate the beginning of the next episode. All values are in minutes. A block is a period of time from the beginning of a pod (or beginning of session) till the beginning of the next pod (or end of episode/session).

"Insert Table 2 about here"

The example shows the behavior of a viewer watching two 24-minute episodes of 'Aquarion'. The viewer's viewing experience was interrupted by 5 ads (light gray shaded box) and 2 minutes of filler content (dark gray shaded box). The black shaded box denotes the beginning of the next episode. We can see evidence of fast-forwarding behavior in block 5 because the session time of 17.66 minutes is less than the sum of ad time (0.66 min) and content time (21 mins). There is evidence of pauses in block 6 because the session time of 1 min is greater than the ad time of 0.5 min. There is no evidence of rewinds in block 6

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

because no content was viewed and ads cannot be rewound. By substituting the values of the example in equation (1), we get,

$$\overbrace{Session\ Time}^{44.82\ \text{minutes}} = \overbrace{Content\ Time}^{43\ \text{minutes}} + \overbrace{Ad\ Time}^{2.83\ \text{minutes}} + \overbrace{Filler\ Content\ Time}^{2\ \text{minutes}} + \overbrace{Pauses - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

On solving the above equation, we find that the sum of the unmeasured variables is -3.01 minutes. This indicates that more time was spent in fast-forwards than in pauses or rewinds in this session.

Ad Delivery

It is important to understand what the platform was doing in terms of ad delivery at the time of our data. As we do not have access to institutional practices at Hulu, we examine the realized data patterns to infer the rules governing ad delivery (technical details on how Hulu collected the viewing data are described in Web Appendix B). We focus on four aspects of ad delivery – the duration of pod (block of ads) exposure, frequency of ad delivery (by examining mean spacing between pods), diversity of ad exposure (based on the industry of the advertiser) and degree of non-conformity to equal spacing (by examining "clumpiness" of pod exposure).

We first plot the distribution of the length of commercial pods, conditional on non-zero seconds of ad exposure, viewed across all sessions in Figure 2a. The median length of time viewed is 30 seconds with range of 6.6 to 55.2 seconds (2.5^{th} to 97.5^{th} percentile). Note that more than 99% of the pods have only one ad. A viewer may not watch a pod completely if she ends the session before the pod ends or refreshes the browser or skips the episode. Hence, the amount of pod duration viewed is less than or equal to the pod length. As the figure shows, the most common pod durations (lengths) are 15 seconds and 30 seconds. In other words, pod durations follow a non-uniform distribution which indicates that Hulu uses a set of rules to set duration – we label this as Hulu's "Length Rule." Next, we plot the density of the spacing (content time viewed including filler content) between pods in an episode across all sessions (Figure 2b). We find that this spacing is also not uniformly distributed. The peak at 0 minutes corresponds to pre-roll ads (ads at the beginning of an episode), and there is also a peak at 6.3 minutes. This bi-modal distribution indicates non-random spacing and we label it as Hulu's "Spacing Rule."

"Insert Figure 2a,2b,2c and 2d about here"

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

We then examine whether there is any systematic pattern in ad delivery across the type of advertiser and length of ad. Using the empirical distribution of ad lengths, we classify all ads into three types: 0-26 seconds, 26-52 seconds, and > 52 seconds. Each ad in our dataset belongs to one of 16 product categories such as CPG, Telecom, etc., resulting in 48 (= 3 X 16) unique combinations.[8] Figure 2c shows the distribution of the percentage of diverse ads viewed in an episode across all sessions. It is not perfectly uniform, suggesting that certain advertisers have preferences over TV shows that they want to show ads on – we label this as Hulu's "Diversity Rule." Finally, we examine the degree to which pods are not equally spaced in an episode using the measure of clumpiness proposed by Zhang et al. (2014) as below:

$$1 + \sum_{i=1}^{n} \frac{\left(\frac{x_i + 0.01}{N + 0.01}\right) \log\left(\frac{x_i + 0.01}{N + 0.01}\right)}{\log(n+1)} \tag{2}$$

where n is number of pods in an episode, x_i is content time viewed till pod i, N is total content time viewed in the episode till the last pod. We add 0.01 to avoid errors because of $\log(0)$ and division by 0. Figure 2d shows the distribution of clumpiness of pods in an episode across all sessions. It is not perfectly uniform – we label this Hulu's "Clumpiness Rule."

The above suggests that Hulu's ad delivery exhibited specific patterns i.e., ads were not delivered randomly. In subsequent analysis, we summarize this non-randomness using the four dimensions above via the Length Rule (LR), Spacing Rule (SR), Diversity Rule (DR) and Clumpiness Rule (CR). We then account for it using instrumental variables (see section "Stage II").[9]

STAGE I

Our research comprises of three stages as illustrated in Figure 1. In the first stage, we construct parsimonious metrics to capture and summarize viewer's control of the consumption experience, thus allowing us to systematically track viewer behavior over time.

Metric Development

Bingeability. As noted earlier, the first metric we develop captures the extent of viewer immersion in the content, potentially leading to non-linear consumption. Immersion in the viewing experience can be likened to experiencing a "flow" state characterized by a combination of focused concentration, intrinsic

[8] Less than 0.15% of total ads are "ad selectors" where a viewer can choose an ad to view from a few options. Hence, we do not use an additional rule to differentiate between "ad selectors" and "non ad-selectors".

[9] While it is possible that there are other forms of non-randomness in ad delivery, our analysis suggested that these four aspects accounted for most of the variation in ad delivery. Any aspect of advertising that is systematic, including relating ad delivery to the story arc (e.g., delivering more ads before a cliffhanger ending), is captured in the large number of fixed effects (viewer, show, genre etc.) we use as features in Stage II.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

enjoyment and time distortion (Ghani & Deshpande, 1994; Schweidel & Moe, 2016). Thus, the stronger the flow state that the viewer is in, the more episodes she is likely to consume within a session. Our metric therefore takes the common industry metric of the raw number of episodes watched in a session (West, 2013) and adjusts it for all activities that indicate that the viewer has fallen out of the flow state e.g., skipping and/or fast-forwarding.[10] In other words, this metric represents the count of episodes in which viewers are immersed in the viewing experience by using the count of "complete unique episodes" watched in a session. Specifically, "complete" refers to episodes that are watched in full i.e., no content is missed, while "unique" refers to the number of distinct episodes watched in the session.[11]

In effect, our metric represents the count of episodes (which are positive integer values) that characterize binge-watching behavior, and hence we name it "Bingeability." It is important to note that we do not define binge-watching, but instead qualify the kind of episodes which should be counted in the industry definition of binge-watching. For example, Netflix conducted a poll and found that its viewers perceive watching 2 to 6 episodes of a TV show in one sitting as binge-watching (West, 2013). We argue that such a count should not be a raw episode count of 2 to 6 episodes but a count that includes only the number of complete and unique episodes watched. Thus, Bingeability is more conservative than a raw episode count, and the product of Bingeability and average episode length is more conservative than a measure of content minutes watched. In order to show that the proposed metric is not identical to a simple count of episodes, we discuss its information content and validity in the subsection on "Metric Validity."

The Bingeability metric is defined as

$$(3) \qquad Bingeability = \sum_{i=1}^{n_e} \mathbb{1} \begin{Bmatrix} Content\ Length_i - 5\ mins \leq\ Content\ Time_i \leq \\ Session\ Time_i - Ad\ Time_i \end{Bmatrix}$$

where, $\mathbb{1}$ is an indicator function, i denotes a unique episode, n_e is the number of unique episodes watched, $Content\ Time_i$ is the time spent watching content for episode i, $Content\ Length_i$ is the length of episode i including opening and end credits, 5 mins is an upper bound on the combined duration of opening and end credits in an episode, $Session\ Time_i$ is the calendar time spent and $Ad\ Time_i$ is the time of ad exposure. The presence of the indicator function ensures that the metric is integer valued. We explain the two conditions in the indicator function below.

[10] Recent academic work e.g., Ameri et al. (2019) and Lu et al. (2017) also study binge-watching of content without looking at the ad scheduling issue. Both studies customize their binge-watching definitions to their idiosyncratic settings – an anime website for the former and Coursera for the latter. In contrast, our objective is to develop a measure of non-linear consumption that can be used by the platform for its decision making, not to define binge-watching.

[11] We ignore repeat viewing behavior (same episode, same Viewer ID, same session) as it is present in only 0.6% of observations.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

i. No skipping:

$$(3a) \qquad Content\ Length_i - 5\ \text{min} \leq Content\ Time_i$$

Skipping means moving ahead to the next episode or ending the session without completely watching the present episode. Skipping content (excluding credits) is indicative of a break in the immersive experience or the 'flow' state of a viewer. Hence, we exclude episodes displaying skipping behavior from the count of Bingeability.

The sum of opening and end credits for TV shows are typically less than 5 minutes which can be considered a lenient upper bound (ABC, 2014; Ingram, 2016). This is subtracted from $Content\ Length_i$ as viewers are less likely to watch credits when they are binge-watching the show (Miller, 2017; Nededog, 2017). After subtracting the maximum possible time involved in opening and end credits, 5 mins, from $Content\ Length_i$, if the difference remains less than or equal to $Content\ Time_i$, then we can conclude that the viewer has not skipped watching content.

ii. No excessive fast-forwarding:

$$(3b) \qquad Content\ Time_i \leq Session\ Time_i - Ad\ Time_i$$

Fast-forwarding means moving ahead faster than normal pace to view future content from the same episode. There may be occasions when the viewer chooses to excessively fast-forward certain portions of an episode. This would result in a greater increase in $Content\ Time_i$ than a difference of $Session\ Time_i$ and $Ad\ Time_i$. Excessive fast-forwarding is indicative of a break in the 'flow' state of a viewer. Hence, we avoid counting episodes in which a viewer carries out excessive fast-forwards. Substituting equation (1) in equation (3b), we can rewrite equation (3b) as follows:

$$(3c) \qquad Fast\ Forward_i \leq Filler\ Content\ Time_i\ +\ Rewind_i +\ Pauses_i$$

The above equation ensures that the amount of time spent in fast-forwards is less than the sum of the time spent watching filler content, in rewinding content and in pauses.[12]

Next, we apply the Bingeability metric to the illustrative example discussed earlier in Table 2, and this computation is shown in Table 3. In this example, the value of content length for each episode is 24 minutes. Time spent watching content in Episode 1 is [$\mathbf{10 + 10 + 2 = 22}$] minutes and in Episode 2 is

[12] We allow viewers to fast-forward filler content because viewers are less likely to be interested in viewing content that has been inserted into their viewing experience by the streaming platform. We also allow viewers to fast-forward content that has been rewound e.g., when a viewer wishes to rewind and go back to a certain section of the episode to get more clarity, and having re-watched that section, now fast-forwards ahead to the point from where the rewind had begun. Such an action need not imply a break in the flow state of a viewer, and hence we do not penalize such behavior. We are also forced to allow the minutes of content fast-forwarded to be less than the time spent in pauses. As the time spent in pauses is an unmeasured variable, we are unable to eliminate all occasions of fast-forwarding behavior. Theoretically, we end up allowing those occasions when a viewer takes frequent breaks but also keeps fast-forwarding content. Such behavioral patterns are unlikely but we cannot rule them out. Hence, we only eliminate occasions of "excessive" fast-forwarding as originally stated in the condition.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

21 minutes. There is no evidence of skipping behavior in either Episode 1 or Episode 2 because the first condition is satisfied. The total time spent in the session for Episode 1 is [$\mathbf{10.66 + 11 + 2.5 = 24.16}$] minutes and for Episode 2 is [$\mathbf{2 + 17.66 + 1 = 20.66}$] minutes. The total ad time for Episode 1 is 1.66 min and for Episode 2 is 1.16 min. We find evidence of excessive fast-forwarding in Episode 2 because the second condition is not satisfied. Thus, the value of Bingeability is 1 as our metric only counts Episode 1 which was viewed completely. The fast-forwarding behavior within Episode 2 (in block 5 – see earlier subsection "Sessions") represents incomplete viewing and hence disqualifies the episode from being used in the Bingeability count.

"Insert Table 3 about here"

Ad Tolerance. Previous research has shown that interruptive stimuli e.g., ads, can influence the enjoyment level of a viewer while watching video content on certain occasions. If the viewer is watching content and adapting to that hedonic experience, then an ad interruption breaks the adaptation pattern preventing enjoyment levels from falling (Nelson et al., 2009). On the other hand, if the viewer is not adapting (to content), then (ad) interruptions can break the flow state by irritating the viewer (Frederick & Loewenstein, 1999). In the first case, the viewer can be expected to watch *more* content after the ad ends. In contrast, in the second case, the viewer can be expected to watch *less* content after the ads ends. Unfortunately, we cannot measure adaptation directly but we use these results as motivation to develop a metric – Ad Tolerance – that captures the willingness of a viewer to watch ads and to watch content after being exposed to ads in a session. Note that the tacit assumption we are making is that consumers are myopic in their viewing behavior i.e., they do not base their current viewing decision on future (expected) ad exposure.[13]

Based on the above, we develop the Ad Tolerance metric by looking at three components of the viewing experience: (i) duration of pod exposure, (ii) amount of content viewed after pod exposure till the end of session and (iii) calendar time elapsed since previous pod exposure. The first component just looks at the viewer's propensity to watch ads – the longer she watches, the more ad tolerant she is. The second component focuses on the content watching behavior after pod exposure. The longer the viewer watches content after pod exposure, the higher her ad tolerance. Finally, the third component is a correction for the

[13] The only objective mechanism by which a viewer could obtain future ad exposure information is via hovering her cursor at the bottom of the video to bring up the progress bar (which fades out quickly) as this bar shows markers denoting pod locations. We checked online forums (reddit.com, slate.com, anandtech.com) and carried out online searches for keywords such as "ad location," "ad position," "future ads" and "ads coming up" for the 2008-09 period. We found a lot of discussion around viewer irritation with ad repetition and video buffering at Hulu, but none around the ability to see future ad locations. This, along with the fact that obtaining this information while viewing is costly, is supportive of our assumption regarding myopic behavior.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

time available to adapt to the content and the absence of ad exposure (described in detail below). The Ad Tolerance metric is constructed as follows:

$$Ad\ Tolerance = \sum_{j=1}^{n_p} (w_1 PodDuration_j + w_2 ContentEnd_j - w_3 (CalendarPod_j - PodDuration_{j-1})) \quad (4)$$

where j is a pod in the session and n_p is number of pods watched in the session. $PodDuration_j$ is the duration of commercial pod j, $ContentEnd_j$ is content watched (including filler content) till the end of the session after watching commercial pod j, $CalendarPod_j$ is calendar time elapsed from the beginning of the previous pod in the same session till the beginning of pod j, $PodDuration_{j-1}$ is the duration of commercial pod j-1 and w_1, w_2, w_3 are the weights associated with the three components in the equation. Initially, we set the value of each of the weights to one (and in Web Appendix C, we show that the optimization outcomes are not sensitive to these weights). Note that though the unit of Ad Tolerance is minutes and its range is the real number line, it cannot be directly interpreted as a temporal measure. Its magnitude represents the willingness of the viewer in a session to watch ads and to watch content after being exposed to ads. A negative value of Ad Tolerance suggests that the viewer stopped watching content immediately after being exposed to a pod which was preceded (at some point) by a long period of no ad exposure. We now explain the importance of each component of the Ad Tolerance metric in equation (4).

i. $PodDuration_j$: *Duration of a pod*

When a viewer is exposed to a commercial pod, each passing second of the pod contributes to the viewer's willingness to be exposed to the pod. This is captured by $PodDuration_j$, the duration of the j^{th} pod that is watched in the session. While a viewer does not have the option to fast-forward, rewind or skip ads, a viewer can partially watch a pod by exiting the session in the middle of the pod, refreshing the browser or skipping to the next episode in sequence. Hence, $PodDuration_j$ captures the willingness of the viewer to be exposed to the pod.

ii. $ContentEnd_j$: *Content time watched till end of the session*

$ContentEnd_j$ measures the time spent watching content till the end of the session after being exposed to pod j. Longer durations suggest higher tolerance for the previous interruption (with Schweidel and Moe (2016) finding empirical evidence that content viewership decreases on average as ad exposure increases). To reduce potential bias in our estimates of $ContentEnd_j$, we operationalize the measure of $ContentEnd_j$ as the minimum of (a) *Content Time* in a block and (b) the difference between *Session Time* and *Ad Time* in a block. As mentioned earlier, a block is a period of time from the beginning of a pod (or beginning of session) till the beginning of the next pod (or end of episode/session). If a viewer

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

keeps excessively fast-forwarding content, *Content Time* would increase without a corresponding increase in *Session Time* (calendar time). As a result, using *Content Time* will positively bias the measure of $ContentEnd_j$ as the viewer is not actually watching content but is only fast-forwarding content. In such situations, the metric adds option (b) which is smaller than option (a), thereby eliminating the above bias. This correction is termed as 'Caveat 1' in the rest of the paper.

iii. $CalendarPod_j - PodDuration_{j-1}$: *Inter-pod calendar time*

In our setting, when a viewer is not watching ads, she is either watching content, fast-forwarding/rewinding content or engaged in a break/pause. During this period, the viewer can be expected to simultaneously adapt to both the content and the absence of ad exposure. The third term captures this period because it is a measure of the calendar time elapsed since the previous pod exposure. This is the time during which the level of potentially unfavorable affective intensity resulting from ad exposure can go down.

$CalendarPod_j$ measures the calendar time from the beginning of the previous pod, j-1, in the same session till the beginning of pod j. For the first pod in the session, $CalendarPod_j$ measures the time from the beginning of the session as there is no previous pod watched in the session. $PodDuration_{j-1}$ is the duration of the j-1 pod that is watched in the session. The difference between $CalendarPod_j$ and $PodDuration_{j-1}$ is the measure of the ad-free time before the beginning of $PodDuration_j$. This measure of ad-free time is subtracted from $ContentEnd_j$ in equation (4) to get the net effect of the affective influence of an interruption on the viewer.

Next, we apply the Ad Tolerance metric to the illustrative example discussed earlier in Table 2, and this computation is shown in Table 4. More illustrative examples are shown in Web Appendix A. In this example, we begin by adding the duration of the first pod which is 0.66 minutes to the amount of content viewed in the remainder of the session (after the end of the pod), which is [$10 + 10 + 2 + 2 + 17 + 0 = 41$] minutes. It is important to note the use of 'Caveat 1' in block 5 (see Table 2) where there is evidence of fast-forwarding behavior. $ContentEnd_j$ is chosen as $Session\ Time - Ad\ Time$, [$17.66 - 0.66 = 17$] minutes, because it is less than $Content\ Time$ of 21 minutes. Then we subtract the difference between the time elapsed since the beginning of the session and duration of the previous pod, which are both 0 minutes in this case. Thus, the total value of the metric for the first pod is 41.66 minutes. Then, we repeat this process for the second pod. The second pod is 0.50 minutes long, to which we add the amount of content viewed in the remainder of the session which is $[10 + 2 + 2 + 17 + 0 = 31]$ minutes. Then we subtract the difference between the time elapsed since the beginning of the previous pod and duration of the previous pod, which is $[10.66 - 0.66 = 10]$ minutes. Thus, the total value of the metric for the second

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

pod is 21.5 minutes. The same process is repeated for each of the remaining pods. On summing up the values corresponding to each pod, we get a total Ad Tolerance value of 67.98 minutes.

"Insert Table 4 about here"

Data Summary via Metrics

For our sample comprising 110,500 sessions,[14] Bingeability ranges from 0 to 57 (median is 1 episode) while Ad Tolerance ranges from −412.17 to 63,449.10 minutes (median is 23.62 minutes) (see Table 5). The frequency distribution of Bingeability and Ad Tolerance in shown in Figures 3a and 3b. The most common value of Bingeability is one (complete episode) in a session. Thus, most of the sessions are not spent watching multiple episodes of the same TV show in our data. The distribution of Ad Tolerance is very right skewed. There is large peak between 0 and 3 minutes for more than 10,500 sessions. More than 16% of these sessions are those in which viewers end the session in less than a minute of calendar time. This suggests that there are many occasions when viewers are averse to seeing ads at the beginning of a session (pre-roll ads).

"Insert Table 5 about here"
"Insert Figure 3a,3b and 3c about here"

The relationship between the two metrics is shown in the jitter plots (around the values of Bingeability) in Figure 3c, where the darker areas indicate regions of high overlap. The correlation between Ad Tolerance and Bingeability is 0.68 over the full range of the two metrics and is 0.60 over the 2.5^{th} to 97.5^{th} percentile range of the two metrics. This provides some model free evidence that both metrics are complementary in terms of describing viewer behavior.

Metric Validity

Given that the two proposed metrics are deterministic transforms of the raw data, it is important for us to establish that they are valid and informative in terms of capturing viewer behavior. In the interest of brevity, we provide a summary of this analysis – full details are reported in Web Appendix D. We first compare the Bingeability metric to the commonly used industry metric for binge-watching – the raw count of episodes, typically unique, watched of the same TV show in one session (West, 2013). Unlike the raw count of episodes, the Bingeability metric considers whether viewers watch each episode completely by

[14] The Ad Tolerance metric is undefined for the 12,117 sessions where there is no ad exposure and so we exclude them.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

explicitly accounting for skipping or excessive fast-forwarding behavior. This allows for a much more precise measure of content consumption. The correlation between Bingeability and raw episode count is 0.85 over the full range and 0.70 over the 2.5th to 97.5th percentile range of Bingeability. The lack of perfect (or close to perfect) correlation suggests that the Bingeability metric captures information distinct from that in episode count. There are no comparable metrics to Ad Tolerance in practice or academic research to the best of our knowledge. In order to test the validity of our metric, we check for evidence of correlation between the metric and other "intuitive" measures of ad tolerance: number of pods shown, minutes of ad exposure and minutes of content viewed. The correlations are 0.78, 0.77 and 0.78 respectively, pointing to the fact that the metric captures distinctive information. Moreover, as shown in Web Appendix D, we find that this metric captures differences in behavioral consumption patterns better than intuitive measures. Overall, for both metrics, the distinctive information captured suggests face validity (cf. Ailawadi et al. (2003)).

STAGE II

In this stage, we use the available information to predict the viewing behavior (summarized by the two metrics) of a new session, for either a current or a new viewer watching an existing or new TV show. In the first step, we lay out the information that is used (Feature Generation) and in the second, we lay out the predictive methods (Model).

Feature Generation

The high granularity of our data allows us to include a rich set of features to help predict viewer behavior during a session. We use current and past viewing activity on Hulu to choose these features. In order to include both weekdays and weekends, we use a seven-day moving window to capture past viewing activity. The features we use fall into four types.

i. Current Behavior

These features (listed in Table 6a) characterize the current behavior of viewers in the session. They include fixed effects for viewer (6157), show (558), genre[15] (18), month (5), week (5), day (2) and time of day (5) as well as continuous variables for episode length of the first episode viewed (1), number of episodes of the TV show ahead in sequence (1) and number of unwatched episodes of the TV show during our sample period (1). These features do not depend on a viewer's historical activity. As our model (in

[15] If a TV show is labelled with multiple genres (0.08% of the sessions), we use the first genre label assigned to it in the data.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

subsection "Model") can handle multicollinearity among the features to make predictions, we include fixed effects for both show and genre, and then later determine the relative importance of the predictors in the section "Results". Since an individual viewer's content consumption and ad response may vary as a function of where a current episode of a TV show is in the show's entire chronology, we include two related measures to capture this. First, we measure the number of episodes of the TV show ahead in sequence (N_1) after the first episode viewed in the current session. Second, we measure the number of potentially unwatched episodes of the TV show (N_2) by subtracting the number of episodes viewed till date (during our sample period) from the total episodes available in our dataset.[16] In total, we have 6,753 features in this type.

"Insert Table 6a about here"

ii. Ad Targeting Rules

The four ad targeting rules (discussed in the earlier subsection "Ad Delivery") can be summarized using features as follows:

- Spacing Rule (SR) is the "mean time between pods in an episode" averaged across all episodes viewed in a session.[17]
- Length Rule (LR) is the "mean pod length in an episode" averaged across all episodes viewed in a session.
- Diversity Rule (DR) is the mean of ad diversity per episode across all episodes viewed in a session.
- Clumpiness Rule (CR) is the mean of clumpiness in pod locations per episode across all episodes viewed in a session.[18]

The absolute value of the correlation between every pair of the rules ranges from 0.16 to 0.51 which shows that the rules are not too strongly correlated, thus providing evidence that each one is capturing a distinct underlying decision rule.

iii. Past Behavior: Watching TV Shows Only

We construct 9 functions to systematically generate 68 features (listed in Table 6b) that characterize many aspects of viewers' TV-viewing behavior at the level of show, day of week, and time of day (cf. Yoganarasimhan (2019)). The features are computed using all TV-show-viewing activity for a user during

[16] Though N_1 and N_2 are correlated (rho=0.73), we use both to capture behavior in the most comprehensive manner, given that we are inferring the inventory at Hulu at any given time (as we do not have access to the actual episode supply). In spite of these measures, we could still miss episodes if they are not viewed by anyone in our dataset at the time of the session and/or if they were available only for a limited time.

[17] We do not consider time from the last pod shown in an episode till the end of an episode because a viewer could have stopped watching an episode at any time and not have waited till the end of the episode.

[18] For 2.5% of sessions where a viewer switches between the same episodes, (e.g., watches episode 1 - episode 2 - episode 1 - episode 2), an episode's ad targeting rule is found by averaging the rule value over each individual occurrence of the episode.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

the one-week window before their current session. For instance, if a viewer decides to watch some TV show on Sunday at 5 pm, we consider all of her sessions watching TV shows that began in the 168 (7*24) hours before Sunday at 5 pm. This moving window of one week is chosen so that we have adequate information of a viewer's recent historical viewing activity that includes both weekdays and weekends. We generate functions that vary with day and time of day to explore whether experiences that occur at specific times in the past are significant predictors of Bingeability and Ad Tolerance.

We explain one function in detail and show how its features are generated. The features for the other functions are generated similarly.

a) *Bingeability Sum (Show, Day, Time of Day):* This function calculates the past one-week sum of Bingeability of the viewer for the **Show** she is about to watch over that **Day** at that **Time of Day.** We consider **Day** as a Weekend or a Weekday and **Time of Day** as one of the five: Early Morning: 7–10am, Day Time: 10am–5pm, Early Fringe: 5pm – 8pm; Prime Time: 8pm – 11pm, Late Fringe: 11pm – 7am (Schweidel & Moe, 2016). For example, if a viewer decides to start watching the TV show *House* on a *Weekend* during *Day Time*, then the function will calculate the sum of Bingeability over all the sessions in the past week when the viewer viewed *House* on the *Weekend* during *Day Time*. More features can be generated by the function when the three variables – *Show, Day or Time of Day*, are dropped in turn from the function using a 2^3 design. Thus, a total of 8 features corresponding to *Bingeability Sum* (BS) can be generated for each session in our sample, and these are shown in Table 6c.

We note that for the first session of each viewer in the panel data, the value of features based on past behavior is 0 because past observations are censored. This is true for 5.6% of the sessions in our sample corresponding to 6,157 viewers. We do not drop these as they help us replicate situations when a new viewer joins the platform.

"Insert Table 6b and 6c about here"

iv. Past Behavior: Watching TV Shows or Movies

Even though the target behavior that we study is viewing of TV show content and ads, we still consider past movie-viewing behavior. This allows us to measure how ad exposure in the past week while watching a movie or a TV show influences the decision to see a TV show in the current session. We construct 11 distinct functions that generate 136 features (listed in Table 6d) which consider historical one-week sessions in which *either* TV shows or movies were seen. We replace 'Show' with 'Title' in the name of

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

these functions to indicate that when 'Title' is absent, the viewer could have watched either a TV show or a movie in the past week.

We explain two functions in detail and show how their features are generated. The features for the other functions are generated similarly.

a) *Pod Count (Pod Length, Title, Day, Time of Day):* This function calculates the past one-week sum of the number of pods of length equal to some **Pod Length,** shown to the viewer for that **Title** over that **Day** at that **Time of day**. Based on the histogram of Pod Length shown earlier in Figure 3a, we divide Pod Length into 3 categories: 1 (1 – 26 sec), 2 (26 – 52 sec) and 3 (>52 sec). We use the same breakdown for the categories as that used for Ad Length in the subsection "Ad Delivery" because more than 99% of the pods in our data have only 1 ad. This function generates a total of 32 features from this 4x2x2x2 design: Pod Length (4: *1, 2, 3,* __) x Title (2: *Title,* __) x Day (2: *Day,* __) x Time of Day (2: *Time of day,* __), where '__' corresponds to 'any value' as shown in Table 6c.
b) *Ad Diversity (Title, Day, Time of Day):* This function finds the past one-week average of the percentage of diverse ads shown in each session (in which there was ad exposure) for the viewer watching that **Title** over that **Day** at that **Time of day.** As we do not have a unique Ad ID for each ad in our dataset, we use a combination of Ad Industry (16 categories such as CPG, Telecom, etc.) and Ad Length (3 categories) to generate 48 unique ad combinations.

"Insert Table 6d about here"

Model

Model Setup. Given the set of chosen features (above), we need to develop a methodology to predict our key summaries of viewing behavior – Bingeability and Ad Tolerance – for a future session. The total number of features generated in the previous subsection "Feature Generation" is large (6,961). In order to capture the effects of this large set of features in the most flexible way, including non-linearities and interactions, we use machine learning methods (Lemmens & Croux, 2006; Neslin et al., 2006; Rafieian & Yoganarasimhan, 2019; Yoganarasimhan, 2019). These predictive methods also have the additional advantage of being scalable, handling many features for many users, and computationally efficient. Since we want to understand importance of different features and interpret those features' relationships with the outcomes, we use tree-based machine learning models. We express our model as follows:

$$Y_t = f_1(X_{1t}, X_{2t}, X_{3t}, X_{4t}, W_{1t}, W_{2t}) + u_t \quad (5)$$

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

where Y is the metric of interest (Ad Tolerance or Bingeability),[19] the subscript t denotes a session and f_1 is a non-linear function of all the features. X_1, X_2, X_3 and X_4, are the Spacing Rule (SR), Length Rule (LR), Diversity Rule (DR) and Clumpiness Rule (CR), respectively (as detailed earlier); W_1 is the matrix of features describing current behavior listed in Table 6a; W_2 is the matrix of features describing past behavior listed in Table 6c and 6d; and u is the error, which is assumed to be additively separable.

We assume W_1 and W_2 to be exogenous as they are determined before the session begins, and we assume there is no autocorrelation between the errors u_t. However, as noted earlier, the data patterns suggest that the X variables, which represent the ad targeting rules, are not set exogenously to the behavior of interest. In other words, they could be endogenous due (primarily) to simultaneity, i.e., X's could be set depending on the value of Y. For example, as Bingeability or Ad Tolerance (Y) increases, the mean spacing between pods (X_1) could increase because the streaming provider may only have a limited inventory of ads to deliver for that show at that time, leading to an average decrease in the frequency of pod spacing (if no other ads are available to compensate). The trade press has noted that low ad inventory was a frequent occurrence at Hulu around the time of our data (Sloane, 2019).

If this potential endogeneity is not corrected for, then our predicted outcomes will be biased, leading to non-optimal ad scheduling recommendations. We correct for this using instrumental variables. These instruments should affect Y only through their effect on X_i , $i = \{1,2,3,4\}$, i.e., be uncorrelated with unobservables u. We leverage the institutional detail that Hulu has been known to match sponsors with specific TV shows (Dubner, 2009). Therefore, ad schedules in an episode of a TV show for a focal viewer are likely to be correlated with ad schedules in the same episode for another viewer (while not depending on the focal viewer's viewing behavior). We construct episode-level instruments Z_i, for each X_i, $i = \{1,2,3,4\}$, in the same spirit as the instruments in Nevo (2000). We define Z_{it} to be the mean of v_i for all other viewer-episode pairs (involving any of the episodes viewed in session t) that began *before* the start of session t, where v_1 = time between pods, v_2 = pod length, v_3 = ad diversity and v_4 = clumpiness. Note that Z_{it} can affect Y_t only through its effect on X_{it}, because the focal viewer is unaware about the value of Z_{it} that was experienced by other viewers. This is a reasonable assumption for two reasons. First, our sample is a random draw from all Hulu viewers, lowering the chance that any two viewers would know each other at all. Second, at the time of our data, there is no discussion around ad delivery on Hulu's Facebook page, which was the brand's major online social media site at the time. In terms of the empirical relationship between X_i and Z_i , we find that the raw correlation between them for $i = \{1,2,3,4\}$ is reasonable at 0.35, 0.25, 0.27 and 0.33.[20]

[19] In Web Appendix E, we show how our approach can be modified to model the two Y's jointly. Given that there isn't a meaningful difference in the final recommendations, the additional benefit of doing so seems to be less than the additional methodological complexity required.

[20] The correlation between the instrument and the endogenous variables does not increase if the instruments are calculated over respective geographical states or regions of the focal viewer's permanent address, which indicates that the ad targeting rules are unlikely to vary by geographical

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Estimation Approach. The first stage of the estimation process can be expressed using a model of X_i as a function of Z_i ($i = \{1,2,3,4\}$), W_1 and W_2 as shown below:

$$X_{it} = g_i(Z_{1t}, Z_{2t}, Z_{3t}, Z_{4t}, W_{1t}, W_{2t}) + e_{it} \quad (6)$$

where, g_i is a non-linear function and e_i is the error term assumed to be additively separable with an expected value of 0. The estimates of the outcome variables from the above first-stage model can then be plugged as inputs to the second-stage model. The second stage of the estimation process can be expressed using a model of Y on $\hat{X}_{it}$ (estimates of X_i from the first-stage) as well as on W_1 and W_2:

$$Y_t = f_2\left(\hat{X}_{1t}, \hat{X}_{2t}, \hat{X}_{3t}, \hat{X}_{4t}, W_{1t}, W_{2t}\right) + u_t \quad (7)$$

where f_2 is a non-linear function, u is the error term of the second-stage ,which is assumed to be additively separable with an expected value of 0, and Y_t represents the values of Bingeability and Ad Tolerance.[21]

The use of instrumental variables along with machine learning methods is nascent, with no prior research in marketing using it, to the best of our knowledge. The machine learning literature has just begun to explore the use of instrumental variable approaches to infer causality. Two notable examples are Hartford et al. (2017) , which uses a deep learning framework with instrumental variables to make counterfactual predictions of the outcome, and Athey et al. (2019), which uses random forests with instrumental variables to find asymptotic marginal effects.

To decide which tree-based method to use,[22] we compare the performance of different methods using simulated data, so that we know the ground truth which is unlike the case with our observed dataset where we do not know the true explanatory power of the features (cf. Hartford et al. (2017)). We consider two popular tree-based machine learning methods known for their ability to get close to the ground truth – Extreme Gradient Boosting (XGBoost) and Random Forests (Breiman, 2001; Chen & Guestrin, 2016). Their performance is also compared with the traditional linear two-stage least squares (2SLS) approach. Our goal is to choose the better performing method for both the first and second stage of the model. Web Appendix F describes the two methods, the simulation, and the results, which show XGBoost gets closest

location of the viewer (assuming the viewer primarily watches content in the state/region of her permanent address which is the only address that is recorded in the data). We also find that show-level instruments (in comparison to episode-level instruments) have a lower correlation with the endogenous variables which suggests that ad characteristics are determined by the platform at the granular episode level and not the show level. Hence, we use episode-level instruments and not show-level instruments.

[21] In order to implement the instrumental variables approach, we can only use observations (sessions) for which we have complete information about $Z_i, i = \{1,2,3,4\}$. We remove 4,536 sessions where no other viewer had viewed those episodes before. Next, we drop 354 viewers who visited the platform exactly once (as their single sessions cannot be randomly assigned to both the training and holdout data).

[22] We also explored other linear models such as LASSO, Ridge Regression and Elastic Net but found that non-linear models fit the data better.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

to the ground truth. Our findings are consistent with past literature and the results of prediction competitions that have found gradient boosting methods, and especially XGBoost, to predict better on average than Random Forests (Olson et al., 2017; Oughali et al., 2019; Synced, 2017). Thus, our results are all based on the XGBoost method (implemented on a 4 core CPU with two threads per core at 3.6 GHz) that takes about 2 minutes to run.

RESULTS

Model Estimation

To estimate the model on the dataset, containing a total of 5,760 viewers, 508 unique shows, and 105,610 sessions (see Table 7), we construct a training dataset for calibration and two separate holdout datasets for estimation. We estimate our model on both future observations of the same set of viewers (Holdout 1) and observations of a completely new group of viewers (Holdout 2). First, we randomly hold out 500 viewers and select the remaining 5,260 viewers for training. Then, among these selected viewers, we select approximately 80% of their initial sessions to form the training sample (74,996 sessions), and 20% of their future sessions (21,497 sessions) to form the first holdout sample (Holdout 1). The remaining 500 viewers, with their 9,117 sessions form the second holdout sample (Holdout 2). As one of our objectives is to allow the streaming platform to build ad schedules for new TV shows that have not yet been viewed but could be viewed by current viewers or new viewers, we estimate the model on this task too. We allow both holdout samples to include sessions with 13 (Holdout 1) and 16 (Holdout 2) new TV shows not in the training data.

"Insert Table 7 about here"

Next, we estimate the first-stage model (6) using the training sample and get the estimates $\hat{X}_i$, $i = \{1,2,3,4\}$ for both the training and holdout samples. The estimates $\hat{X}_i$ are then plugged into the second stage of the model (7). We estimate the second-stage model using the training sample and obtain predictions for the outcomes in the holdout samples. The parameters of the XGBoost model are selected using 5-fold cross-validation repeated 10 times (Web Appendix G provides details on the cross-validation process and parameter tuning). The estimates of the outcome variables will be used as inputs to the ad scheduling process (detailed in section "Stage III").

A frequent critique of machine learning methods is that they operate as a "black-box" and yield results that are not interpretable. We try to address this via the use of two descriptive methods – "feature

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

importance" and "partial dependence" below. The former can be seen as analogous to the "average effect" of a covariate (coefficient times mean covariate) in a traditional regression setting while the latter can be seen as analogous to the "marginal effect" of a covariate (coefficient).

Feature Importance

As the name denotes, this method allows us to identify the features in equation (7) that are most predictive of the outcomes. A commonly used metric to do this is "Variance Reduction" (Hastie et al., 2009). This is the "gain" achieved when the tree is split on a feature, defined as the maximum reduction in RMSE (for continuous outcomes, like Ad Tolerance) or Negative Log Likelihood (for discrete outcomes, like Bingeability). We identify the features that are most frequently split during model training. Then we compute the gain of a set of multiple related features by summing up the gain for each individual feature in that set. The percentage gain for each set of features used to split the tree is reported in Tables 8a and 8b for the top 10 sets of predictive features for Bingeability and Ad Tolerance respectively.

The most important predictors of Bingeability are past predictors of 'Bingeability Sum', and 'Number of episodes ahead in sequence, N_1' and viewer fixed effects. The most important predictors of Ad Tolerance are viewer fixed effects, past predictors of 'Ad Tolerance Sum' and the past predictors of 'Pod End'. The fact that individual fixed effects and the sums of past outcomes are important in predicting the outcomes for a new session is not itself surprising, but this process quantifies their relative importance and identifies the other important features. The total gain contribution of the four estimated ad targeting rules, $\widehat{SR}, \widehat{LR}, \widehat{DR}$ & $\widehat{CR}$, is 9.0% for Bingeability and 8.2% for Ad Tolerance. This indicates that the four ad targeting rules have an important role to play in predicting the value of the metrics. The clumpiness of ads ($\widehat{CR}$) is the most important advertising pattern for predicting Bingeability (with a 4.8% gain) while the frequency of pod delivery ($\widehat{SR}$) is the most important advertising pattern for predicting Ad Tolerance (with a 4.3% gain) (Table 8c).

"Insert Table 8a,8b and 8c about here"

Partial Dependence

We use *partial dependence plots* (Friedman, 2001) to examine the (partial) relationship between the features and the outcomes, and to the best of our knowledge we are introducing this practice to marketing. Let $X = \{X_1, \dots, X_d\}$ be the set of all features in the training sample, and $f(X)$ be the corresponding prediction function. If X can be partitioned into a set of features of interest X_s and its complement set X_c, then the partial dependence of the outcome on X_S is defined as follows:

$$f_s(X_S) = E_{X_c}[\hat{f}(X_s, X_c)] = \int f(X_s, X_c) p_c(X_c) dX_c$$

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

where, $p_c(X_c)$ is the marginal probability density function of X_c. The above equation can be estimated from a set of training data by averaging out the effects of all the other features X_c in the model, while taking into account any correlations among features in X_s (Friedman, 2001; Greenwell, 2017). Empirically, for a single feature of interest, consider an observation's value of that feature, and create an otherwise identical copy of the dataset except substitute that value in for all other observations' values of that feature. For the newly edited data, obtain the model's predictions for each observation and average the predictions across all observations. Then repeat this for each observation of that feature, plotting feature values versus average prediction values. This can be better understood as a two-step process:

i. For $i = \{1, \dots, n\}$, where n is the number of observations in the training data,

a) Replace each value in X_S (n-dimensional vector) with X_{S_i} (constant)

b) Compute predicted values of the n outcome variables

c) Find average of the n predicted values = $\bar{f}_S(X_{S_i})$

ii. Plot $\{X_{S_i}, \bar{f}_S(X_{S_i})\}$ for $i = \{1, \dots, n\}$ to get the partial dependence plot.

To ease the computational burden, we compute the partial dependence over the deciles of the feature in addition to its 2.5th and 97.5th percentile. Figure 4a shows the relationship between Bingeability and its most important feature, *Bingeability Sum (same Show, any Day, any Time of day)*. This feature represents the sum of Bingeability across all sessions shown to the viewer in the past week for the same Show (as the current session) viewed on any Day at any Time of day. The figure shows that an increase in Bingeability for a show from 0 episodes to 15 episodes over the past week predicts an average increase in Bingeability for the same show in the current session by 0.6 episodes. The relationship between Ad Tolerance and its most important feature (other than viewer fixed effects), *Ad Tolerance Sum (same Title, any Day, any Time of day)*, is shown in Figure 4b. This feature calculates the sum of Ad Tolerance across all sessions shown to the viewer in the past week for the same Title (as the current session) viewed on any Day at any Time of day. The figure shows that an increase in Ad Tolerance for a title from -16 minutes to 3,513 minutes over the past week predicts an average increase in Ad Tolerance for the same title in the current session by 710 minutes. Both relationships (in Figures 4a and 4b) provide evidence of state dependence between the past and current sessions of a viewer for the same TV show.

"Insert Figure 4a and 4b about here"

As our goal is to make ad scheduling recommendations, we need to understand the relationships between the ad targeting rules and our two outcome variables. The partial relationships between the ad

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

targeting rules that are most predictive, clumpiness ($\widehat{CR}$) and spacing ($\widehat{SR}$), and the predicted values of Bingeability and Ad Tolerance respectively, are shown in Figures 4c and 4d. Lower clumpiness values, i.e., more equally spaced pods, predict higher Bingeability (Figure 4c). Moreover, the extent of the influence of $\widehat{CR}$ (over its 2.5th to 97.5th percentile range) on Bingeability is $\pm$ 0.44 episodes. The extent of the influence of $\widehat{SR}$ (over its 2.5th to 97.5th percentile range) on Ad Tolerance is $\pm$ 18.95 minutes (Figure 4d), with most of the change occurring from the 90th percentile (7.7 minutes) to 97.5th percentile (8.6 minutes) of spacing. This suggests that, on average, spacings longer than 7.7 minutes can overly adapt viewers to the content and/or absence of ads and increase their aversion to ads.

"Insert Figure 4c and 4d about here"

It is also possible that the ad targeting rules may interact, so we examine the partial dependence of two predictors jointly. We consider our first pair of predictors of interest to be $X_{s1} = \{\hat{X}_1, \hat{X}_4\} = \{\widehat{SR}, \widehat{CR}\}$, the predicted ad targeting rules that are most important in predicting Bingeability and then the second pair of predictors $X_{s2} = \{\hat{X}_1, \hat{X}_2, \} = \{\widehat{SR}, \widehat{LR}\}$, since these two rules are most important in predicting Ad Tolerance. The partial dependences of the estimated values of Bingeability and Ad Tolerance on each pair of their important predictors are shown in Figures 4e and 4f. Figure 4e shows that the magnitude of the influence of the top two ad targeting rules (over their 2.5th to 97.5th percentile range) on Bingeability is $\pm$ 0.24 episodes. Furthermore, Figure 4e shows that higher values of $\widehat{SR}$ and lower values of $\widehat{CR}$ predict higher Bingeability. Similarly, Figure 4f shows that the magnitude of the influence of the top two ad targeting rules (over their 2.5th to 97.5th percentile range) on Ad Tolerance is $\pm$ 15.94 minutes, which is less than the size of the partial dependence on $\widehat{SR}$ alone, $\pm$ 18.95 minutes, found in Figure 4d. Furthermore, Figure 4d also shows that lower values of $\widehat{SR}$ and higher values of $\widehat{LR}$ predict higher Ad Tolerance.

"Insert Figure 4e and 4f about here"

Finally, we look at the effect of the pairwise interactions (six) across all the four estimated ad targeting rules $X_{s3} = \{\hat{X}_1, \hat{X}_2, \hat{X}_3, \hat{X}_4\} = \{\widehat{SR}, \widehat{LR}, \widehat{DR}, \widehat{CR}\}$. The partial dependences of the estimated values of Bingeability and Ad Tolerance on X_{s3} are calculated over the quintiles of each variable in addition to their 2.5th and 97.5th percentile to ease computational burden. The extent of the influence of the four ad targeting rules (over their 2.5th to 97.5th percentile range) on Bingeability is $\pm$ 0.25 episodes, which is about the same as $\pm$ 0.24 episodes found in Figure 4e. Thus, there is almost no additional impact on Bingeability. Similarly, the extent of the influence of the four ad targeting rules (over their 2.5th to 97.5th

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

percentile range) on Ad Tolerance is ± 33.08 minutes, which is more than the extent of ± 15.94 minutes found in Figure 4f.

From Table 7, the median value of Bingeability in the data is 1 episode and that of Ad Tolerance is 23.69 minutes. The partial dependence analysis is useful in that it tells us the impact of different variables (or sets of variables) on the outcome variables. For example, based on the above, we know that the Ad Targeting rules, in combination can effect a maximum change of 25% (0.25/1.00) on median Bingeability and a maximum change of 140% (33.08/23.69) on median Ad Tolerance. Overall, these analyses show that the ad targeting rules, individually and together, have a material impact on viewer behavior as captured via the two outcomes.

STAGE III

With summaries of behavior predicted and the importance of the features that predict those summaries understood, in the third stage we use the predicted values of the behavioral summary metrics to make ad scheduling recommendations. We do this in two steps. First, we provide a guide to the streaming provider on how to use these predictions with a decision tree, and then we use an optimization procedure to recommend a better ad schedule in any given session. In order to illustrate the properties of our generated ad schedule for each session in the holdout samples, we contrast it with the current ad schedule (observed in the data) and an alternative ad schedule based on a naïve heuristic.

Ad Decision Tree

We propose an "Ad Decision Tree" (Figure 5) to identify the types of sessions where ads may enhance – or at least not detract from – content consumption. The Ad Decision Tree takes in the predictions of Bingeability and Ad Tolerance obtained from the model and recommends action. The first decision split in the Ad Decision Tree is to check whether the predicted value of Bingeability is greater than a threshold, T. If the predicted value of Bingeability for the session is less than the threshold, then the streaming platform is advised to not show any ads in the session. This is because there is not much incentive for a free ad-supported only streaming platform to show ads in a session if the ads are predicted to prevent the viewer from completing a desired number of episodes (represented by the chosen threshold value for Bingeability). By ensuring that a viewer is predicted to watch at least beyond that threshold, the streaming platform will be able to provide a minimum level of engagement with the content on its platform. We examine the impact of increasing the threshold in the subsection "Decision Support System," but for now, we start by choosing the lowest Bingeability threshold of 0 episodes i.e., show ads for all sessions.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

"Insert Figure 5 about here"

If the predicted value of Bingeability is greater than or equal to the threshold, we move to another part of the tree and check the sign of the predicted value of Ad Tolerance. Negative values of Ad Tolerance capture occasions where viewers stopped watching content after being exposed to a pod, which itself was preceded (at some point) by a longer period of no ad exposure. On the other hand, a positive value of Ad Tolerance indicates occasions where ads were shown more frequently to a viewer and the viewer continued to watch content.

If the predicted value of Ad Tolerance is > 0, then we solve a novel optimization procedure discussed in subsection "Optimization". If the predicted value of Ad Tolerance is ≤ 0, then it is unclear how tolerant a viewer is towards seeing a pod of ads, so our proposed decision tree recommends testing and then adapting to what is learned. To test whether the viewer can have both Ad Tolerance > 0 and Bingeability ≥ 1, the streaming platform is advised to show pods within the first half of each episode to resemble occasions of frequent ad exposure. To ensure an overall minimum ad exposure within the first half of an episode, it would be best to show pods at an interval of a quarter of the episode length with "regular interruptions" (discussed further in subsection "Optimization"). Based on the viewer's response to the ad exposure in the first half of the episode, if the viewer continues to have Bingeability ≥ 1, then the viewer's Ad Tolerance is updated to > 0 and the rest of the optimization procedure can be implemented.

The recommendations made by the Ad Decision Tree for the observations in the two holdout samples are summarized in Table 9. We find that for most of the sessions in both holdout samples (94% of observations in Holdout 1 and 97% of observations in Holdout 2 – see Set C, Table 9), the recommendation to the streaming provider is to use the proposed optimization procedure.

"Insert Table 9 about here"

Optimization

In this research, we have set the objective of a streaming provider to maximize ad exposure (to earn more ad revenue) subject to the constraint of not detracting from the consumption experience. We can express the maximization of the objective function for a given session as follows:

$$\max f(n,d) = \sum_{j=1}^{n} d_j \quad \text{where} \quad \sum_{j=1}^{n} s_j + s' = \overbrace{\tilde{\tilde{b}e}}^{\text{expected content watched}} \tag{8}$$

where, n is the number of pods shown in a session, d_j is the duration (length) of pod j, s_j is the spacing (content time shown) between pod j-1 (or beginning of session if j=1) and pod j, s' is the duration of

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

content time shown after the end of pod n, $\hat{b}$ is the estimated Bingeability from the model, and e is the average episode length of all episodes of the TV show watched in that session in our dataset.

Our findings in subsection "Partial Dependence" showed that lower values of clumpiness (i.e., more equal spacings between pods) result in higher values of Bingeability. In addition, past literature has shown that viewers are less likely to adapt to irregular sources of interruptions, such as dormitory noise or aircraft noise (Frederick & Loewenstein, 1999). Hence, it is likely that having regularity in interruptions would assist the adaptation process and increase Bingeability. Unequal spacing s_j between pods and unequal duration d_j for each pod are sources of irregularity. To remove irregularities within a session, we let the duration of each pod d_j be equal to d and let the duration of each spacing s_j and s' be equal to s. Consequently, we rewrite the optimization as follows:

$$\max f(n,d) = nd \quad \text{where} \quad s(n+1) = \hat{b}e \tag{9}$$

where the product of spacing between pods, s, and number of pods plus one, $n+1$, should equal the product of predicted Bingeability, $\hat{b}$, and average episode length, e. The constraint $s(n+1) = \hat{b}e$ allows only mid-roll ads (i.e., no pre-roll ads or post-roll ads). This is because prior work has found that viewers are more likely to completely view mid-roll ads, followed by pre-roll ads and finally post-roll ads (Krishnan & Sitaraman, 2013). Note that the subsection on "Decision Support System" relaxes this constraint to allow for pre-roll ads.

Our objective function is subject to the constraint of not detracting from the content consumption experience i.e., not exceeding the predicted Ad Tolerance for a session. Using equation (4) and a series of stepwise substitutions shown in Web Appendix H (Part 1), this constraint can be expressed as follows:

$$\hat{a} = w_1 nd + w_2 \left(n\hat{b}e - \frac{n(n+1)}{2}s\right) - w_3 ns \tag{10}$$

where w_1, w_2, w_3 are the three weights, originally present in equation (4), and $\hat{a}$ is the predicted value of Ad Tolerance. We also have additional constraints that there must be at least one pod, and the duration of a pod must be non-zero. Therefore, the constrained optimization problem in equation (9) can be expressed as follows along with all its constraints:

$$\max f(n,d) = nd \quad \text{where} \quad s(n+1) = \hat{b}e$$

$$\text{such that } \hat{a} = w_1 nd + w_2 \left(n\hat{b}e - \frac{n(n+1)}{2}s\right) - w_3 ns,\ n \geq 1, \text{ and } d > 0$$

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Since we are setting spacing to be a constant function of expected total episode content viewed, we replace s with $\frac{\hat{b}e}{n+1}$ in the constraints, and then we can re-express our constrained optimization problem as

$$\max f(s,d) = nd \tag{11}$$

$$\text{such that } \hat{a} = w_1 nd + w_2\left(\frac{n\hat{b}e}{2}\right) - w_3\left(\frac{n\hat{b}e}{n+1}\right), n \geq 1, \text{ and } d > 0$$

By applying the Lagrange function to the optimization problem, we get the following expression:

$$L(n,d,\lambda_1,\lambda_2,\lambda_3) = nd - \lambda_1\left(\hat{a} - w_1 nd - w_2\left(\frac{n\hat{b}e}{2}\right) + w_3\left(\frac{n\hat{b}e}{n+1}\right)\right) + \lambda_2(n-1) + \lambda_3 d$$

with the following six constraints: (1) $\frac{\partial L}{\partial n} = 0$ (2) $\frac{\partial L}{\partial d} = 0$, (3) $\lambda_2(n-1) = 0, \lambda_3 d = 0$ (4) $n \geq 1,\ d > 0$ (5) $\hat{a} = w_1 nd + w_2\left(\frac{n\hat{b}e}{2}\right) - w_3\left(\frac{n\hat{b}e}{n+1}\right)$ (6) $\lambda_1, \lambda_2, \lambda_3 \geq 0$

We use the fifth constraint to solve for n, so we get a quadratic equation in n:

$$n^2\left(w_1 2d + w_2\hat{b}e\right) + n\left(w_1 2d - (2w_3 - w_2)\hat{b}e - 2\hat{a}\right) - 2\hat{a} = 0 \tag{12}$$

As $d > 0$, and $\hat{a} > 0$, $\hat{b} \geq 1$ (from the Ad Decision Tree), the above equation has one positive root and one negative root of n. Solving the other constraints of the Lagrange Function does not give solutions within the acceptable parameter space. Next, we set the weights, w_1, w_2, w_3, to 1, as originally done in subsection "Metric Development," although these can be set differently, which we consider in Web Appendix C. The two unknown parameters in equation (12) are d and n. We fix d at 30 seconds,[23] the median pod duration in our dataset, and then solve equation (12) for the optimal $\tilde{n}$, and use its positive root which can be expressed as follows:

$$\tilde{n} = \frac{-\left(1 - \hat{b}e - 2\hat{a}\right) + \sqrt{\Delta}}{2\left(1 + \hat{b}e\right)}, \text{where } \Delta = \left(\hat{b}^2e^2 + 12\hat{a}\hat{b}e - 2\hat{b}e + 4\hat{a}^2 + 4\hat{a} + 1\right) \text{ and } \sqrt{\Delta} > 0 \tag{13}$$

[23] We also run the optimization using a fixed pod duration of 15 seconds instead of 30 seconds as the distribution of pod length (Figure 2a) shows a second peak at 15 seconds. The recommended spacing using 15 seconds and that using 30 seconds is almost identical (and the difference on average is less than 6 seconds).

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Therefore, we have found the recommended number of pods $\tilde{n}$ from the optimization routine, and this implies that the recommended spacing $\tilde{s} = \frac{\hat{b}e}{\tilde{n}+1}$. Hence, our optimization procedure recommends the pod frequency $\tilde{s}$ for a viewer's session holding pod duration d constant.[24,25]

In order to understand the effect of the estimates of Bingeability, $\hat{b}$, and Ad Tolerance, $\hat{a}$, on the recommended number of pods $\tilde{n}$ and pod spacing, $\tilde{s}$ (which is proportional to $\frac{1}{\tilde{n}}$), we take the partial derivatives of $\tilde{n}$ in (12) with respect to $\hat{b}$ and then with respect to $\hat{a}$. The expressions of the partial derivatives are shown in Web Appendix H (Part 2). The partial derivatives suggest that the streaming provider should on average increase pod spacing (decrease pod frequency) when a viewer is expected to have lower Ad Tolerance or higher Bingeability, holding the other constant. Similarly, the streaming provider should on average decrease pod spacing (increase pod frequency) when a viewer is expected to have higher Ad Tolerance or lower Bingeability, holding the other constant.

To summarize all of these session-by-session ad spacing recommendations, we consider their full distribution. The density of the recommended spacing for the two holdout samples helps illustrate the range of recommendations made by the optimization routine (Figures 6a and 6b). The median value of the recommended spacing for Holdout 1 (future sessions of the viewers in the training sample) is 4.33 minutes, and its 2.5th to 97.5th percentile range is from 0.61 to 9.80 minutes. The median value of the recommended spacing for Holdout 2 (new viewers) is 4.43 minutes, and its 2.5th to 97.5th percentile range is from 0.71 to 9.43 minutes.[26] While this may seem like very frequent ad exposure, it is not very different from that in the data (below). In addition, Nelson et al. (2009) show ads every 2 minutes in their experiments and current industry practice is experimenting with comparable or even shorter ad spacing (Gessenhues, 2018). Note that the optimization procedure takes about 10 seconds.

"Insert Figure 6a,6b,7a,7b,8a and 8b about here"

Recommended Schedule: Comparison with Data

In this section, we compare the recommended spacing $\tilde{s}$ for a session with the average observed spacing $\bar{s}$ in the session. The average observed spacing for a session is calculated across each of its observed spacings, s_j , which is the content time shown between pod j-1 (or beginning of session if j=1) and pod j. We

[24] We express the constraint as a quadratic equation in $\tilde{n}$, and not $\tilde{s}$, because the product of the roots in the quadratic equation of $\tilde{n}$ is always negative, giving us one positive and one negative root, and helping us choose the positive root. The product of the roots in the quadratic equation of $\tilde{s}$ is always positive, making it harder to choose the appropriate positive root.

[25] We do not directly recommend number of pods, $\tilde{n}$, because the number of pods to be shown is not under direct control of the streaming provider. The streaming provider can only set the spacing (content time) after which a pod must be shown. The total number of pods that the viewer will end up viewing depends on the endogenous decision of the viewer to stop viewing content.

[26] We also examined recommendations by show length and genre. For show episodes < 30 mins, the spacing was 4.44 mins (4.26 mins) for Holdout 1 (Holdout 2), while for show episodes > 30 mins, it was 4.22 mins (4.57 mins). For Comedy shows, it was 4.60 mins (4.49 mins), for Drama 4.40 mins (4.76 mins) and for Science Fiction 4.95 mins (5.11 mins).

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

do not consider the content viewed from the end of the last pod till the end of the session, s', as a viewer could have ended the session before the end of an episode thus biasing the value of s'.

The density of the average observed spacing, $\bar{s}$, for these sessions in both holdout samples is shown in Figures 7a and 7b. The median value of the average observed spacing for Holdout 1 (future sessions of current viewers) is 6.43 minutes and its 2.5th to 97.5th percentile range is from 0 to 13.30 minutes. The median value of the average observed spacing for Holdout 2 (new viewers) is 6.96 minutes and its 2.5th to 97.5th percentile range is from 0 to 14.40 minutes. The peak at 0 minutes corresponds to sessions where there was only a pre-roll ad (ads at beginning of a session) and hence the spacing is 0 minutes.

We use the ratio of our recommended spacing to the average observed spacing in the data to highlight the difference of our approach. The distribution of the ratio of recommended spacing and average observed spacing for these sessions in the two holdout samples is shown in Figures 8a and 8b (and sessions with pre-roll ads only are dropped to avoid division by 0). The median value of the recommended ratio is 0.66 and 0.61 for Holdout 1 and 2, respectively. The optimization recommends a shorter spacing than observed (when the ratio is less than 1) in 81% and 86% of the sessions (Figure 8a, 8b). In these sessions, the streaming provider is recommended to show ads more frequently than current practice to maximize ad exposure, thus increasing revenue. On those occasions when the ratio is greater than 1, the streaming provider is recommended to show ads less frequently (with a longer spacing) than current practice to avoid compromising the content consumption experience and promote viewer engagement with the content on the platform.

Decision Support System

The ad decision tree can be used as a decision support system by the platform. Specifically, the platform can define critical thresholds of Bingeability and obtain the recommended ad delivery schedules to explore the inherent tradeoffs between content consumption and ad exposure for its viewers. The threshold is set so that for sessions with predicted Bingeability below the threshold, T, there should be no ads served. The recommended number of ads, $\tilde{n}$, is compared with the observed ad exposure, n, in Table 10a and Figure 9a for different values of the threshold, T.

Using $\tilde{s}$, $\hat{b}$ and e, we can derive the recommended spacing rule $\tilde{X}_1$ which is the "mean recommended spacing between pod exposures in an episode" averaged across all episodes predicted to be viewed in a session. Similarly, using $\tilde{s}$, $\hat{b}$ and e, we can also derive the recommended clumpiness rule $\tilde{X}_4$, which is the recommended clumpiness of pods throughout an episode, averaged across all episodes predicted to be viewed in a session. In the Bingeability model (equation (7)), we then replace $\hat{X}_1$ (Spacing Rule) and replace $\hat{X}_4$ (Clumpiness Rule) with their newly recommended values $\tilde{X}_1$ and $\tilde{X}_4$, respectively. We also replace $\hat{X}_2$ (Length Rule) with 0.5 (median pod duration) and keep $\hat{X}_3$ (Diversity Rule) as it is.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Then we find the optimized predictions of Bingeability based on our recommended ad schedule, which we denote as, $\tilde{b}_{withad}$. Next, for those observations which had initial predictions of Bingeability, $\hat{b}$, below the threshold T (where the platform is advised to not show ads[27]) we train the Bingeability model *without* the four ad targeting rules ($\hat{X}_1, \hat{X}_2, \hat{X}_3, \hat{X}_4$), and then make revised predictions, $\tilde{b}_{woad}$. We report the net incremental change in $\tilde{b}_{withad} + \tilde{b}_{woad}$ (= $\tilde{b}$) as compared to observed Bingeability b, and initial predicted Bingeability $\hat{b}$, in Tables 10b and 10c respectively, for different values of the threshold. These comparisons are also shown in Figures 9b and 9c for Holdout 1 and Holdout 2 respectively.

"Insert Table 10a,10b and 10c about here"

"Insert Figure 9a,9b and 9c about here"

The results for Holdout 1 (future sessions of current viewers) and Holdout 2 (new viewers) show the tradeoff between content consumption (measured through Bingeability) and ad exposure. From Table 10a and Figure 9a, we see that there is a net increase in ad exposure for thresholds (of predicted Bingeability) ≤ 0.8 for observations in Holdout 1 and for thresholds ≤ 0.9 for observations in Holdout 2. A threshold of 0 results in the maximum increase in ad exposure for both Holdout 1 and Holdout 2. From Tables 10b & 10c and Figures 9b & 9c, we see that the maximum increase in content consumption for Holdout 1 is for a threshold of 0, and the maximum increase (or lowest decrease) in content consumption for Holdout 2 is for a threshold of 1.6.

Overall, for future sessions of current viewers, if the platform uses a threshold of 0 to show ads, the platform gets more ad exposures and viewers see more content. This results in a 71.2% increase in ad exposure, as compared to what was observed, and a 5.17% increase in Bingeability as compared to the initially predicted Bingeability before optimization (or a 5.33% increase in Bingeability as compared to observed Bingeability).[28] On the other hand, for new viewers, there is a tension: the platform is better off in terms of ads shown if it uses a threshold of 0 to show ads which results in a 79.1% increase in ad exposure; whereas viewers are better off in terms of content viewed if the platform uses a threshold of 1.6 to show ads which results in a 1.03% increase in Bingeability as compared to the initially predicted Bingeability (or a decrease of 0.80% in Bingeability as compared to observed Bingeability). This indicates that for new viewers for whom preferences are unknown, there is no single threshold T that can lead to the best outcome for both the platform and the viewer. The best that can be done in this case is to compare the

[27] It is important to note that we also do not show ads for those sessions for which $\hat{b} > T$ and $\hat{\tilde{b}} < \tilde{s}$, i.e., if predicted value of Bingeability is greater than the threshold but less than the recommended spacing, we are unable to show ads.

[28] While a 71% increase seems large, it is within the range of the observed data - for Holdout 1, the recommended range of ad exposure is once every 0.01-61.01 minutes (data is 0.00-108.43 minutes) and for Holdout 2 is 0.01-28.94 minutes (data is 0.00-103.38 minutes).

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

optimized Bingeability with initial predicted Bingeability using the same set of features, giving a 1% increase in content consumption.

It is important to note that for Holdout 1, the best threshold of 0 corresponds to *showing ads* for most sessions which results in a net increase in content consumption by 5.2%. This is higher than the net increase in content consumption of 2.1% for a threshold of 9 that corresponds to *not showing ads* for most sessions. This indicates that the decision to show ads for future sessions of current viewers (Holdout 1) under the optimized ad schedule can make viewers better off as compared to a decision to not show ads.

We also consider the impact specifically of allowing pre-roll ads. Casual observation suggests that that it may increase ad exposure but lower content consumption. To test this, we allow for pre-roll ads by modifying the constraint in equation (9) to $sn = \hat{be}$, and then we run the optimization routine followed by the steps outlined in the Decision Support System for a threshold of 0. We find that ad exposure increases substantially (23%) as compared to ad exposure under our recommendation. However, content consumption decreases as compared to our recommendation by about 0.7% across both holdout samples. The decrease in content consumption is driven by the average decrease in mean spacing between pods that results from allowing pre-rolls ads (in addition to mid-roll ads) for the same level of predicted Ad Tolerance. Thus, allowing pre-roll ads results in much higher ad exposure but comes at a cost of a very small reduction in content consumption, a trade-off that a platform may be willing to make.

We also test the performance of a naive heuristic that computes pod spacing as a ratio of the total content time to the total number of pods for a viewer in a given week (see Web Appendix I for details). The best this heuristic can do is to increase content consumption at the expense of decreasing ad exposure (compared to observed practice) i.e., not delivering a win-win recommendation for the platform and its viewers.

CONCLUSION

This paper adds to the small but growing body of work that investigates the implications of increase in consumer control vis-à-vis content consumption on streaming media. To the best of our knowledge, this paper is the first attempt at providing a solution for advertising scheduling in such settings. Specifically, it provides an approach for streaming providers to explore the tradeoff between content consumption and ad exposure in order to provide a balanced viewing experience. The recommendations from this approach are available at the granular level of an individual viewer-session. The approach also uses state-of-the-art methods such as machine learning, but more importantly allows for causal inference via the use of instrumental variables and provides increased interpretability of the estimates.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

In the first stage of the three-stage approach, we develop two new metrics – Bingeability and Ad Tolerance – to capture the interplay between content consumption and ad exposure for each session. We need to do this as there is little standardization around the measurement of content consumption and ad exposure in streaming media settings. Our metrics are motivated by the consumer psychology literature on flow states and hedonic adaptation as well as observed consumer behavior (in these settings). In the second stage, we first use feature generation to summarize the current and past viewing environment of each consumer over a moving one-week window. We then use a novel tree-based instrumental variable approach to predict the value of the metrics. Using feature importance and partial dependence analyses, we provide insights into the relative importance of various features in predicting viewer consumption patterns. In the third stage, we pass the predictions from the previous stage through a decision tree and an optimization routine. This is followed by the construction of a decision support system which allows the platform to explore the tradeoff between content consumption and ad delivery for both current and new viewers. The platform can then make choices around its ad schedule for each session given its objective function. It is important to note that "win-win" ad schedules are possible e.g., for current viewers, we are able to find schedules that simultaneously allow for higher content consumption (a 5.2% increase in Bingeability) at higher levels of ad exposure (a 71.2% increase).

Our approach could potentially be applied to other ad supported environments, especially where consumers have control over content consumption e.g., news media consumption. Our decision support system can also be integrated into an online experimentation platform, where recommendations can be tested in live settings and where the results from experiments can be used to improve the performance of predictive models.

Our work does suffer from some limitations. First, while we believe that our approach is general, it is calibrated on data from just one streaming provider. Second, our optimization algorithm simplifies ad scheduling. While it provides conservative results, it can be improved (at the cost of complexity). Third, even though free ad-supported streaming platforms continue to grow, there are now combinations of free/paid ad-supported and paid ad-free models available within the same platform. Figuring out ad scheduling in these settings would necessitate modifications to our approach. Fourth, we cannot link our optimal ad exposure to final purchase due to lack of data. Finally, given the increasing availability of different online streaming options on multiple devices, newer patterns of non-linear consumption could emerge, perhaps requiring the development of other metrics. We hope that future work can address these limitations.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

REFERENCES

ABC. (2014). End Credit Guidelines. Retrieved from http://www.abc.net.au/tv/independent/doc/ABC_Commissioned_Productions_Credit_Guidelines_2014.pdf

Ailawadi, K. L., Lehmann, D. R., & Neslin, S. A. (2003). Revenue premium as an outcome measure of brand equity. *Journal of Marketing, 67*(4), 1-17.

Ameri, M., Honka, E., & Xie, Y. (2019). The Effects of Binge-Watching on Media Franchise Engagement. *Available at SSRN: https://ssrn.com/abstract=2986395 or http://dx.doi.org/10.2139/ssrn.2986395*.

Armental, M. (2019, January 10). Amazon's IMDb Launches Ad-Supported Streaming-Video Service. Retrieved from https://www.wsj.com/articles/amazons-imdb-launches-ad-supported-streaming-video-service-11547161586

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148-1178.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Cakebread, C. (2017, September 15). Here are all the reasons why Americans say they binge-watch TV shows. Retrieved from https://www.businessinsider.com/reasons-why-americans-binge-watch-tv-shows-chart-2017-9

Chae, I., Bruno, H. A., & Feinberg, F. F. (2018). Wearout or Weariness? Measuring Potential Negative Consequences of Online Ad Volume and Placement on Website Visits. *Working Paper*.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Danaher, P. J., Lee, J., & Kerbache, L. (2010). Optimal internet media selection. *Marketing Science, 29*(2), 336-347.

Deloitte. (2018, March 20). Meet the MilleXZials: Generational Lines Blur as Media Consumption for Gen X, Millennials and Gen Z Converge. Retrieved from https://www2.deloitte.com/us/en/pages/about-deloitte/articles/press-releases/digital-media-trends-twelfth-edition.html

Dubé, J.-P., Hitsch, G. J., & Manchanda, P. (2005). An empirical model of advertising dynamics. *Quantitative marketing and economics, 3*(2), 107-144.

Dubner, S. (2009, May 13). Your Hulu Questions, Answered. Retrieved from http://freakonomics.com/2009/05/13/your-hulu-questions-answered/

eMarketer. (2018, August 16). Audience for Connected TV Grows, but Ad Spending Has Lagged. Retrieved from https://www.emarketer.com/content/audience-for-connected-tv-grows-but-ad-spending-has-lagged

Frederick, S., & Loewenstein, G. (1999). 16 Hedonic Adaptation. *Well-Being. The foundations of Hedonic Psychology/Eds. D. Kahneman, E. Diener, N. Schwarz. NY: Russell Sage*, 302-329.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Gessenhues, A. (2018, June 29). Is YouTube serving up more pre-roll & mid-roll video ads? Retrieved from https://marketingland.com/is-youtube-serving-up-more-pre-roll-mid-roll-video-ads-243505.

Ghani, J. A., & Deshpande, S. P. (1994). Task characteristics and the experience of optimal flow in human—computer interaction. *The Journal of Psychology, 128*(4), 381-391.

Greenwell, B. M. (2017). pdp: an R Package for constructing partial dependence plots. *The R Journal, 9*(1), 421-436.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). *Deep IV: A flexible approach for counterfactual prediction.* Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.): New York: Springer.

Ingram, K. (2016, July 7). A brief history of TV shows' opening credit sequences. Retrieved from http://theweek.com/articles/632836/brief-history-tv-shows-opening-credit-sequences

Johnson, L. d. (2019, August 14). Customer Experience Key to Streaming Advertising Success. Retrieved from https://www.admonsters.com/customer-experience-key-streaming-advertising-success/

Krishnan, S. S., & Sitaraman, R. K. (2013). *Understanding the effectiveness of video ads: a measurement study.* Paper presented at the Proceedings of the 2013 conference on Internet measurement conference.

Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research, 43*(2), 276-286.

Liyakasa, K. (2018). Netflix Will Have Ads, And Other Predictions From Top TV Ad Chiefs. Retrieved from https://adexchanger.com/digital-tv/netflix-will-ads-predictions-top-tv-ad-chiefs/

Lu, T., Bradlow, E., & Hutchinson, J. W. (2017). Binge Consumption of Online Content. *Carnegie Mellon University, Working Paper*.

Melki, G., Cano, A., Kecman, V., & Ventura, S. (2017). Multi-target support vector regression via correlation regressor chains. *Information Sciences, 415*, 53-69.

Miller, L. S. (2017, March 16). Netflix Shouldn't Let Fans Skip Movie Credits, But We'll Allow It For TV Shows. Retrieved from http://www.indiewire.com/2017/05/netflix-skip-intro-bad-for-film-good-for-tv-1201817946/

Nededog, J. (2017, March 17). Some lucky Netflix members have a cool new 'skip intro' button to make binge-watching better. Retrieved from http://www.businessinsider.com/netflix-tests-skip-intro-button-to-improve-binge-watching-2017-3

Nelson, L. D., Meyvis, T., & Galak, J. (2009). Enhancing the television-viewing experience through commercial interruptions. *Journal of consumer research, 36*(2), 160-172.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204-211.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy, 9*(4), 513-548.

Ogasawara, T. (2011, June 23). Hulu Plus Sort of Available for Android Phones: Hello Android Fragmentation. Retrieved from https://www.adweek.com/digital/hulu-plus-sort-of-available-for-android-phones-hello-android-fragmentation/

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv:1708.05070*.

Oughali, M. S., Bahloul, M., & El Rahman, S. A. (2019). *Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models.* Paper presented at the 2019 International Conference on Computer and Information Sciences (ICCIS).

Oxford Dictionary (2018). Definition of binge-watch in US English. Retrieved from https://en.oxforddictionaries.com/definition/us/binge-watch

Patel, S. (2018, October 1). The anti-Netflix: Free, ad-supported video streaming services are growing. Retrieved from https://digiday.com/media/free-video-streaming-services-publishers-tv-ambitions/

Rafieian, O., & Yoganarasimhan, H. (2019). Targeting and Privacy in Mobile Advertising. *Available at SSRN: https://ssrn.com/abstract=3163806 or http://dx.doi.org/10.2139/ssrn.3163806*.

Sahni, N. S. (2015). Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quantitative Marketing and Economics, 13*(3), 203-247.

Schweidel, D. A., & Moe, W. W. (2016). Binge watching and advertising. *Journal of Marketing, 80*(5), 1-19.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Sherman, A. (2019). NBC is removing 'The Office' from Netflix in 2021 and putting it on its new streaming service. Retrieved from https://www.cnbc.com/2019/06/25/nbc-to-remove-the-office-from-netflix.html

Sloane, G. (2019). Hulu puts a cap on ad loads. Retrieved from https://adage.com/article/media/hulu-cuts-ad-breaks-half/317174/

Sommerlad, J. (2018). Netflix will never host advertising or enter battle for live news and sport, ceo says. Retrieved from https://www.independent.co.uk/life-style/gadgets-and-tech/news/netflix-advertising-live-broadcasting-mobile-streaming-30-second-trailers-reed-hastings-a8245701.html

Synced. (2017, October 22). Tree Boosting With XGBoost — Why Does XGBoost Win "Every" Machine Learning Competition? Retrieved from https://medium.com/syncedreview/tree-boosting-with-xgboost-why-does-xgboost-win-every-machine-learning-competition-ca8034c0b283

West, K. (2013). Unsurprising: Netflix Survey Indicates People Like To Binge-Watch TV. Retrieved from http://www.cinemablend.com/television/Unsurprising-Netflix-Survey-Indicates-People-Like-Binge-Watch-TV-61045.html

Yoganarasimhan, H. (2019). Search personalization using machine learning. *Management Science*.

Zhang, Y., Bradlow, E. T., & Small, D. S. (2014). Predicting customer value using clumpiness: From RFM to RFMC. *Marketing Science, 34*(2), 195-208.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Session (minutes)						
N	Min	2.5%	Median	Mean	97.5%	Max
122,617	0.02	1.82	42.70	56.06	236.51	1573.03

Table 1: Summary of Sessions

24 min episode of Aquarion	0.66	10	0.50	10	0.50	2		2	0.66	21		0.50	0
	10.66		11		2.5			2	17.66			1	
	|------Block 1 ---------|--------Block 2--------|--Block 3----|							|-B4-|	|----------Block 5----------|---Block 6--|				

Table 2: Example timeline (in minutes) of viewing behavior in a session

In the first row, 'light gray shaded boxes' denote *Ad Time*, 'white shaded boxes' denote *Content Time*, and the 'dark gray shaded box' denotes *Filler Content Time*. In the second row, 'white shaded dashed line boxes' denote *Session Time*, and the 'black shaded boxes' indicate the beginning of the next episode. All values are in minutes.

Example	**Expression:** $\sum_{i=1}^{n_e} \mathbb{1}\left\{\begin{matrix} Content\ Length_i - 5\ mins \leq\ Content\ Time_i \leq \\ Session\ Time_i - Ad\ Time_i \end{matrix}\right\}$		**Bingeability**
	No Skipping: $Content\ Length_i - 5\ min \leq\ Content\ Time_i$	No Excessive Fast-forwarding: $Content\ Time_i \leq Session\ Time_i - Ad\ Time_i$	
24 min episode of Aquarion	Episode 1: $24 - 5 \leq 22$ Episode 2: $24 - 5 \leq 21$	Episode 1: $22 \leq\ 24.16 - 1.66$ Episode 2: $21 \nleq 20.66 - 1.16$	1

Table 3: Computation of the Bingeability Metric

Example	**Expression:** $\sum_{j=1}^{n_p}(PodDuration_j + ConEnd_j - (CalPod_j - PodDuration_{j-1}))$	**Ad Tolerance** (minutes)
24 min episode of *Aquarion*	• Pod 1: $0.66 + (10 + 10 + 2 + 2 + 17 + 0) - (0 - 0) = 41.66$ • Pod 2: $0.50 + (10 + 2 + 2 + 17 + 0) - (10.66 - 0.66) = 21.5$ • Pod 3: $0.50 + (2 + 2 + 17 + 0) - (11 - 0.50) = 11$ • Pod 4: $0.66 + (17 + 0) - (4.50 - 0.50) = 13.66$ • Pod 5: $0.50 + 0 - (21 - 0.66) = -19.84$	67.98

Table 4: Computation of the Ad Tolerance Metric

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Viewers	6,157	
TV shows	558	
Sessions	110,500	
	Bingeability (count)	Ad Tolerance (minutes)
Min	0	-412.17
2.5%	0	-24.27
Median	1	23.62
97.5%	5	1178.22
Max	57	63,449.10

Table 5: Metric Summary Statistics

Current Variables	**No. of features**	**Description**
Viewer ID	6157	Viewer Fixed Effects
Show name	558	Show Fixed Effects
Genre	18	Genre Fixed Effects
Month	5	Month Fixed Effects (Feb, Mar, Apr, May, Jun)
Week	5	Week Fixed Effects {1 (Day 1 to 7), 2 (Day 8 to 14), 3 (Day 15 to 21), 4 (Day 22 to 28), 5 (Day 29 to 31)}
Day	2	Day Fixed Effects (Weekend and Weekday)
Time of Day (cf. Schweidel & Moe, 2016)	5	Time of Day Fixed Effects (Early morning: 7–10am, Day Time: 10am–5pm, Early Fringe: 5pm – 8pm; Prime Time: 8pm – 11pm, Late Fringe: 11pm – 7am)
First Episode Length	1	Episode length of the first episode seen in the session
Number of episodes of the TV show ahead (remaining) in sequence (N_1)	1	Season Number and Episode Number of all the episodes of a TV show establish a chronological order
Number of potentially unwatched episodes of the TV show during our sample period (N_2)	1	Subtracting the number of episodes viewed till date (during the sample period) from the total episodes available in the dataset

Table 6a: Current Predictors

Electronic copy available at: https://ssrn.com/abstract=3510785
Electronic copy available at: https://ssrn.com/abstract=3498039

Functions	No. of features	Description
Bingeability Sum (Show, Day, Time of Day)	8	Sum of (historical) Bingeability of the viewer for that **Show** over that **Day** at that **Time of day**
Bingeability Indicator (Show, Day, Time of Day)	8	Indicator of whether the viewer has *Bingeability Sum* > 0 for that **Show** over that **Day** at that **Time of day**
Bingeability Session Count (Show, Day, Time of Day)	8	Sum of the number of sessions of the viewer over which Bingeability > 0 for that **Show** over that **Day** at that **Time of day**
Episode Count Sum (Show, Day, Time of Day)	8	Sum of the number of episodes viewed (even partially) by the viewer for that **Show** over that **Day** at that **Time of day**
Episode Session Count (Show, Day, Time of Day)	8	Sum of the number of sessions of the viewer over which Episode Count > 0 for that **Show** over that **Day** at that **Time of day**
Genre Session Count (Day, Time of Day)	4	Sum of the number of sessions over which the viewer has seen that genre over that **Day** at that **Time of day**
Episode Revert[29] Count (Show, Day, Time of Day)	8	Sum of the number of times the viewer reverts to an episode that has been watched in the same session for that **Show** over that **Day** at that **Time of day**
Filler Content Count (Show, Day, Time of Day)	8	Sum of the number of filler content episodes (<15 mins in length) viewed (even partially) by the viewer while watching that **Show** over that **Day** at that **Time of day**
Episode Length (Show, Day, Time of Day)	8	Average episode length of the **Show** viewed by a viewer over that **Day** at that **Time of day**

Table 6b: Functions for watching only TV shows

Function	Description
Bingeability Sum (Show, Day, Time of Day)	*BS* for 'House' over 'Weekend' at 'Day Time'
Bingeability Sum (__ , Day, Time of Day)	*BS* for **any** Show over 'Weekend' at 'Day Time'
Bingeability Sum (Show, __ , Time of Day)	*BS* for 'House' over **any** Day at 'Day Time'
Bingeability Sum (Show, Day, __)	*BS* for 'House' over 'Weekend' at **any** Time of Day
Bingeability Sum (__ , __ , Time of Day)	*BS* for **any** Show over **any** Day at 'Day Time'
Bingeability Sum (Show, __ , __)	*BS* for 'House' over **any** Day at **any** Time of Day
Bingeability Sum (__ , Day, __)	*BS* for **any** Show over 'Weekend' at **any** Time of Day
Bingeability Sum (__ , __ , __)	*BS* for **any** Show over **any** Day at **any** Time of Day

Table 6c: Eight features of Bingeability Sum (BS)

[29] Episode Reversion is when, after finishing a few episodes, a viewer starts watching the next episode, but decides to go back and see an episode already seen while staying in the same session. This is different from the more common behavior of rewinding content while watching an episode.

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

Functions	No. of features	Description
Clicks (Title, Day, Time of Day)	8	Sum of ad clicks by the viewer for that **Title** over that **Day** at that **Time of day**
Ad Proportion[30] (Title, Day, Time of Day)	8	Average ad proportion (over all sessions) for the viewer for that **Title** over that **Day** at that **Time of day**
Pod Count (Pod Length, Title, Day, Time of Day)	32	Sum of number of pods of length **Pod Length** shown to the viewer for that **Title** over that **Day** at that **Time of day**
Pod Session Count (Pod Length, Title, Day, Time of Day)	32	Sum of number of sessions where the viewer was exposed to a given **Pod Length** for that **Title** over that **Day** at that **Time of day**
Ad Diversity (Title, Day, Time of Day)	8	Average % of unique ads per session (in which ads are shown) for the viewer for that **Title** over that **Day** at that **Time of day**
Pod End[31] (Title, Day, Time of Day)	8	Sum of the number of times a viewer ends a pod before it is finished for that **Title** over that **Day** at that **Time of Day**
Calendar Time Spent (Title, Day, Time of Day)	8	Sum of calendar time (session time) spent watching that **Title** over that **Day** at that **Time of day**
Time Between Sessions (Title, Day, Time of Day)	8	Average time between sessions for the viewer for that **Title** over that **Day** at that **Time of day**
Ad Tolerance Sum (Show, Day, Time of Day)	8	Sum of (historical) Ad Tolerance of the viewer for that **Show** over that **Day** at that **Time of day**
Positive Ad Tolerance Indicator (Show, Day, Time of Day)	8	Indicator of whether the viewer has *Ad Tolerance Sum* > 0 for that **Show** over that **Day** at that **Time of day**
Positive Ad Tolerance Session Count (Show, Day, Time of Day)	8	Sum of the number of sessions of the viewer over which Ad Tolerance > 0 for that **Show** over that **Day** at that **Time of day**

Table 6d: Functions for watching TV shows or Movies

[30] Ad Proportion = Ad Time / (Ad Time + Content Time)

[31] A viewer can end a pod (not completely watch it) under a few situations by either ending the session or refreshing the browser or skipping the episode. For a pod to be classified as "ended," we consider all cases where the viewer watches less than 5 seconds of the Pod Length as a case of Pod End.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Viewers	5,760	
TV shows	508	
Sessions	105,610	
	Bingeability (count)	Ad Tolerance (minutes)
Min	0	-412.17
2.5%	0	-24.39
Median	1	23.69
97.5%	5	1182.34
Max	57	63,449.10

Table 7: Summary statistics for dataset used in the model

Rank	Predictor / Function	Type	No. of Features	Gain%
1	Bingeability Sum	Past Predictor	8	18.01
2	Number of episodes ahead in sequence (N_1)	Current Predictor	1	17.64
3	Viewer ID	Current Predictor	300	13.09
4	$\widehat{SR}, \widehat{DR}, \widehat{LR}, \widehat{CR}$	Ad Targeting Rules	4	9.04
5	Show name	Current Predictor	102	7.13
6	First Episode Length	Current Predictor	1	6.16
7	Ad Tolerance Sum	Past Predictor	8	5.44
8	Episode Session Count	Past Predictor	6	3.29
9	Genre	Current Predictor	8	3.22
10	Ad Diversity	Past Predictor	7	2.99

Table 8a: Top 10 sets of predictors of Bingeability

Rank	Predictor / Function	Type	No. of Features	Gain%
1	Viewer ID	Current Predictor	107	31.03
2	Ad Tolerance Sum	Past Predictor	8	10.26
3	Pod End	Past Predictor	8	9.24
4	$\widehat{SR}, \widehat{DR}, \widehat{LR}, \widehat{CR}$	Ad Targeting Rules	4	8.18
5	Bingeability Sum	Past Predictor	8	6.93
6	Number of episodes ahead in sequence (N_1)	Current Predictor	1	6.91
7	Ad Diversity	Past Predictor	8	4.48
8	Episode Count	Past Predictor	8	3.64
9	Ad Proportion	Past Predictor	8	2.87
10	Pod Count	Past Predictor	25	2.59

Table 8b: Top 10 sets of predictors for Ad Tolerance

Electronic copy available at: https://ssrn.com/abstract=3510785
Electronic copy available at: https://ssrn.com/abstract=3498039

	Bingeability	Ad Tolerance
Spacing Rule $\widehat{SR}$	2.04	4.26
Length Rule $\widehat{LR}$	1.43	1.47
Diversity Rule $\widehat{DR}$	0.77	1.38
Clumpiness Rule $\widehat{CR}$	4.79	1.06

Table 8c: Gain % of the Ad Targeting Rules

Set	**Prediction**	**Holdout 1 % of predic-tions**	**Holdout 2 % of predic-tions**	**Recommendation**
A	Bingeability $< \boldsymbol{T}$	0%	0%	Do not show ads
B	Bingeability $\geq \boldsymbol{T}$ & Ad Tolerance ≤ 0	6.3%	3.3%	Show pods at an in-terval of a quarter of the episode length
C	Bingeability $\geq \boldsymbol{T}$ & Ad Tolerance > 0	93.7%	96.7%	Solve Optimization

Table 9: Recommendation Summary for Threshold $T = 0$

Bingeability Threshold (T)	**Holdout 1 Future sessions**	**Holdout 2 New Viewers**
9	-99.7%	-99.9%
5	-98.9%	-99.6%
2	-82.3%	-82.8%
1.6	-71.2%	-67.2%
1	-27.4%	-11.7%
0.9	-13.8%	3.6%
0.8	1.3%	19.3%
0	71.2%	79.1%

Table 10a: Percentage change in optimized ad exposure $(\tilde{n})$ *as compared to observed ad exposure* (n)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Bingeability Threshold (T)	Holdout 1 Future sessions			Holdout 2 New Viewers		
	Sessions with ads ($\hat{b} \geq T \,\&\, \hat{b} \geq \tilde{s}$)	Sessions without ads ($\hat{b} < T \mid \hat{b} < \tilde{s}$)	Net effect	Sessions with ads ($\hat{b} \geq T \,\&\, \hat{b} \geq \tilde{s}$)	Sessions without ads ($\hat{b} < T \mid \hat{b} < \tilde{s}$)	Net effect
9	-11.75%	2.31%	2.23%	42.92%	-1.65%	-1.58%
5	3.18%	2.50%	2.51%	-7.93%	-1.35%	-1.41%
2	5.76%	3.26%	3.64%	-5.66%	0.07%	-0.87%
1.6	6.85%	3.07%	3.97%	-5.93%	1.16%	-0.80%
1	2.07%	3.54%	2.72%	-7.71%	5.11%	-3.19%
0.9	2.42%	4.29%	3.05%	-5.82%	3.97%	-3.33%
0.8	2.93%	5.60%	3.55%	-5.37%	6.93%	-3.43%
0	5.31%	222.85%	5.33%	-2.86%	-59.91%	-2.88%

Table 10b: Percentage change in optimized Bingeability ($\tilde{b}$) as compared to observed Bingeability (b)

Bingeability Threshold (T)	Holdout 1 Future sessions			Holdout 2 New Viewers		
	Sessions with ads ($\hat{b} \geq T \,\&\, \hat{b} \geq \tilde{s}$)	Sessions without ads ($\hat{b} < T \mid \hat{b} < \tilde{s}$)	Net effect	Sessions with ads ($\hat{b} \geq T \,\&\, \hat{b} \geq \tilde{s}$)	Sessions without ads ($\hat{b} < T \mid \hat{b} < \tilde{s}$)	Net effect
9	-17.35%	2.19%	2.07%	-14.24	0.26%	0.23%
5	-6.21%	2.54%	2.35%	-12.08%	0.52%	0.40%
2	1.06%	3.94%	3.48%	-3.13%	1.74%	0.95%
1.6	2.29%	4.30%	3.80%	-2.32%	2.27%	1.03%
1	0.02%	5.92%	2.56%	-4.89%	4.76%	-1.41%
0.9	0.62%	7.57%	2.89%	-4.24%	6.37%	-1.55%
0.8	1.25%	11.03%	3.38%	-3.68%	9.22%	-1.66%
0	5.16%	61.19%	5.17%	-1.10%	32.76%	-1.09%

Table 10c: Percentage change in optimized Bingeability ($\tilde{b}$) as compared to initial predicted Bingeability ($\hat{b}$)

Electronic copy available at: https://ssrn.com/abstract=3510785
Electronic copy available at: https://ssrn.com/abstract=3498039

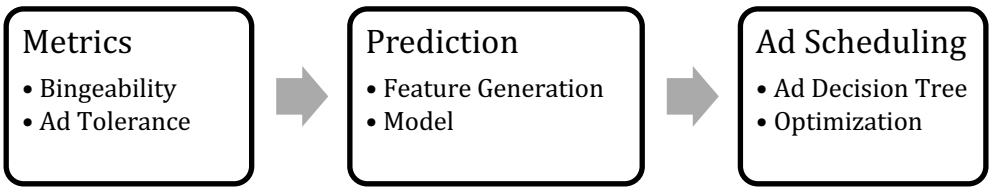


Figure 1: Three-Stage Architecture

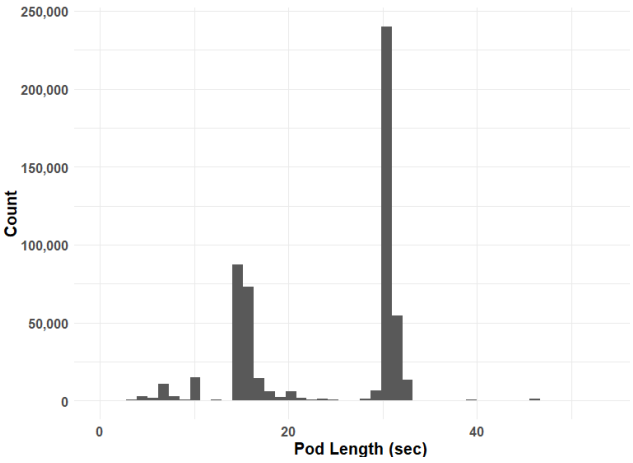


Figure 2a: Histogram of Pod Length (0^{th} to 97.5^{th} percentile)

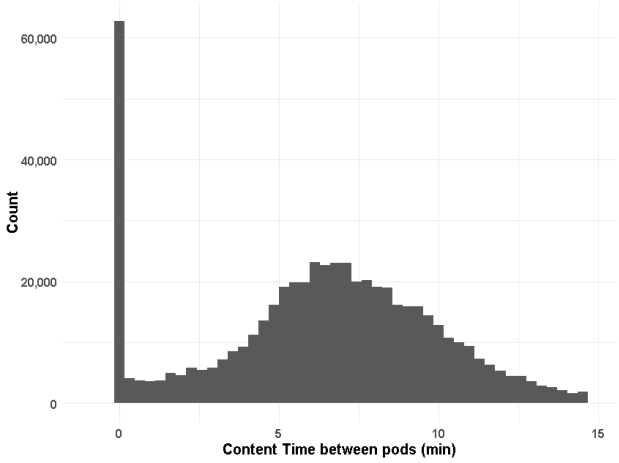


Figure 2b: Histogram of Pod Spacing (min) (0^{th} to 97.5^{th} percentile)

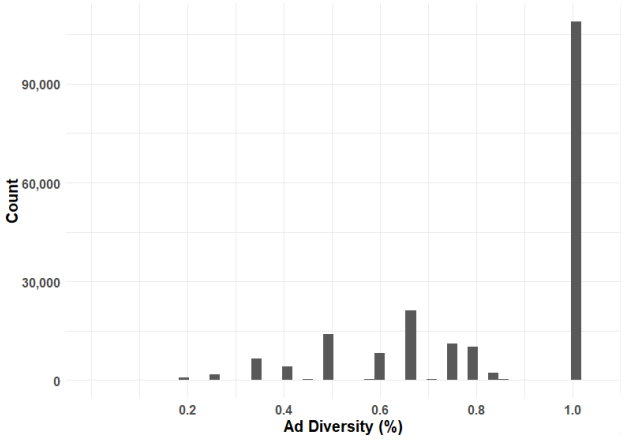


Figure 2c: Histogram of Ad Diversity (%) (0^{th} to 100^{th} percentile)

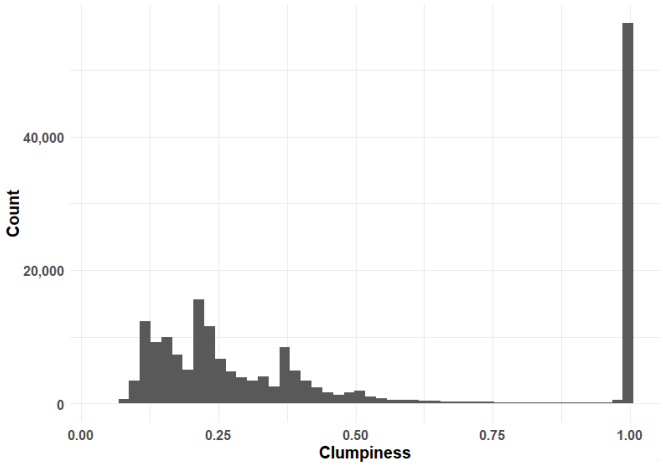


Figure 2d: Histogram of Clumpiness (0^{th} to 100^{th} percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

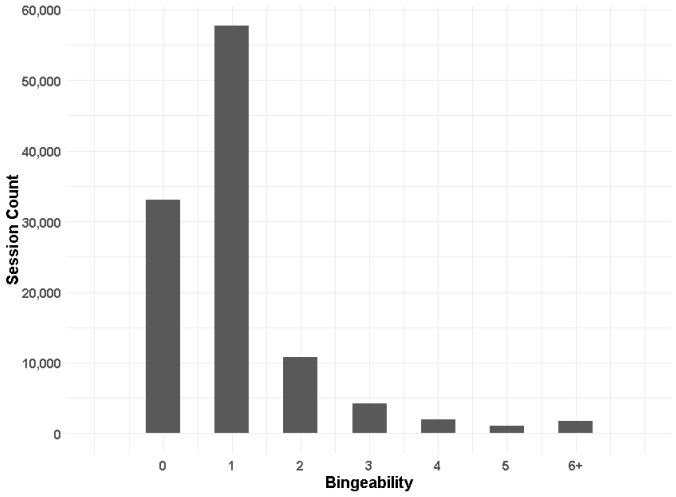


Figure 3a: Histogram of Bingeability

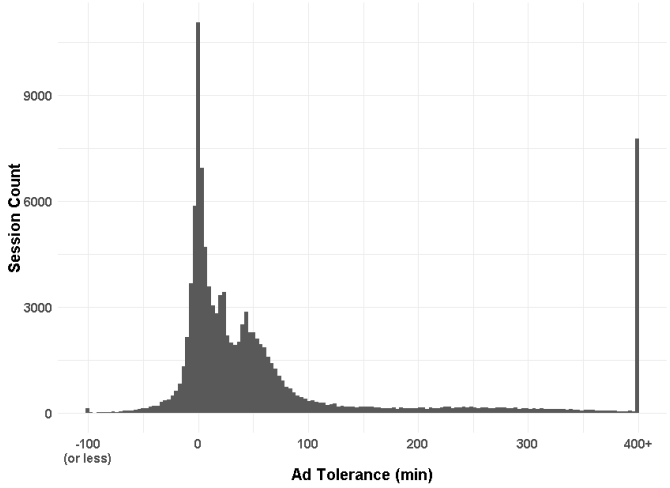


Figure 3b: Histogram of Ad Tolerance

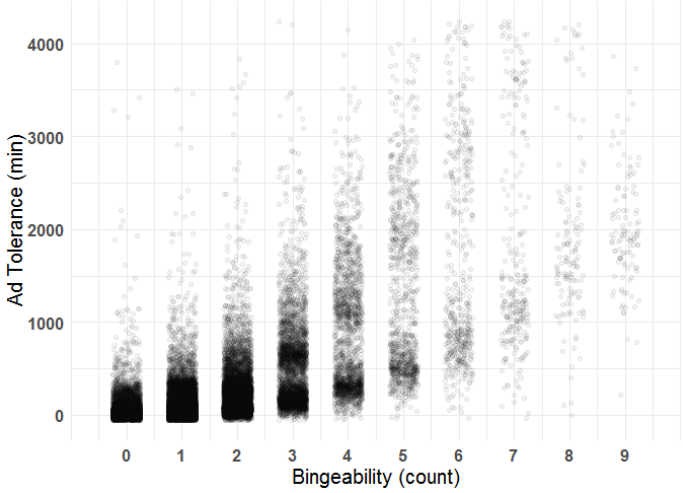


Figure 3c: Ad Tolerance vs Bingeability (0.5th to 99.5th percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

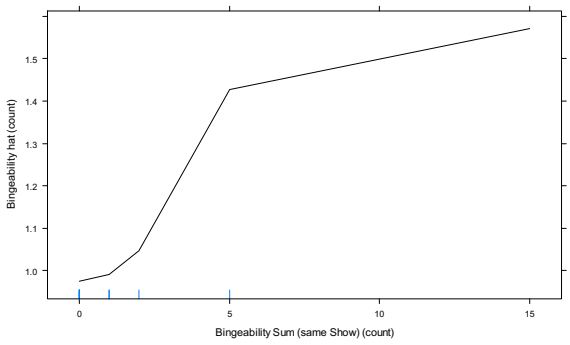


Figure 4a:
Partial Dependence of Bingeability on Bingeability Sum
(same Show, any Day, any TOD)
(2.5th to 97.5th percentile)

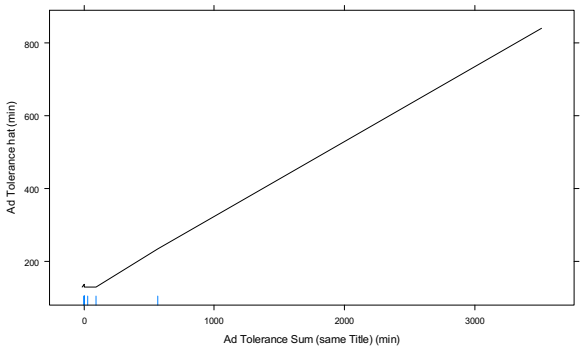


Figure 4b:
Partial Dependence of Ad Tolerance on Ad Tolerance Sum
(same Title, any Day, any TOD)
(2.5th to 97.5th percentile)

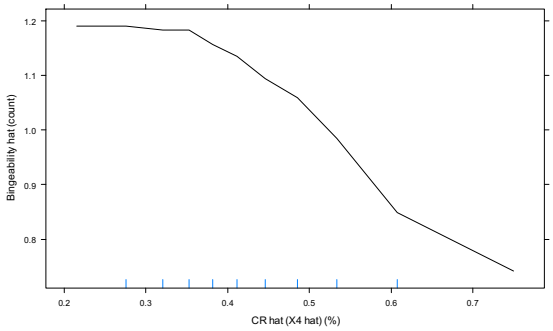


Figure 4c:
Partial Dependence of $\widehat{CR}$ on Bingeability
(2.5th to 97.5th percentile)

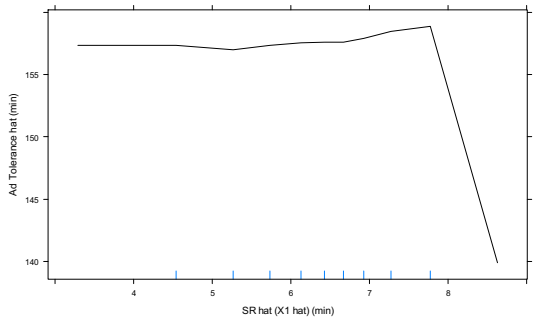


Figure 4d:
Partial Dependence of $\widehat{SR}$ on Ad Tolerance
(2.5th to 97.5th percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

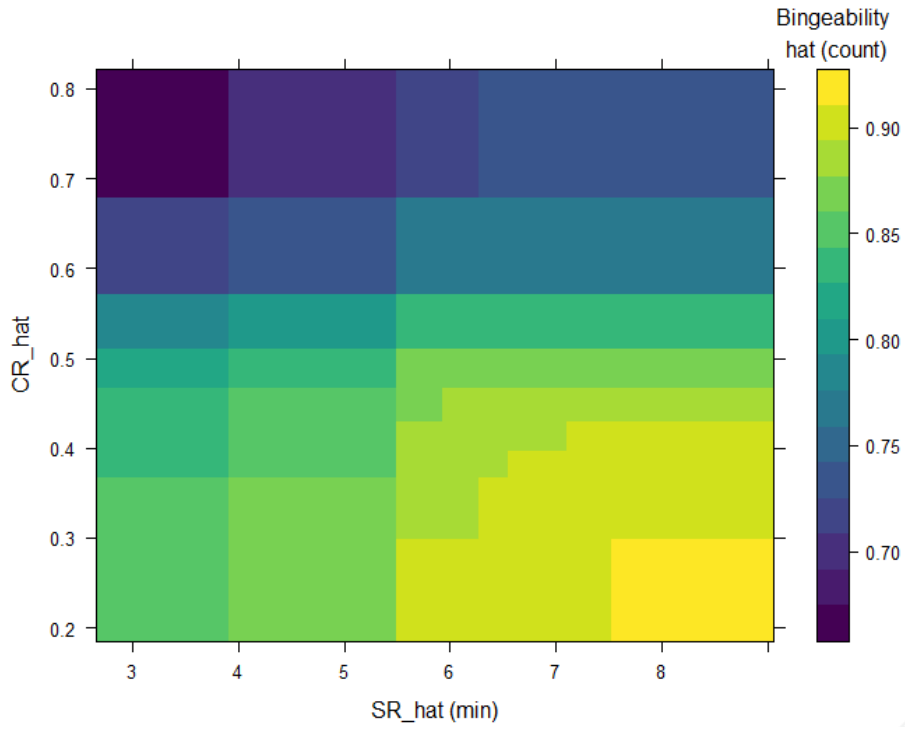


Figure 4e: Partial Dependence of Bingeability on its two most important Ad Targeting Rules, X_{s1} (2.5^{th} to 97.5^{th} percentile)

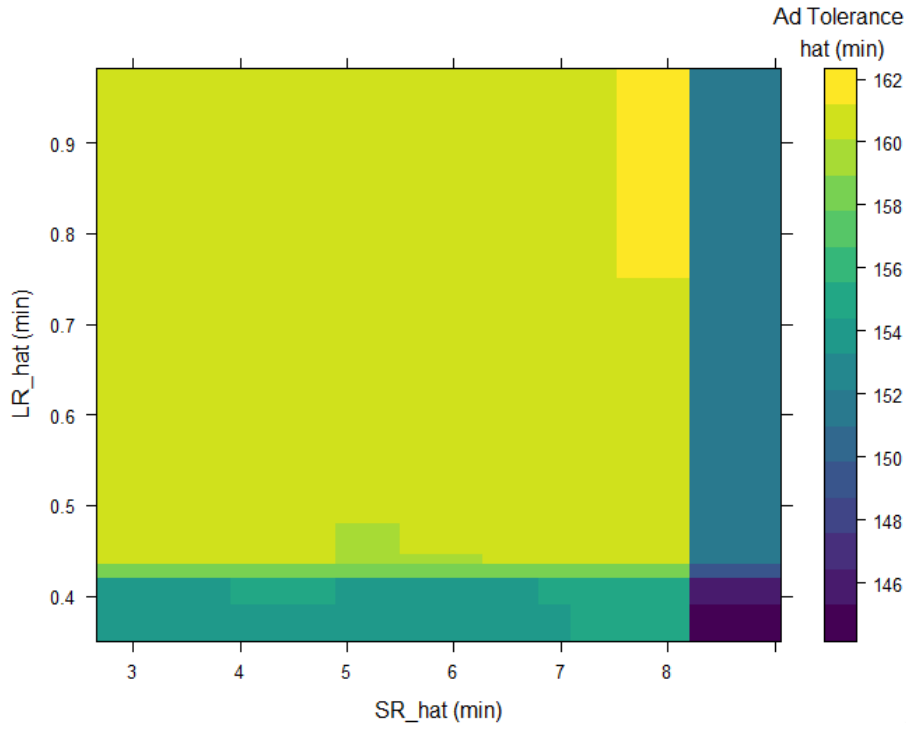


Figure 4f: Partial Dependence of Ad Tolerance on its two most important Ad Targeting Rules, X_{s2} (2.5^{th} to 97.5^{th} percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

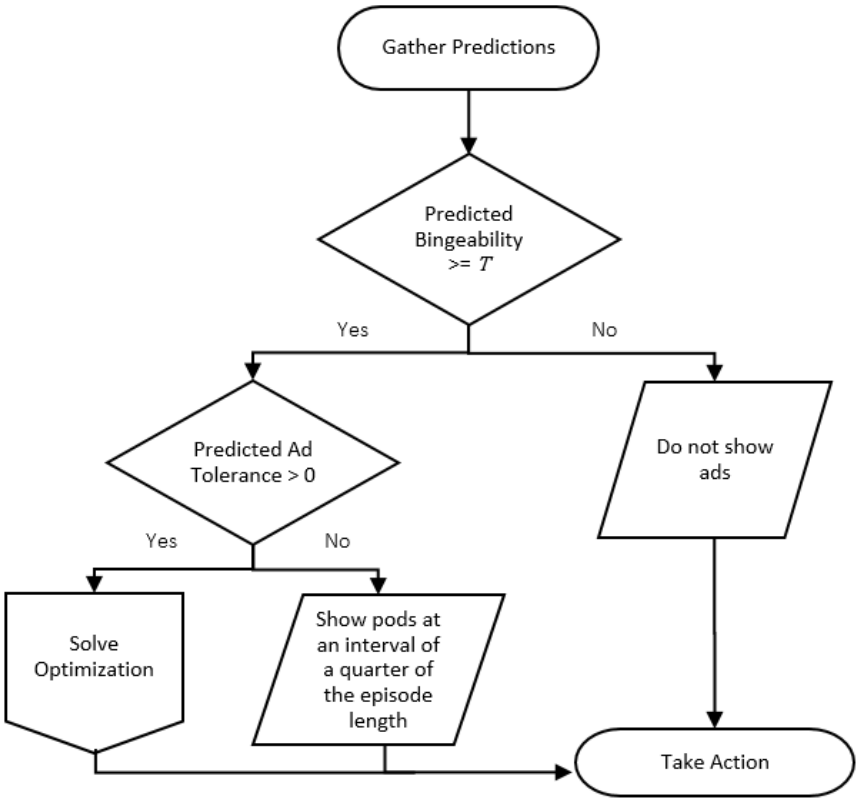


Figure 5: Ad Decision Tree

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

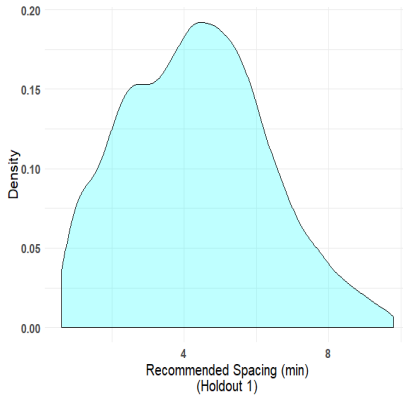


Figure 6a: Density of Recommended Spacing Holdout 1 (2.5th to 97.5th percentile)

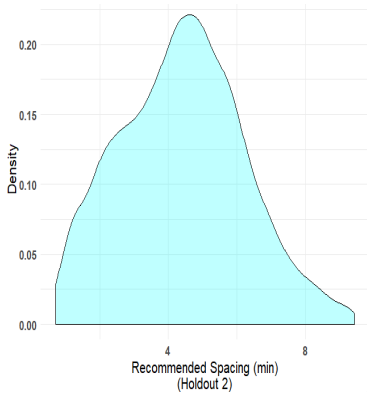


Figure 6b: Density of Recommended Spacing Holdout 2 (2.5th to 97.5th percentile)

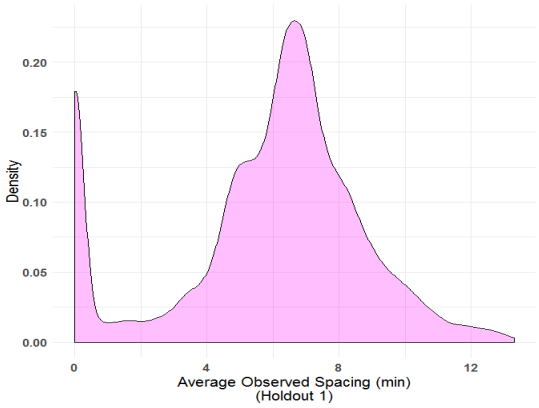


Figure 7a: Density of Average Observed Spacing Holdout 1 (2.5th to 97.5th percentile)

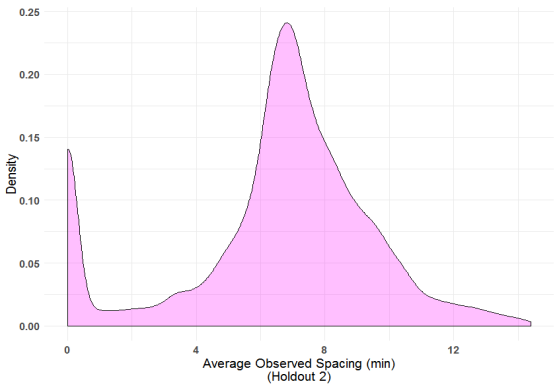


Figure 7b: Density of Average Observed Spacing Holdout 2 (2.5th to 97.5th percentile)

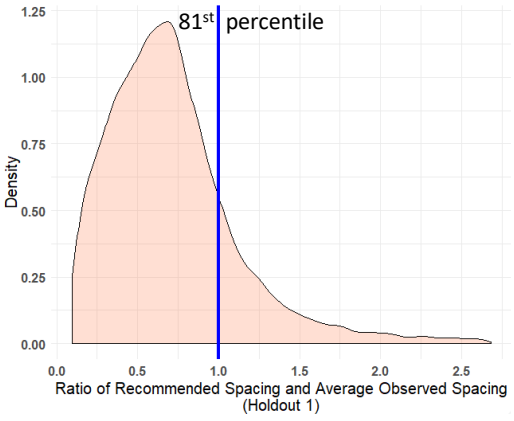


Figure 8a: Density of the Ratio of Recommended Spacing and Average Observed Spacing Holdout 1 (2.5th to 97.5th percentile)

Figure 8b: Density of the Ratio of Recommended Spacing and Average Observed Spacing Holdout 2 (2.5th to 97.5th percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

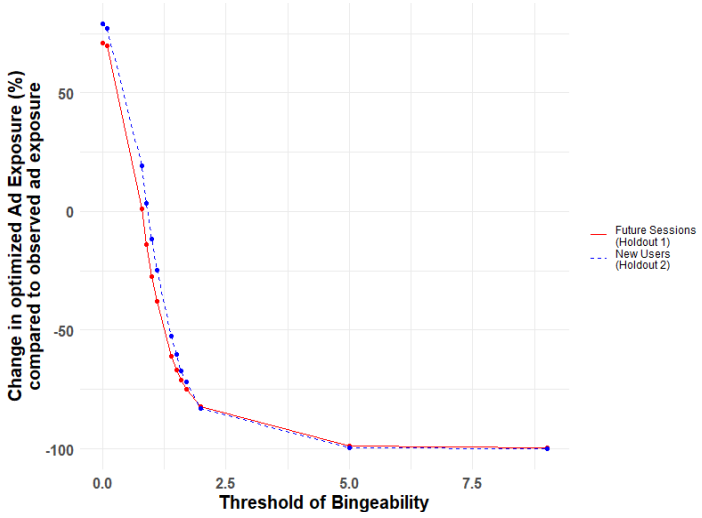


Figure 9a:
Percentage change in optimized ad exposure ($\tilde{n}$) as compared to observed ad exposure (n)

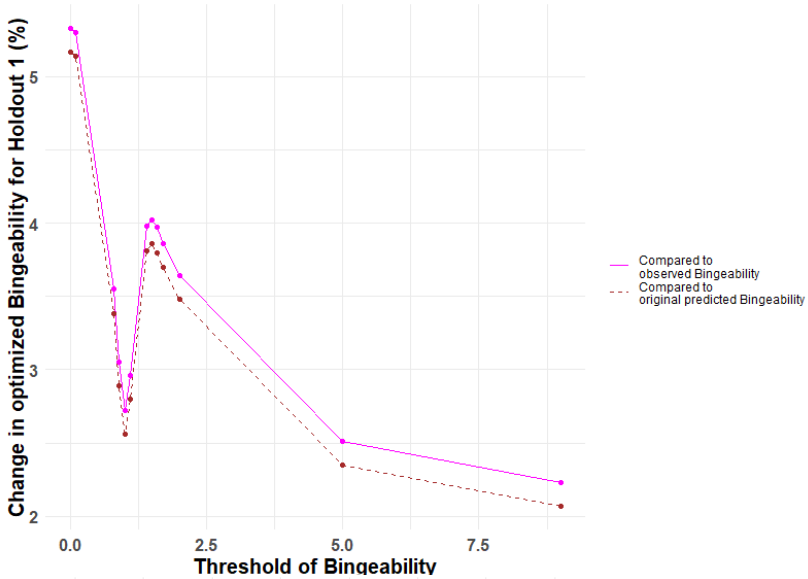


Figure 9b:
Percentage change in optimized Bingeability ($\tilde{b}$) for Holdout 1

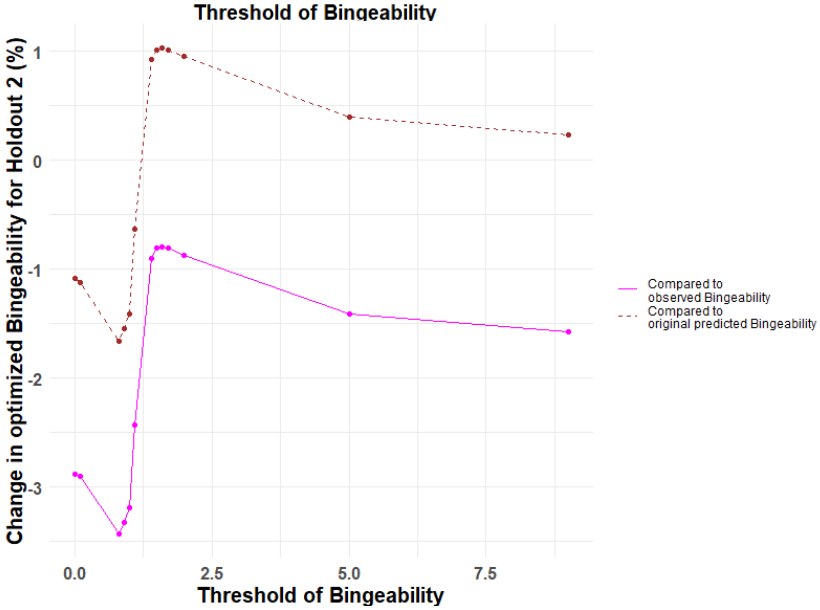


Figure 9c:
Percentage change in optimized Bingeability ($\tilde{b}$) for Holdout 2

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX A: DETAILS ON BINGEABILITY AND AD TOLERANCE

We present three more illustrative examples of session viewing behavior and application of the Bingeability and Ad Tolerance metric. In the first row of each illustration, 'light gray shaded boxes' denote *Ad Time*, and 'white shaded boxes' denote *Content Time*. In the second row of each illustration, 'white shaded dashed line boxes' denote *Session Time*, and the 'black shaded boxes' (in Example C) indicate the beginning of the next episode. All values are in minutes.

Example A

<table>
<tr><td>A
23 min episode of Family Guy</td><td colspan="2">13 | 0.50 | 6 | 0.50 | 2
13 | 6.5 | 2.50
|----------Block 1----------|------Block 2 -------|----Block 3-----|</td></tr>
<tr><td rowspan="2">Bingeability: 1</td><td>No Skipping:
$Content\ Length_i - 5\ \text{min} \leq Content\ Time_i$</td><td>No Excessive Fast-forwarding:
$Content\ Time_i \leq Session\ Time_i - Ad\ Time_i$</td></tr>
<tr><td>Episode 1: $23 - 5 \leq 21$</td><td>Episode 1: $21 \leq 22 - 1$</td></tr>
<tr><td>Ad Tolerance: -8 min</td><td>$\sum_{j=1}^{n_p}(PodDuration_j + ConEnd_j - (CalPod_j - PodDuration_{j-1}))$</td><td>Pod 1: $0.5 + (6 + 2) - (13 - 0) = -4.5$
Pod 2: $0.5 + 2 - (6.5 - 0.5) = -3.5$</td></tr>
</table>

Example A shows the behavior of a viewer watching one 23-minute episode of 'Family Guy'. The first row shows blocks of time spent watching content (in white) and ads (in light gray). The viewer's viewing experience was interrupted by two ads that were 0.50 minutes long. The first ad was shown after the viewer viewed 13 minutes of content, and the second ad was shown after the viewer viewed 6 additional minutes of content. After the last ad, the viewer viewed 2 more minutes of content and the session ended. The second row denotes the calendar time spent corresponding to the blocks of time in the first row. In Example A, the calendar time spent in each block is equal to the sum of content time and ad time in the corresponding block. By substituting the values of Example A in equation (1), we get,

$$\overbrace{Session\ Time}^{22\ \text{minutes}} =$$

$$\overbrace{Content\ Time}^{21\ \text{minutes}} + \overbrace{Ad\ Time}^{1\ \text{minute}} + \overbrace{Filler\ Content\ Time}^{0\ \text{minutes}} + \overbrace{Pauses - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

As the value of the measured variable on the Left-Hand-Side of the above equation is the same as sum of the measured variables on the Right-Hand-Side of the equation, the sum of the unmeasured variables is 0 minutes.

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

Second, both the conditions of the Bingeability metric are satisfied. As we see no evidence of skipping or excessive fast-forwarding behavior, the value of Bingeability is 1. Third, we discuss the construction of the Ad Tolerance metric in detail for Example A. We begin by adding the duration of the first pod which is 0.5 minutes to the amount of content viewed in the remainder of the session (after the end of the pod), which is $6 + 2 = 8$ minutes. We then subtract the calendar time that has elapsed since the beginning of the session, $CalPod_j$, which is 13 minutes. As there was no pod before this, we have a null value for $PodDuration_{j-1}$. Thus, the total value of the metric for the first pod is -4.5 minutes. Now, we repeat the same process for the second pod which is also 0.5 minutes in duration. To this we add the content time viewed in the remainder of the session which is 2 minutes. We then subtract the difference between the calendar time elapsed since the beginning of the previous pod and the duration of the previous pod, which is $6.5 - 0.5 = 6$. Thus, the total value of the metric for the second pod is -3.5 minutes. On summing up the values corresponding to each pod, we get a total Ad Tolerance value of $-4.5 - 3.5 = -8$ minutes. A negative value of Ad Tolerance suggests that the viewer ended a session after exposure to a commercial pod which was preceded (at some point) by a long period of no ad exposure. This is true in Example A where content time between pods (or the period of no ad exposure) was initially large at 13 minutes, and then reduced to 6 minutes, followed by 2 minutes.

Example B

<table>
<tr><td rowspan="2">B
43 min episode of Chuck</td><td colspan="2"><table>
<tr><td>3</td><td>0.5</td><td>6</td><td>0.25</td><td>8</td><td>0.25</td><td>9</td><td>0.25</td><td>6</td><td>0.25</td><td>7</td><td>0.25</td><td>2</td></tr>
<tr><td colspan="2">7</td><td colspan="3">6.5</td><td colspan="3">8.25</td><td colspan="3">10</td><td colspan="2">5</td><td colspan="2">7.25</td><td colspan="2">10</td></tr>
</table>
|-B1-|------Block 2 ----|--------Block 3--------|--------Block 4----------|--------Block 5-----|---------Block 6-----|----Block 7----|</td></tr>
<tr><td rowspan="2">Bingeability: 1</td><td>No Skipping:
$Content\ Length_i - 5\ \text{min} \leq Content\ Time_i$</td><td>No Excessive Fast-forwarding:
$Content\ Time_i \leq Session\ Time_i - Ad\ Time_i$</td></tr>
<tr><td>Episode 1: $43 - 5 \leq 41$</td><td>Episode 1: $41 \leq 54 - 1.75$</td></tr>
<tr><td>Ad Tolerance: 74.25 min</td><td>$\sum_{j=1}^{n_p}(PodDuration_j + ConEnd_j - (CalPod_j - PodDuration_{j-1}))$</td><td>Pod 1: $0.5 + (6 + 8 + 9 + 4.75 + 7 + 2) - (7 - 0) = 30.25$
Pod 2: $0.25 + (8 + 9 + 4.75 + 7 + 2) - (6.5 - 0.5) = 25$
Pod 3: $0.25 + (9 + 4.75 + 7 + 2) - (8.25 - 0.25) = 15$
Pod 4: $0.25 + (4.75 + 7 + 2) - (10 - 0.25) = 4.25$
Pod 5: $0.25 + (7 + 2) - (5 - 0.25) = 4.5$
Pod 6: $0.25 + 2 - (7.25 - .25) = -4.75$</td></tr>
</table>

Example B shows the behavior of a viewer watching one 43-minute episode of 'Chuck'. The viewer's viewing experience was interrupted by 6 ads shown in the light gray shaded boxes. The content

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3496039

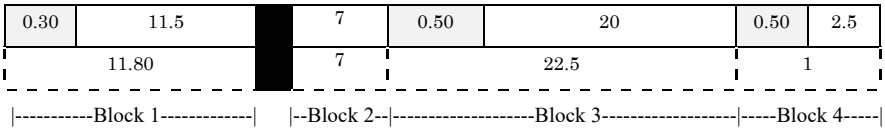
time spent in the first block is 3 minutes, but the calendar time is 7 minutes. A higher value of calendar time suggests that time was spent in pauses or rewinds in this block. This is similarly observed in block 4 and block 7. In block 5, the calendar time spent is 5 minutes, which is less than the sum of ad time and content time (totaling 6.25 minutes) in the corresponding block. A lower value of calendar time suggests that time was spent in fast-forwards in this block. By substituting the values of Example B in equation (1), we get,

$$\overbrace{Session\ Time}^{54\ \text{minutes}} =$$

$$\overbrace{Content\ Time}^{41\ \text{minutes}} + \overbrace{Ad\ Time}^{1.75\ \text{minutes}} + \overbrace{Filler\ Content\ Time}^{0\ \text{minutes}} + \overbrace{Pauses\ - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

On solving the above equation, we find that the sum of the unmeasured variables is 11.25 minutes. This indicates that more time was spent in pauses or rewinds than in fast-forwards in this session. Second, both the conditions of the Bingeability metric are satisfied. As we see no evidence of skipping or excessive fast-forwarding behavior, the value of Bingeability is 1. Third, we adopt a similar process to calculate Ad Tolerance as done in Example A. It is important to note the use of 'Caveat 1' in block 5 of Example B where there is evidence of fast-forwarding behavior: $ConEnd_j$ is chosen as *Session Time – Ad Time,* $5 - 0.25 = 4.75$ minutes, because it is less than *Content Time* of 6 minutes. We get a total Ad Tolerance value of 74.25 minutes.

Example C

<table>
<tr><td>C
45 min episode of Rescue Me</td><td colspan="2"> 
</td></tr>
<tr><td rowspan="2">Bingeability: 1</td><td>No Skipping:
$Content\ Length_i - 5\ \text{min} \leq Content\ Time_i$</td><td>No Excessive Fast-forwarding:
$Content\ Time_i \leq Session\ Time_i - Ad\ Time_i$</td></tr>
<tr><td>Episode 1: $45 - 5 \not\leq 11.5$
Episode 2: $45 - 5 \not\leq 29.5$</td><td>Episode 1: $11.5 \leq 11.80 - 0.30$
Episode 2: $29.5 \leq 30.5 - 1$</td></tr>
<tr><td>Ad Tolerance: 20.80 min</td><td>$\sum_{j=1}^{n_p}(PodDuration_j + ConEnd_j - (CalPod_j - PodDuration_{j-1}))$</td><td>Pod 1: $0.30 + (11.5 + 7 + 20 + 0.5) - (0 - 0) = 39.30$
Pod 2: $0.50 + (20 + 0.50) - (18.80 - 0.30) = 2.5$
Pod 3: $0.50 + 0.50 - (22.50 - 0.50) = -21$</td></tr>
</table>

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Example C shows the behavior of a viewer watching two 45-minute episodes of ‘Rescue Me’. However, the viewer watches only 11.5 minutes of the first episode and 29.5 minutes of the second episode. There is also evidence of pauses or rewind in block 3 and fast-forwarding in block 4 because there is mismatch between the calendar time spent and the sum of content time and ad time in those blocks. It is important to note that evidence of fast-forwarding behavior in block 4 could be for the content that was rewound in block 3. This is because each block in the illustration does not denote unique content being viewed due to possible rewinds and fast-forwards by the viewer. By substituting the values of Example C in equation (1), we get,

$$\overbrace{Session\ Time}^{42.3\ \text{minutes}} =$$

$$\overbrace{Content\ Time}^{41\ \text{minutes}} + \overbrace{Ad\ Time}^{1.30\ \text{minutes}} + \overbrace{Filler\ Content\ Time}^{0\ \text{minutes}} + \overbrace{Pauses\ - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

On solving the above equation, we find that the sum of the unmeasured variables is 0 minutes, but as mentioned earlier we find definite evidence of fast-forwards, and pauses or rewinds. Second, the first condition (no skipping) of the Bingeablity metric is not satisfied in both Episode 1 and 2. As none of the episodes in the session were viewed completely, the value of Bingeability is 0. Third, we adopt a similar process to calculate Ad Tolerance as done in Example A. We also use ‘Caveat 1’ in block 4 where there is evidence of fast-forwarding behavior. We get a total Ad Tolerance value of 20.80 minutes.

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX B: HULU DATA COLLECTION METHODOLOGY

In the raw data, a 'playback ping' from the Hulu server records the amount of content viewed since the previous 'playback ping.' Similarly, the 'revenue ping' records the amount of ad viewed since the previous 'revenue ping.' 'Playback ping' and 'revenue ping' occur at periodic brief intervals and need not be in chronological order with respect to the other. For example, a 'playback ping' could fall in between successive 'revenue pings.' As content cannot be viewed in between an ad, we record the content viewed till this 'playback ping' as occurring before the commencement of that respective block of 'revenue pings.' In situations when the 'playback ping' occurs after the last 'revenue ping' (in a block of consecutive revenue pings), we carry out the following data manipulation: "Calculate the calendar time and content time captured between these two pings, and then take the difference between the two. If the difference is negative, we add the absolute value of the difference to the amount of content viewed before the commencement of the ad." Thus, for these brief instances (where the difference is negative) right after the end of an ad, we assume no presence of fast-forwarding behavior because it is less likely. In addition, on 6.7% of the occasions, the amount of ad (pod) watched is registered as greater than the ad (pod) length due to potential errors in the recording of data by the streaming provider. In these cases, we increase the ad (pod) length to match the ad (pod) watched.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX C: USING DIFFERENT WEIGHTS IN THE AD TOLERANCE METRIC

As mentioned in subsection "Metric Development", we had set each of the weights of the three components of the Ad Tolerance metric to 1. We test a few different scenarios using other combinations of weights and analyze their effect on the optimized frequency of ad exposure. This is shown in Table C1. The first scenario assumes that viewers weigh the time spent watching a pod twice as much as the other two components of the metric. The second scenario assumes that viewers weigh the time spent watching content after the end of a pod, half as much as the other two components. The third scenario assumes that viewers weigh the difference between the calendar time elapsed since the beginning of the previous pod and the duration of the previous pod, twice as much as the other two components of the metric. For each scenario, we calculate new values of the Ad Tolerance metric, and update the past predictors in Table 6d that correspond to functions for Ad Tolerance Sum, Positive Ad Tolerance Indicator and Ad Tolerance Session Count. Next, we run the first-stage and second-stage of the model, follow the steps of the Ad Decision Tree and then run the optimization procedure. The recommended spacing for the set of observations in each scenario is compared with the recommend spacing for the corresponding set of observations in the original scenario that had all weights set to 1. The mean absolute difference (MAD) for these comparisons are also shown in Table C1. The low value of MAD ($\leq$ 1 minute) indicates that our optimization process is robust to the choice of values of the weights (in the range considered).

Scenario	**Description**	$\boldsymbol{w_1}$	$\boldsymbol{w_2}$	$\boldsymbol{w_3}$	**MAD (minutes) Holdout 1**	**MAD (minutes) Holdout 2**
1	$\boldsymbol{PodDuration_j}$ is weighed 2 times the other components	2	1	1	0.88	0.73
2	$\boldsymbol{ConEnd_j}$ is weighed 0.5 times the other components	2	1	2	0.96	0.89
3	$\boldsymbol{CalPod_j - PodDuration_{j-1}}$ is weighed 2 times the other components	1	1	2	1.00	0.91

Table C1: Different weight combinations of the components of the Ad Tolerance Metric; MAD (min) on comparing recommended spacing in each scenario with the original recommended spacing

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX D: METRIC VALIDITY

D.1 Bingeability

We apply both the Episode Count metric and the Bingeability metric to our sample and examine the cases of mismatch between them in Table D1. Bingeability is different from Episode Count for 45.4% of the sessions, 89.8% of viewers, 96.2% of TV shows and 94.4% of genres. While the mismatch seems to be frequent, it is mainly a consequence of skipping behavior and not excessive fast-forwarding behavior. Skipping behavior is 26 times more likely than excessive fast-forwarding behavior across all sessions.

	N (Total count)	Skipping (S)	Excessive Fast-Forwarding (FF)	Both ($S \cap FF$)	Total ($S \cup FF$)
Sessions	110,500	45.0%	1.7%	1.3%	45.4%
Viewers	6,157	89.7%	13.6%	13.4%	89.8%
TV shows	558	96.2%	46.0%	46.0%	96.2%
Genre	18	94.4%	88.8%	88.8%	94.4%

Table D1: Evidence of Skipping and Excessive Fast-Forwarding

We show the relationship between the 0.05th and 99.5th percentile range of Bingeability and Episode Count in Figure D1. The darker the color of the square, more are the number of points located there. For example, when Episode Count is 7, there are more instances when Bingeability is 7 than 1.

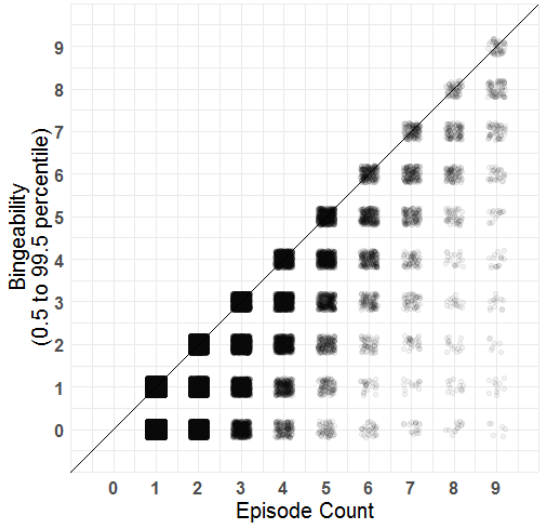


Figure D1: Bingeability versus Episode Count (0.5th to 99.5th percentile of Bingeability)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

Now, we compare the trend in viewership of episodes across all 558 TV shows on a weekly (and monthly) basis using both Episode Count and Bingeability. Trends for both the metrics are compared to check whether they are both increasing, decreasing or constant. We find that the weekly (and monthly) trends in viewership popularity are mismatched 21.1% (and 14.7%) of the time across 72.8% (and 26.3%) of TV shows viewed in our sample. This tells us that inferences about the trend in viewership can change based on the metric one decides to use. By counting episodes which are not completely watched, Episode Count typically overstates the popularity of a TV show. Bingeability quantifies the immersive experience and presents a more conservative estimate of the popularity level. The Bingeability metric by itself can be useful to various streaming platforms, production studios, advertisers and data measurement companies who would like to measure the trend in popularity of TV shows streamed on platforms.

D.2 Ad Tolerance

We check whether the Ad Tolerance metric can capture differences in behavioral consumption patterns. For illustration, consider six sessions of six different viewers in Table D2 where each session is spent viewing 15 minutes of content in addition to ad exposure that is assumed to be randomly delivered. For ease of exposition, we assume there are no instances of fast-forwards, rewinds or pauses in each session.

Session	Illustration	Ad Tolerance (min)	Ad Exposure (min)	Number of Pods
Viewer 1	2.5 | 0.50 | 12.5 |--B1-|----------Block 2 ----------|	10.5	0.50	1
Viewer 2	7.5 | 0.50 | 7.5 |----------Block 1----------|----------Block 2 ----------|	0.5	0.50	1
Viewer 3	12.5 | 0.50 | 2.5 |----------Block 1----------|-----Block 2-----|	-9.5	0.50	1
Viewer 4	2.5 | 0.50 | 2.5 | 0.50 | 2.5 | 0.50 | 7.5 |-B 1--|----Block 2----|----Block 3-----|----------Block 4----------|	26.5	1.50	3

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

Viewer 5	3.75 | 0.50 | 3.75 | 0.50 | 3.75 | 0.50 | 3.75 |---B 1---|-------Block 2------|-------Block 3-------|-------Block 4------|	12.75	1.50	3
Viewer 6	7.5 | 0.50 | 2.5 | 0.50 | 2.5 | 0.50 | 2.5 |-------Block 1--------|----Block 2-----|----Block 3-----|-----Block 4----|	4	1.50	3

Table D2: Examples of viewing behavior in a session

Viewer 1 is exposed to an ad of length 0.5 minutes after viewing 2.5 minutes of content. After the end of the ad, the viewer views 12.5 more minutes of content and the session ends. Viewer 2 is exposed to an ad in the middle of her session while Viewer 3 is exposed to an ad after viewing 12.5 minutes of content. Across the first three sessions, we observe that Viewer 3 had the most time (12.5 min) to adapt to the absence of ads and viewed the least amount of content (2.5 min) after the final ad. Hence, Viewer 3 can be expected to have the lowest Ad Tolerance. Similarly, across the first three sessions, Viewer 1 had the least time (2.5 min) to adapt to the absence of ads and viewed the most content (12.5 min) after it, and hence can be expected to have the highest Ad Tolerance.

In the last three sessions, Viewer 4 is exposed to 3 ads in small intervals in the first half of her session, Viewer 5 is exposed to 3 ads at equally spaced intervals and Viewer 6 is exposed to 3 ads in small intervals in the second half of the session. Across all the six sessions, we observe that Viewer 3 had the most time (12.5 min) to adapt to the absence of ads and was exposed to only one ad in total. Hence, Viewer 3 can be expected to have the lowest Ad Tolerance overall. While both Viewer 1 and Viewer 4 had the least amount of time (2.5 min) to adapt until the first ad, Viewer 4 was exposed to two additional ads and still ended up watching 15 minutes of content in total. Hence Viewer 4 can be expected to be have the highest Ad Tolerance. Overall, we can observe that if ads are bunched together in the beginning of a session, Ad Tolerance is the highest, whereas if the session ends shortly after viewing an ad which was preceded by a long period of no ad exposure, then the Ad Tolerance is the lowest.

We can compare the Ad Tolerance metric for the six sessions with the simple measures of ‘minutes of ad exposure’ and ‘number of pods’ in Table D2. We observe that the simple measures are unable to distinguish between the first three cases or between the last three cases, whereas the Ad Tolerance metric gives us a unique value for each of the six cases. Thus, we showed that Ad Tolerance is able to capture differences in behavioral consumption patterns (assuming randomness in ad delivery) in an intuitive manner, thereby lending further validity to the construction of the metric. Non-randomness in ad delivery is

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

controlled with the help of instrumental variables in our model, which is detailed in the subsection "Model".

Lastly, using our dataset, we conduct a principal component analysis (with varimax rotation) on the two metrics and the 'number of pods', 'minutes of ad exposure' and 'minutes of content viewed'. We choose three factors and present their loadings for each variable in Table D3. We find that the factor loadings for 'number of pods', 'minutes of ad exposure' and 'minutes of content viewed' are very similar and are dominated by the first factor. On the other hand, Bingeability and Ad Tolerance are dominated by the third and second factor respectively. This analysis further demonstrates that the two metrics are capturing different latent constructs.

	Factor 1	Factor 2	Factor 3
Bingeability	0.44	0.30	0.85
Ad Tolerance	0.41	0.87	0.29
Number of pods shown	0.82	0.37	0.38
Minutes of ad exposure	0.81	0.40	0.38
Minutes of content viewed	0.76	0.40	0.43

Table D3: Factor Loadings from a Principal Component Analysis

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX E: MODELLING CORRELATION BETWEEN OUTCOMES

We model correlation between the two outcomes—Bingeability and Ad Tolerance—using the regressor chain approach (Melki et al., 2017). It involves incorporating the predicted value of an outcome as a covariate to predict another outcome which results in the formation of a chain. This can be formalized by modifying equation (7) in subsection "Model" as follows:

$$Y_{1t} = f_2\left(\hat{Y}_{2t}, \hat{X}_{1t}, \hat{X}_{2t}, \hat{X}_{3t}, \hat{X}_{4t}, W_{1t}, W_{2t}\right) + u_t$$
$$Y_{2t} = f_2\left(\hat{Y}_{1t}, \hat{X}_{1t}, \hat{X}_{2t}, \hat{X}_{3t}, \hat{X}_{4t}, W_{1t}, W_{2t}\right) + u_t$$

where $\hat{Y}_{1t}$ and $\hat{Y}_{2t}$ are the predictions from the original model that are added as covariates to predict Y_{2t} and Y_{1t} respectively. Such an approach allows us to capture the correlational influence of one outcome on the other. The final predictions of the outcomes from the regressor chain approach are then used as inputs to the Ad Decision Tree, which is followed by running the optimization procedure. Subsequently we construct a corresponding Decision Support System whose results are shown in Figures E1, E2 and E3 which are analogous to Figures 9a, 9b and 9c.

The graphs show that for future sessions of current viewers (Holdout 1), the platform and its viewers are better-off if the platform uses a threshold of 0 to show ads. This results in a 47.5% increase in ad exposure as compared to observed ad exposure and a 7.19% increase in Bingeability as compared to initial predicted Bingeability (or a 9.77% increase in Bingeability as compared to observed Bingeability). Similarly, for new viewers (Holdout 2), the platform and its viewers are better-off if the platform uses a threshold of 0 to show ads. This results in a 51.0% increase in ad exposure as compared to observed ad exposure and a 0.93% increase in Bingeability as compared to initial predicted Bingeability (or a 1.14% decrease in Bingeability as compared to observed Bingeability). This indicates that by capturing correlation between outcomes for new viewers for whom viewer fixed effects are not known, a best threshold of 0 to show ads can be achieved. In addition, the presence of an increase of 0.93% in comparison with initial predicted Bingeability but a decrease of 1.14% in comparison with observed Bingeability suggests that there are unobserved covariates that influence observed Bingeability, whose effect cannot be completely captured by modelling correlation between outcomes.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

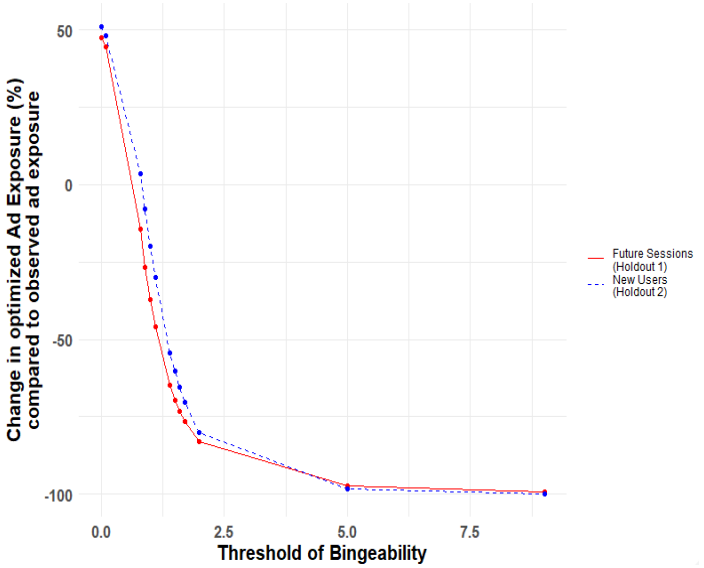


Figure E1:
Percentage change in optimized ad exposure ($\tilde{n}$) as compared to observed ad exposure (n)

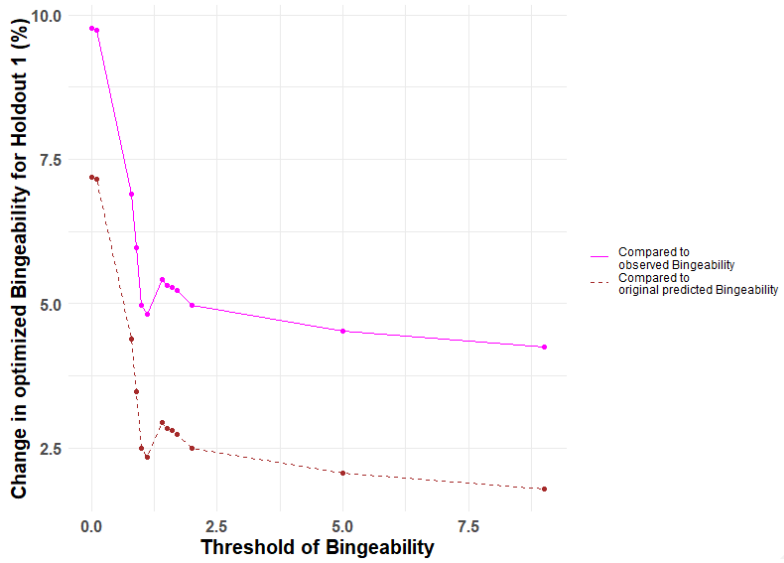


Figure E2:
Percentage change in optimized Bingeability ($\tilde{b}$) for Holdout 1

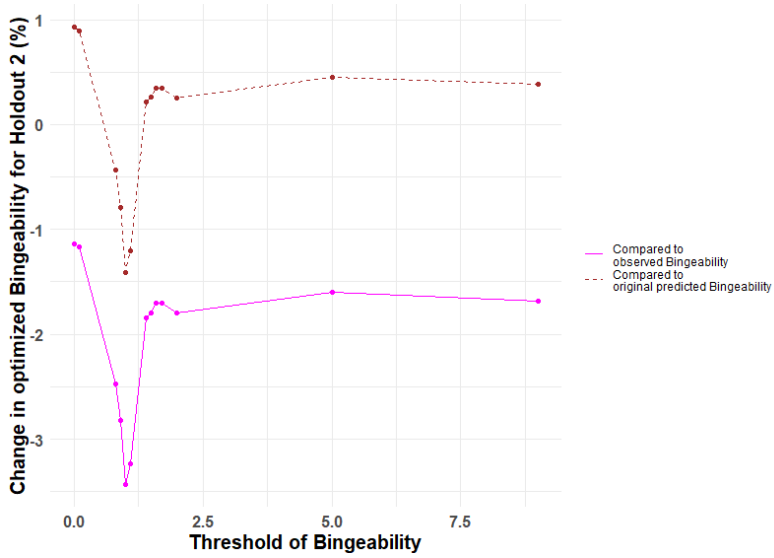


Figure E3:
Percentage change in optimized Bingeability ($\tilde{b}$) for Holdout 2

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX F: TREE BASED METHODS AND SIMULATED DATA

Boosting and Random Forests

Boosting or Boosted Regression Trees refer to a weighted linear combination of regression trees, with each tree trained greedily in sequence to improve the final output (Friedman, 2002). This output can be presented as follows:

$$F_N(x) = \sum_{k=1}^{N} \alpha_k f_k(x)$$

where, $f_k(x)$ is the function modelled by the k^{th} regression tree, and α_k is the weight associated with it. The value of f_k and α_k are learnt during model training. We adopt a recent extension of gradient boosting called Extreme Gradient Boosting (XGBoost) because it is a powerful method for making predictions with structured data (Chen & Guestrin, 2016). For the set of points (x_i, y_i), and a loss function $l(y_i, \hat{y_i})$, the XGBoost model minimizes an objective function $\mathcal{L}$ to find the step-wise value of $f_k(x)$. For our application, the loss function $l(y_i, \hat{y_i})$ is the least squares error when the outcome is Ad Tolerance which is continuous, and negative log likelihood when the outcome is Bingeability which is a count. The objective function $\mathcal{L}$ can be represented as follows:

$$\mathcal{L} = \Sigma_i l(y_i, \hat{y_i}) + \Sigma_k \Omega(f_k)$$

where Ω is the regularization parameter that penalizes the complexity of the model (see Chen and Guestrin (2016) for details on the objective function). If we set the regularization parameter to 0, we would get the traditional gradient boosting model. At each step, Newton's method computes the new value of f_k that minimizes the average value of the objective function. The step-wise iterations can be shown as follows:

$$F_k(x) = F_{k-1}(x) - \eta (H_k)^{-1} . g_k$$

where η is the learning rate, g_k with components $g_{ik} = \left[\frac{d\mathcal{L}(y_i, F(x_i))}{dF(x_i)}\right]_{F(x_i)=F_{k-1}(x_i)}$ is the gradient of the objective function and H_k is the second order gradient of the objective function. XGBoost has a faster computation time than conventional gradient boosting because it employs parallel processing using all the cores of the computer.

Random Forests refer to the average of thousands of distinct regression trees (Breiman, 2001). Unlike gradient boosting which uses weak learners or shallow trees at each step, Random Forests average

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

multiple deep trees. Each regression tree is different because it is constructed on a different training sample by sub-sampling on both observations and covariates. The output of Random Forests can be represented as follows:

$$F_{avg}(x) = \frac{1}{N}\sum_{k=1}^{N} f_k(x)$$

where, $f_k(x)$ is the function modelled by the k^{th} regression tree, which is learnt during model training.

Simulation

We simulate data to match the distribution of our real data. Using simulated data, we can create the ground truth, i.e. we know the extent to which the outcome variable is influenced by the observed covariates in the model. Hence, we can compare performance of different models in terms of their ability to make predictions that are closest to the ground truth. We adopt an approach similar to Hartford et al. (2017) to create our simulated data. We let the source of endogeneity be represented by $v \sim N(1,0.1)$ and the four instruments be represented by $z_1 \sim N(1, 0.1)$, $z_2 \sim N(1, 0.1)$, $z_3 \sim N(1, 0.1)$ and $z_4 \sim N(1, 0.1)$. The spacing rule (sr) is represented as follows:

$$sr \sim \max\left(W_1 + 2v + 13z_1 - 17\,, 0\right)$$
$$W_1 \sim \alpha_1 N(9.1\,, 1.5) + (1-\alpha_1)U(0,0)$$
$$\alpha_1 \sim Bern(0.85)$$

The correlation between sr and z_1 is 0.34 which is close to the correlation of 0.35 in the real data. The median value of the simulated and real distribution of sr is the same at 6.6 min. The length rule (lr) is represented as follows:

$$lr \sim W_2 - \frac{v}{4} + \frac{z_2}{4}$$
$$W_2 \sim \gamma_2 N(0.38, 0.01) + (1-\gamma_2)\big(\alpha_2 N(0.25, 0.002) + (1-\alpha_2)N(0.5, 0.002)\big)$$
$$\alpha_2 \sim Bern(0.4)$$
$$\gamma_2 \sim Bern(0.2)$$

The correlation between lr and z_2 is 0.22 which is close to the correlation of 0.25 in the real data. The median value of the simulated and real distribution of lr is the same at 0.42 min. The diversity rule (dr) is represented as follows:

$$dr \sim W_3 - v + z_3$$
$$W_3 \sim \alpha_3 U(0.15, 1) + (1-\alpha_3)U(1,1)$$
$$\alpha_3 \sim Bern(0.5)$$

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

$$dr = \begin{cases} 0.05, & dr \leq 0.05 \\ 1, & dr \geq 1 \end{cases}$$

The correlation between sr and z_3 is 0.25 which is close to the correlation of 0.27 in the real data. The median value of the simulated and real distribution of dr is 0.86 and 0.87 respectively. The clumpiness rule (cr) is represented as follows:

$$cr \sim W_4 - 1.1v + 1.3z_4$$
$$W_4 \sim \gamma_4\big(\alpha_4 N(0.05, 0.02) + (1-\alpha_4)U(0.3,0.99)\big) + (1-\gamma_4)U(1,1)$$
$$\alpha_4 \sim Bern(0.9)$$
$$\gamma_4 \sim Bern(0.8)$$
$$cr = \begin{cases} 0, & dr \leq 0 \\ 1, & dr \geq 1 \end{cases}$$

The correlation between cr and z_4 is 0.29 which is close to the correlation of 0.33 in the real data. The median value of the simulated and real distribution of cr is 0.33 and 0.30 respectively.

The outcome variable Bingeability (y_1) is simulated to have a complex non-linear relationship with the covariates. It is represented as follows:

$$y_1 \sim Poisson(\lambda)$$
$$\lambda = \max\Big(\frac{\alpha_{it}}{50} + 0.2sr_t - 0.8lr_t - 1.5dr_t + 0.5W_{1_t}^{W_{2_t}} - 2W_{2_t}^2 + exp\big(W_{3_t}\big) - W_{4_t} - 0.2 + u_t, 0\Big)$$
$$u \sim N(\rho v, 1-\rho^2)$$

u is the error term that is correlated with sr, lr, dr and cr; and ρ is the level of endogeneity which we set at 0.9. W_1,W_2, W_3 and W_4 are exogenous covariates which were defined earlier in the equation of each ad targeting rule. α_i corresponds to viewer fixed effects and we simulate 500 viewer fixed effects as follows:

$$\alpha_i = N\left(\frac{i}{10}, 0.01\right), i = \{1, \dots, 500\}$$

We ensure that the sign of the correlation between the outcome variable and the four endogenous variables in the simulated data is the same as that in the observed data.

The outcome variable Ad Tolerance (y_2) is also simulated to have a complex non-linear relationship with the covariates. It is represented as follows:

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

$$y_2 = \frac{\alpha_{it}}{10} - 0.05(sr - 0.9)^2 + 4000(lr - 0.4)^2 + 300e^{-4(dr+0.5)^2} - 25 + 10W_1^{W_2} - 20W_2^2 + 10\exp(W_3) - 50W_{4_t} + u$$

$$u \sim N(\rho v, 1 - \rho^2)$$

The level of endogeneity ρ is set at 0.9 as before. We again ensure that the sign of the correlation between the outcome variable and the four endogenous variables in the simulated data is the same as that in the observed data.

The first stage of the model can be represented as follows:

$$X_{1t} = sr_t = g_1\left(z_{1_t}, z_{2_t}, z_{3_t}, z_{4_t}, W_{1_t}, W_{2_t}, W_{3_t}, W_{4_t}, \alpha_{it}\right) + e_{1t}$$

$$X_{2t} = lr_t = g_2\left(z_{1_t}, z_{2_t}, z_{3_t}, z_{4_t}, W_{1_t}, W_{2_t}, W_{3_t}, W_{4_t}, \alpha_{it}\right) + e_{2t}$$

$$X_{3t} = dr_t = g_3\left(z_{1_t}, z_{2_t}, z_{3_t}, z_{4_t}, W_{1_t}, W_{2_t}, W_{3_t}, W_{4_t}, \alpha_{it}\right) + e_{3t}$$

$$X_{4t} = cr_t = g_4\left(z_{1_t}, z_{2_t}, z_{3_t}, z_{4_t}, W_{1_t}, W_{2_t}, W_{3_t}, W_{4_t}, \alpha_{it}\right) + e_{4t}$$

where, the subscript t denotes a session; e_{1t}, e_{2t}, e_{3t}, and e_{4t} are the error terms which are all equal to v_t in our simulation. The second stage of the model can be represented as follows:

$$y_{jt} = f_2\left(\hat{X}_{1t}, \hat{X}_{2t}, \hat{X}_{3t}, \hat{X}_{4t}, W_{1t}, W_{2t}, W_{3_t}, W_{4_t}, \alpha_{it}\right) + u_t$$

where y_j is either Bingeability (y_1) or Ad Tolerance(y_2), and $\hat{X}_{1t}, \hat{X}_{2t}, \hat{X}_{3t}, \hat{X}_{4t}$ are the fitted values from the first stage. Next, we need to determine the counterfactual function or ground truth against which the performance of different models can be compared. Let us represent this counterfactual function as h, and for each outcome variable the counterfactual function can be represented as follows:

$$h_1 \sim Poisson(\lambda_h)$$

$$\lambda_h = \max\left(\frac{\alpha_i}{50} + 0.2sr - 0.8lr - 1.5dr + 0.5W_1^{W_2} - 2W_2^2 + \exp(W_3) - W_4, 0\right)$$

$$h_2 = \frac{\alpha_i}{10} - 0.05(sr - 0.9)^2 + 4000(lr - 0.4)^2 + 300e^{-4(dr+0.5)^2} + 10W_1^{W_2} - 20W_2^2 + 10\exp(W_3) - 50W_4$$

Note, we removed the intercept terms and the endogenous error to get the equations of the ground truth h_1 and h_2. Next, we simulate different sizes of the data, as represented in Table E1, and split it into an 80% training sample and 20% holdout sample. We test the performance of three different models: XGBoost, Random Forests and Linear Regression (2SLS), in terms of their ability to get close to the ground truth, h. Note that the same model is applied on both the first and second stage. Model performance on the holdout

Electronic copy available at: https://ssrn.com/abstract=3510705
Electronic copy available at: https://ssrn.com/abstract=3498039

sample is compared in terms of the RMSE between $\hat{y}_1$ and h_1, and $\hat{y}_2$ and h_2 which is shown in Table F1. We find that XGBoost performs better than Random Forests and Linear Regression (2SLS) in getting close to the ground truth for both small and large data sizes. Hence, we use the XGBoost model to analyze our observed data.

	Bingeability			Ad Tolerance		
Data Size	Linear Regression	Random Forests	XGBoost	Linear Regression	Random Forests	XGBoost
5,000	2.12	1.93	1.90	50.48	34.75	25.97
25,000	1.90	1.88	1.84	48.81	31.74	24.72
50,000	1.89	1.88	1.82	48.04	29.55	24.82
75,000	1.88	1.88	1.82	48.43	28.97	24.61
100,000	1.87	1.87	1.81	48.64	29.50	24.66

Table F1: Comparison of Model Performance (RMSE) on holdout sample

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX G: CROSS-VALIDATION FOR XGBOOST

The parameters of the XGBoost model for the observed data are set by cross-validation. We carry out 5-fold cross validation on the training sample by dividing viewers into five different folds. This process is repeated 10 different times with random splits made on the training data to determine the 5 folds. The parameters that are tuned are as follows:

- Maximum depth of a tree: {4,6}
- Minimum threshold for loss reduction, γ: {0,5}
- Regularization parameter on weights of a leaf: {0,1}
- Row subsampling fraction: {0.8,1}
- Column subsampling fraction at the node level: {0.8,1}

The minimum number of observations on each node of a leaf is set to 1, and the value of the learning rate η is set by judgement to ensure that the cross-validation process does not take unduly long to finish. We have $2^5 = 32$ distinct parameter combinations and 10 iterations for each parameter combination. As an exhaustive grid search for a total of 320 iterations over 74,996 observations (in the training sample) will take unduly long to finish, we use an efficient three step process to decide the final parameter combination to be used to tune the training sample.

- Step 1: We use a "fractional factorial design" to design $2^{5-2} = 8$ combinations of the parameters that are balanced and orthogonal. Then we run the cross-validation routine 10 times for these 8 combinations for a total of 80 iterations. Then we average the performance measure (e.g. RMSE or Negative Log Likelihood) across the 10 iterations for each of the 8 combinations and rank the combinations in order of best to worst performance.
- Step 2: Next, we analyze the performance across the 8 orthogonal combinations and identify other potential parameter combinations that could result in an improved cross-validation performance. We run the cross-validation routine 10 times for each of these newly identified parameter combinations. Then we average the performance measure across the 10 iterations for each of the newly identified combinations.
- Step 3: The parameter combination that leads to the lowest average value of the performance measure across the 10 repetitions for the parameter combinations in Step 1 and Step 2 is chosen to train the model.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX H: OPTIMIZATION PROCEDURE

Part 1

Our objective function is subject to the constraint of not detracting from the content consumption experience. This constraint corresponds to the equation of the Ad Tolerance metric, originally shown in equation (4), which is reproduced below.

$$Ad\ Tolerance = \sum_{j=1}^{n_p} (w_1 PodDuration_j + w_2 ContentEnd_j - w_3 (CalendarPod_j - PodDuration_{j-1}))$$

This constraint ensures that the optimization routine (of ad maximization) takes cognizance of the predicted values of Ad Tolerance and Bingeability, thus preventing the routine from making scheduling recommendations that can cause a reduction in the amount of content viewed. Now, we substitute the variables from equation (9) in the above equation, i.e. $n_p = n$; $PodDuration_j = d$; $ContentEnd_j = \hat{b}e - js$ where j is pod number; and $PodDuration_{j-1} = d$. Hence, we can rewrite the constraint corresponding to the Ad Tolerance metric as follows:

$$Ad\ Tolerance = \sum_{j=1}^{n} \left(w_1 d + w_2 (\hat{b}e - js) - w_3 (CalendarPod_j - d) \right)$$

To substitute values into $CalendarPod_j$, we use equation (1) which is reproduced below:

$$\overbrace{Session\ Time}^{\text{Measured}} = \overbrace{Content\ Time + Ad\ Time + Filler\ Content\ Time}^{\text{Measured}} + \overbrace{Pauses\ - Fast\ Forward + Rewind}^{\text{Unmeasured}}$$

Using the variables in equation (9), the above equation can be rewritten as follows:

$$\overbrace{CalendarPod_j}^{\text{Measured}} = \overbrace{s + d + f_j}^{\text{Measured}} + \overbrace{\tilde{u}_j}^{\text{Unmeasured}}$$

where, f_j is duration of filler content viewed from the beginning of the pod j-1 till the beginning of pod j and $u_j = (pauses - fast\ forward + rewind)_j$ is the sum of the unmeasured variables from the beginning of pod j-1 till the beginning of pod j. A viewer is not expected to be immersed in the viewing experience while watching filler content; hence we allow the viewer to skip it or fast forward it by setting f_j to 0. As the unmeasured variables—Pauses, Fast Forward and Rewind—are directly under viewer control and cannot be controlled by the streaming provider, we set u_j to 0. Hence, we can rewrite the constraint (equation (10)) as follows:

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

$$\hat{a} = \sum_{j=1}^{n} \left(w_1 d + w_2 (\hat{b}e - js) - w_3 (s + d - d) \right)$$

where the predicted value of Ad Tolerance is shown as $\hat{a}$. After summing over the variables, we get

$$\hat{a} = w_1 nd + w_2 \left(n\hat{b}e - \frac{n(n+1)}{2} s \right) - w_3 ns$$

Part 2

The partial derivative of $\tilde{n}$ with respect to $\hat{a}$ is shown below:

$$\frac{\partial \tilde{n}}{\partial \hat{a}} = \frac{1 + 2\hat{a} + 3\hat{b}e + \sqrt{\Delta}}{(1 + \hat{b}e)\sqrt{\Delta}} > 0$$

The above equation is always > 0 because $\hat{b} \geq 1, \hat{a} > 0, \sqrt{\Delta} > 0$ and $e > 0$. Thus, controlling for $\hat{b}$ and e, an increase in Ad Tolerance results in an increase (decrease) in the number of ads $\tilde{n}$ (spacing $\tilde{s}$) $\left(\because \ \tilde{s} \propto \frac{1}{\tilde{n}} \right)$. The partial derivative of $\tilde{n}$ with respect to $\hat{b}$ is shown below:

$$\frac{\partial \tilde{n}}{\partial \hat{b}} = \frac{e(\hat{b}e - 3\hat{a}\hat{b}e + \sqrt{\Delta}(1 - \hat{a}) + \hat{a} - 2\hat{a}^2 - 1)}{(1 + \hat{b}e)^2 \sqrt{\Delta}}$$

On substituting the values of $\hat{b}, \hat{a}$ and e from the observations in Set C (from Table 9) into the above equation, we find that the partial derivative is almost always negative. For the few instances when $\hat{a} \in (0, 0.5 \text{ min})$, the value of the partial derivative is positive. Thus, controlling for $\hat{a}$ and e, an increase in Bingeability almost always results in a decrease (increase) in the number of ads $\tilde{n}$ (spacing $\tilde{s}$). The interpretation of the partial derivative for the few instances when $\hat{a} \in (0, 0.5 \text{ min})$ can be understood as the effect of the algorithm to ensure a minimum level of Ad Tolerance before recommending a decrease (increase) in the number of ads $\tilde{n}$ (spacing $\tilde{s}$) for an increase in Bingeability, $\hat{b}$.

Overall, the partial derivatives help illustrate the direction of the influence of the metrics on the recommended spacing $\tilde{s}$.

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

WEB APPENDIX I: RECOMMENDED SCHEDULE VERSUS A NAÏVE HEURISTIC

We develop a naïve heuristic based on viewer response to ad delivery that could be used to recommend ad spacing. As mentioned in the "Introduction" section, viewer response to ad delivery has been studied in past work by Schweidel and Moe (2016) who find that ad exposure is negatively correlated with content consumption. Hence, one naïve heuristic for a session (i) is the ratio of total time spent watching TV shows in the past week by the viewer (before the commencement of the session) to the total number of pods shown to that viewer while watching TV shows in the past week. It can be represented as follows:

$$Naive\ Spacing_i = \frac{Total\ Content\ Time_i}{Total\ Number\ of\ Pods_i}$$

The heuristic is naive for mainly the following reasons (1) it does not incorporate the frequency (or spacing) of pod exposure in the viewing experience, as done by the Ad Tolerance metric (2) it does not account for fast-forwarding or skipping behavior, as done by the Bingeability metric, and (3) it does not control for the non-randomness in ad delivery, as done in our model using instrumental variables.

A density distribution of the naive spacing for those sessions in Set C (from Table 9) is shown in Figures I1a and I1b for Holdout 1 and Holdout 2 respectively. We ignore those sessions which are the first sessions of the viewers, because the value of the naïve spacing metric will result in a 'divide by 0' error. The median value of the naive spacing for Holdout 1 (future sessions of the viewers in the training sample) is 8.30 minutes and its 2.5^{th} to 97.5^{th} percentile range is from 4.19 to 17.96 minutes. The median value of the naive spacing for Holdout 2 (new viewers) is 9.16 minutes and its 2.5^{th} to 97.5^{th} percentile range is from 5.01 to 18.11 minutes.

The distribution of the ratio of (optimized) recommended spacing and naïve spacing is shown in Figures I2a and I2b. We ignore those sessions where the naïve spacing is 0 minutes to avoid a 'divide by 0' error and also because they suggest showings ads continuously without any show content in between which is not meaningful. The median value of the ratio is 0.50 in Figure I2a and 0.47 in Figure I2b. As the median is a lot less than 1 (which is the 90^{th} percentile in Figure I2a and 95^{th} percentile in Figure I2b), we can conclude that the naïve heuristic suggests a lower frequency (longer spacing) a lot more often than the (optimized) recommendation, thus losing out on opportunities to maximize ad exposure.

Finding the ratio of the naïve spacing with the average observed spacing reveals that the median of the distribution is 1.25 for each holdout sample. This demonstrates that on average the naïve schedule recommends showing ads less frequently (with a longer spacing) as compared to observed practice. We quantify the decrease in ad exposure by using the naïve heuristic as done in the subsection "Decision Support System" for a Bingeability threshold of 0. Ad exposure increases by -35% for Holdout 1 and by

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039

−40% for Holdout 2 as compared to observed practice. On the other hand, content consumption increases by 11.56% as compared to initial predicted Bingeability (and by 12.47% compared to observed Bingeability) for Holdout 1 and by 4.62% compared to initial predicted Bingeability (and 1.44% compared to observed Bingeability) for Holdout 2. As the naïve heuristic is unable to make both the platform and the viewers better off, it is inferior to the optimized ad schedule.

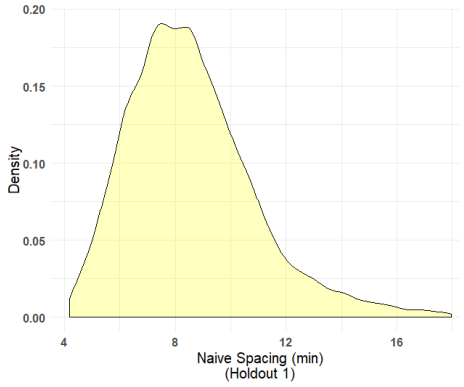


Figure I1a: Density of Naïve Spacing Holdout 1 (2.5th to 97.5th percentile)

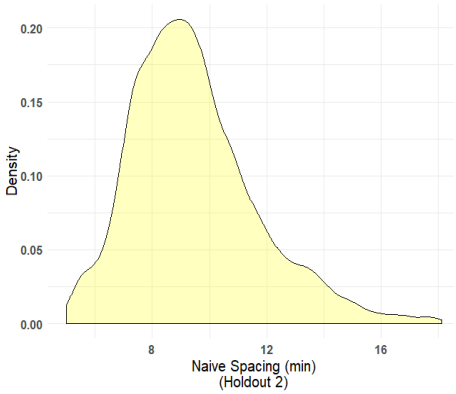


Figure I1b: Density of Naïve Spacing Holdout 2 (2.5th to 97.5th percentile)

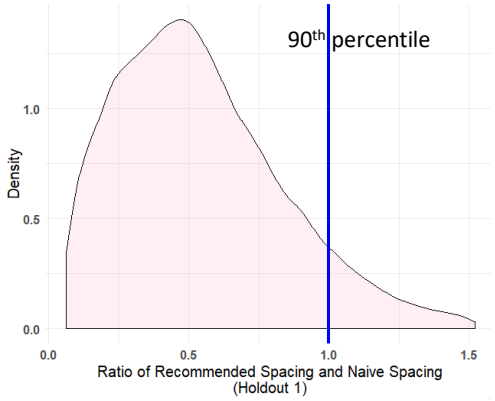


Figure I2a: Density of the Ratio of Recommended Spacing and Naïve Spacing Holdout 1 (2.5th to 97.5th percentile)

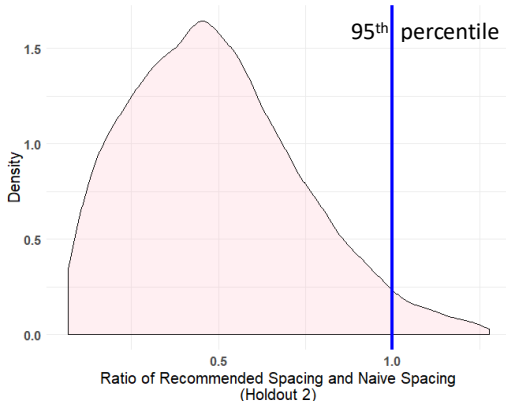


Figure I2b: Density of the Ratio of Recommended Spacing and Naïve Spacing Holdout 2 (2.5th to 97.5th percentile)

Electronic copy available at: https://ssrn.com/abstract=3510795
Electronic copy available at: https://ssrn.com/abstract=3498039