New York University - Stern School of Business
Working paper CBA-16-02

Who's Watching TV?

Jessica Clark

IOMS Department

NYU Stern School of Business

Jean François Paiement

AT&T Research

Foster Provost

IOMS Department

NYU Stern School of Business

October 24, 2016

Abstract

TV viewership data available at the individual set-top box level has enabled new methods for estimating the demographics of shows' audiences, but it is impossible to tell with certainty which household members are watching TV in multi-person households. We address this problem through four main contributions. First, we develop a novel method for estimating the likelihood that each individual in a multi-person household is watching. Second, we derive a set of tasks at which models must succeed in order to demonstrate that they have solved the core problem, since there are no ground-truth labels. Third, we evaluate our new method as well as two current state-of-the-art heuristic methods. Fourth, we conduct some example analyses of viewership in the context of living with others. Our solution has implications for advertisers, researchers who seek better understanding TV viewership, and anyone using data generated by shared devices or accounts. A major TV provider is planning on deploying this method for use in their TV ad-targeting system. No personally identifiable information (PII) was gathered or used in conducting this study. To the extent any data was analyzed, it was anonymous and/or aggregated data, consistent with the carrier's privacy policy.

1 Introduction

Television audience research has traditionally been conducted using data from Nielsen or other competitors which use aggregate opt-in panel data to report various demographics' viewership. There are a few disadvantages to using such data: they require users to self-report viewership instead of directly observing behavior; their panels are relatively small, leading to unreliable estimates for either TV shows that are not commonly watched or demographic groups that are sparsely populated or very specific; and they have historically been slow to measure media consumption via novel channels such as on mobile devices (Steel 2014). Additionally, Nielsen in particular has recently been found to have ratings errors (Carter and Steel 2014). Finally, panel-based data is very expensive and TV providers now have an alternative data source that serves the same purpose.

More recently, ad networks and television providers are able to collect digital channel-change data for each individual set-top box (STB). An STB is the device through which customers access digital cable programming on their televisions. These data have the advantage of being available at the individual STB level and therefore overcome some of the weaknesses inherent in panel-based data (Nielsen and similar data providers collect data at the individual level but report aggregate numbers to their users). Thus, state-of-the-art algorithms for understanding TV audiences utilize STBs (Balakrishnan et al. 2012, Kitts and Au 2014, Spangler et al. 2003); however, it is not currently possible for cable[1] providers to understand which individual associated with an STB is the one watching.

Standard practice in research and industry is to use all STBs in households that include at least one member estimated to be of the target demographic for audience profiling, forecasting, and media plan optimization (Balakrishnan et al. 2012, Kitts and Au 2014). That is, the assumption is that everyone in the household is watching TV, all of the time that it is on. As we demonstrate later on in this paper, this practice results in an inaccurate picture of what various demographics watch. Not knowing who is watching in the multi-person households is a core problem that diminishes stakeholders' trust in using individual STB data rather than panel-based data such as that provided by Nielsen (Stack 2014), despite the disadvantages of using Nielsen or similar competitors' data described above. This is because Nielsen does collect information on the viewing behavior of

[1] We use the term "cable" loosely here; our work pertains to television viewing sources with the characteristic that it is possible to measure what is being watched at what time.

individuals within households.

To the best of our knowledge, this is the first paper to develop a solution that increases the quality of the STB data by estimating the probability that each member of the household is watching. We present a novel framework for modeling that first uses single-person households, then adjusts the result to the specifics of individual multi-person households. We also formalize two state-of-the-art heuristics to use as baselines.

The heuristic estimators we use as the basis for comparison may seem simple, especially when compared to the highly sophisticated and complex models developed and used in online advertising. The sophistication of ad targeting on TV has lagged behind that of online advertising, where the vast amount of highly granular data enables nearly instantaneous and highly specific individual action and measurement. In contrast, television advertising schedules are often decided months in advance (most of the time even before the show schedule is set), and there has, until now, been little transparency into the behaviors of individuals within households. This is particularly troubling because televisions (unlike laptops or phones) tend to be shared by multiple family members. Additionally, people frequently watch TV in groups (Morrison and Krugman 2001).

Because it is not currently possible to separate the viewership of individuals within STBs (and therefore establish ground truth), evaluating the performance of any model for this problem is non-trivial. Therefore, we derive a set of three tasks related to the core problem for which we *do* have ground truth. Any model that can be said to succeed at estimating which person is watching TV must necessarily have high performance on all three tasks; this evaluation framework enables us to rule out unsatisfactory solutions and compare among models. The novel model we develop is the only one we test that has acceptable performance on all three tasks, and in fact it dominates the two state-of-the-art heuristic estimators at all tasks.

A final goal of this work is to contribute to the understanding of television viewership in a social context. Watching TV is understood to be a social activity (Morrison and Krugman 2001), and recommender systems for groups of people (as opposed to individuals) is an active research topic (Chaney et al. 2014). Thus, we use our model to investigate three questions: (1) whose tastes are the most likely to dominate viewership in households, (2) how individuals living in multi-person households differ from otherwise-similar individuals who live alone, and (3) which channels are most likely to be watched by all household members versus only one household member.

The rest of the paper is organized as follows. The next section formally describes a mathematical formulation of the predictive problem along with the notation and terminology that we use going forward. It also describes the models we develop for prediction: the two current state-of-the-art heuristic estimators, and the new model we develop. Section 3 describes some properties that successful models should possess, and shows a set of metrics that capture those properties. We use that set of metrics for evaluating the models we have developed in this work. Next, Section 4 describes our experimental procedure and some promising results. Section 5 demonstrates a few example analyses that could be facilitated using the resulting model. Section 6 ties our new model to existing methods used for predictive modeling of television viewership, as well as the machine learning literature that it relates most closely to. We conclude in Section 7 with a discussion of further applications of this particular model, the broader implications of this type of modeling beyond the TV advertising world, and some intriguing future directions.

2 Modeling Framework

2.1 Problem Statement and Notation

We assume two major data components are available for each STB i: first, the number of seconds that it was tuned to channel c in some time period. The second data component is demographic information for each individual within a household that contains STB i. This information will help us to distinguish among the members of the household. Our specific data were provided by a major IPTV entertainment provider and contain anonymized viewership information for individual STBs as well as demographic features for the customers in those STBs.[2]

The higher-level goal is to predict the likelihood that each person associated with an STB watches a TV "segment" (a show, a channel, etc.). We formalize the question as: given a channel c, an STB i, and a person i_j associated with STB i who has demographic feature vector $\mathbf{x}_{i_j}$, what is the probability that i_j watches c for more than 30 minutes per week[3] on average during the time period of interest?

Let $b_{c,i}$ denote the event that STB i watches channel c for more than 30 minutes per week on

[2]No personally identifiable information (PII) was gathered or used in conducting this study. To the extent any data was analyzed, it was anonymous and aggregated data.

[3]We have observed that the results described in this paper are qualitatively similar if this threshold is varied.

average. Similarly, y_{c,i_j} is the binary label for individual j within STB i representing the event that individual i_j watches channel c for more than 30 minutes per week on average. The $b_{c,i}$ labels are observed in the training data; the y_{c,i_j} are what we are trying to predict, but are not observed for multi-person households. Each method that we develop will result in two estimates:

1. Estimates of $\mathbb{P}(y_{c,i_j})$ for each individual j associated with multi-person STB i. This is the quantity that we would like to estimate in this paper.

2. An estimate for $\mathbb{P}(b_{c,i})$, the label of household i. Although we directly observe this quantity on historical data, the estimate is useful in evaluation.

An important assumption is that if an STB is tuned to a channel, then at least one person associated with the STB must be watching it, and if the TV is off then nobody is watching. The event $b_{c,i}$ is equivalent to the union of the events $y_{c,i_1}, y_{c,i_2}, \ldots$. This means that the probability that STB i watched channel c is the probability that individual 1 *or* individual 2 *or* individual 3, etc. watched. Specifically, we will utilize the approximation given in Equation 1, which simply states that the probability that STB i watches channel c is equal to the probability that any of the individuals within i watch channel c:

$$\mathbb{P}(b_{c,i}) = \mathbb{P}(y_{c,i_1} \cup y_{c,i_2} \cup \ldots) \tag{1}$$

A key challenge of this problem is that there are no ground truth labels for who is actually watching. Therefore, we have derived a set of three tasks that (A) are related to the core problem of interest—models that fail at these tasks are unacceptable, and doing better at any of these tasks implies doing better at the core problem, and (B) are possible to evaluate using elements of the training data. As part of our business understanding, via discussions with stakeholders, we have determined that successful models should (i) have high predictive performance on a held-out set of weeks, (ii) should match some benchmark demographic distributions, and (iii) should produce reasonable estimates for the total possible audience of viewers. Section 3 goes into more detail on these tasks, how we evaluate whether models are successful at them, and how the tasks can be at odds with each other.

NB: there is an important distinction between single-person and multi-person households.

Multi-person households are the target of our analysis: the purpose of this research is to distinguish among viewers in these households. Therefore, all evaluation is done on multi-person households only. Because we know which person is watching in the single-person households, they are highly useful, and so we use both multi-person and single-person STBs for training. Put another way, the multi-person STBs constitute the focal training and test data; the single-person STBs function as auxiliary training data, drawn from a related but different distribution.

2.2 Heuristic Estimators

This section formalizes two current industry state-of-the-art methods for estimating the demographic composition of television audiences (Spangler et al. 2003). We use these as baselines against which to compare the model we have developed for this paper.

2.2.1 Whole Household Estimator

The first baseline strategy is to assign the STB's viewership equally to all individuals within the household. That is, for each person i_j associated with each STB i:

$$y_{c,i_j} = \begin{cases} 1 & \text{if } b_{c,i} = 1 \\ 0 & \text{if } b_{c,i} = 0 \end{cases}$$

We use a very simple predictive model for generating estimates: we predict that STB i will watch channel c in the test period if and only if it was tuned to c in the training period. This heuristic is denoted as the Whole-Household Estimator (WHE). WHE is straightforward to compute. For each channel c and (multi-person) STB i:

1. Estimate $\mathbb{P}(b_{c,i}) = 1$ if $b_{c,i} = 1$ in the training period, otherwise $\mathbb{P}(b_{c,i}) = 0$.
2. Predict $\mathbb{P}(y_{c,i_j})$ to be $\mathbb{P}(b_{c,i})$ for each individual j who is associated with i.

WHE corresponds to the current state-of-the-art method for assigning viewership among people associated with an STB (Balakrishnan et al. 2012, Kitts and Au 2014), and the reaction of stakeholders to WHE's estimates was the original motivation for this research. In the setting of generating media plans (lists of recommended channels on which advertisers should place ads to

maximize exposure to target demographic audiences), utilizing this heuristic results in non-intuitive recommendations (and therefore it does not do well at task (ii) mentioned in Section 2.1 above). This issue stems from multi-person households' viewership.

For instance, households containing men in their 30's and 40's frequently also contain women because such people are likely to be married. Men and women tend to have somewhat different viewership preferences; however, because the viewership is assigned equally within households, men can be credited with watching the channels that are likely instead being watched by their wives. Thus, although very accurate along some dimensions, WHE's recommendations are frequently unacceptable to external stakeholders. Consider, as an example, an advertiser who wants to to reach 30-45 year-old men. They would object to a recommendation to advertise on the Lifetime network (which at one point had the slogan "television for women"). Moreover, they would reject an analytical *model* that made such recommendations.

2.2.2 Single-Person Household Estimator

The second baseline heuristic is to estimate viewership based on the viewing behavior of only the single-person STBs, and is designated going forward as the Single-Person Household Estimator (SPHE). This method is our formalization of a standard heuristic alternative that "corrects" the non-intuitive predictions that come from using WHE. End users frequently discard the multi-person STBs in order to determine the channels which their desired demographic audience watches the most. We formalize this idea by learning predictive models for $\mathbb{P}(y_{c,i_j}|\mathbf{x}_{i_j})$ for STBs that contain only one individual, using the individual's characteristics as features, and denote the function learned for channel c as $\phi_c(\mathbf{x}_{i_j})$. We estimate:

$$\mathbb{P}(y_{c,i_j}|\mathbf{x}_{i_j}) = \phi_c(\mathbf{x}_{i_j}) \tag{2}$$

Next, we predict the multi-person STBs' individuals' probabilities of watching using the resulting models. The prediction of STB i's probability of viewing is then computed as the probability of any one individual watching. Assuming that the individuals watch independently, we then estimate $\mathbb{P}(b_{c,i})$ as the product across the N_i individuals associated with STB i:

$$\mathbb{P}(b_{c,i}) = 1 - \prod_{j=1}^{N_i}(1 - \phi_c(\mathbf{x}_{i_j})) \tag{3}$$

Equation 3 represents the probability that within STB i, person 1 watches OR person 2 watches, etc.

Computationally, the single-person modeling is a relatively standard prediction task, so we have many options for how to approach it. Engineering good features using the demographic traits available is one part. Also, standard modeling decisions apply, such as whether to use nonlinear or linear models, and how much and what type of regularization to use. These parameters can be tuned using cross-validation across single-person households. The quality of the single-person models is key to success using this method.

We thus build $L2$-regularized logistic regression models to predict $\mathbb{P}(b_{c,i_j}|\mathbf{x}_{i_j})$ via learning a function ϕ_c for each channel c using all of the single-person STBs. The choice of model type here and standard parameters are optimized by comparing the AUC using 3-fold cross validation over the training data. Note that both WHE and SPHE treat all channels' viewerships independently.

While SPHE may produce estimates that are apparently more intuitive based on the demographics, there are problems with it. First, SPHE ignores demographic groups which are inherently associated with multiple people in a household, such as "housewives." Equally as important, it ignores differences in behavior that may arise when individuals live together. This could be due to inherent differences in the demographic or taste characteristics of people who live with others rather than alone (for instance, men and women who are married may be more conservative than other people of their age who live alone). In addition, living together alters viewing patterns such that people who watch TV together make different choices. For example, it has been empirically demonstrated that husbands are more likely to influence wives' television viewership choices than vice versa (Yang et al. 2006).

2.3 The Mixed Estimator

The development of the following method is our first main contribution, which draws upon the strengths of the two baseline methods above. That is, this method utilizes the single-person households' probabilities of watching the various channels, as does SPHE, but also leverages within-

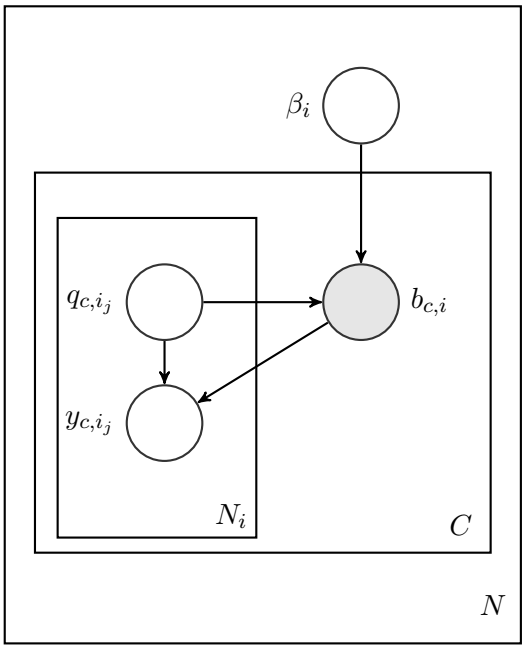

Figure 1: A graphical representation of the Mixed Estimator model.

household information, as does WHE. We will denote this model going forward as the Mixed Estimator (ME). A graphical representation of ME is shown in Figure 1.

β_i is a vector of parameters specific to STB i. Let $\mathbf{q}_{c,i}$ be a vector of scores that are believed to correlate with the y_{c,i_j}, which we will explain in greater detail below. We start by proposing a probabilistic process by which the STB labels $b_{c,i}$ and individual labels y_{c,i_j} are generated:

For each STB i:

1. Draw $\beta_i \sim \mathcal{N}(0, \frac{1}{2k}\mathbf{I}_k)$.

2. For each of the C channels c:

   (a) Draw $b_{c,i} \sim \text{Bernoulli}(p_{c_i})$, where

   $$p_{c,i} = \frac{1}{1 + \exp -\beta_i \cdot \mathbf{q}_{c,i}} \tag{4}$$

   (b) Draw y_{c,i_j} from $\mathbb{P}(y_{c,i_j}|b_{c,i}, q_{c,i_j})$, a Bernoulli distribution conditioned on the STB label (defined below in Equations 11-13).

In this generative process, only the $b_{c,i}$ are observed in the training data. The q_{c,i_j} can be thought of as a sort of "prior" for the individual binary labels y_{c,i_j} and can be populated using any data source that the user believes to be correlated with the (unobserved) label y_{c,i_j} for a particular person. One possibility includes ratings from Nielsen or another aggregate (external) data source; here, in order to not require external data (which can be expensive or difficult to obtain) we utilize the probabilities learned from the single-person STBs. That is, use ϕ_c as defined in Equation 2:

$$q_{c,i_j} = \phi_c(\mathbf{x}_{i_j}) \tag{5}$$

The STB-level parameters β_i are learned separately for each STB i using logistic regression, such that the Logistic likelihood function is maximized:

$$\beta_i = \underset{\beta}{\operatorname{argmin}} \prod_c (p_{c,i})^{b_{c_i}} (1 - p_{c,i})^{1-b_{c,i}} \tag{6}$$

$$p_{c,i} = \frac{1}{1 + \exp(-\beta \cdot \mathbf{q}_{c,i})} \tag{7}$$

Next, the model estimates $\mathbb{P}(y_{c,i_j}|b_i)$. These values correspond to the individual-level probabilities, and are not labeled in the training data. We use the assumption given in Equation 1 to constrain our estimates: if $b_{c,i} = 1$, then there must be at least one j such that $y_{c,i_j} = 1$; otherwise, all $y_{c,i_j} = 0$. We have additionally assumed that the y_{c,i_j}'s are conditionally independent given $b_{c,i}$, so we have:

$$\prod_j \mathbb{P}(y_{c,i_j} = 0 | b_{c,i} = 1) = 0 \tag{8}$$

$$\Rightarrow \mathbb{P}(y_{c,i_k} = 1 | b_{c,i} = 1) = 1, \text{ for some } k \tag{9}$$

Thus, there must be at least one individual i_k in STB i that is watching the TV whenever it is tuned to channel c. To decide which individual that is, the model uses both the scores q_{c,i_j} and the parameters β_i learned in Equation 6. In logistic regression, $\exp(\beta_{i_j})$ represents the increase

in the odds that $b_{c,i} = 1$ when the predictor q_{c,i_j} increases by one unit (all other predictors being equal). Therefore, a rough interpretation of these coefficients is that they represent the degree to which each individual i_j's viewership patterns match those of STB i as a whole. Equation 6, then, mixes the q_{c,i_j} by reweighting them proportionally to the weights β_{i_j}. The degree to which $b_{c,i}$ is influenced by each q_{c,i_j} is proportional to both $\exp \beta_{i_j}$ and q_{c,i_j}. We define the *relevance* r_{c,i_j} of each individual i_j for channel c as

$$r_{c,i_j} = \exp(\beta_{i_j}) p_{c,i_j} \tag{10}$$

The relevances determine the individual conditional probabilities $\mathbb{P}(y_{c,i_j} = 1 | b_{c,i})$. Let k be chosen such that $r_{c,i_k} \geq r_{c,i_j}$ for all $j \neq k$. We constrain:

$$\mathbb{P}(y_{c,i_k} = 1 | b_i = 1) = 1 \tag{11}$$

$$\mathbb{P}(y_{c,i_j} = 1 | b_i = 1) = \frac{r_{c,i_j}}{r_{c,i_k}}, \text{ for } j \neq k \tag{12}$$

$$\mathbb{P}(y_{c,i_j} = 1 | b_i = 0) = 0, \text{ for all } j \tag{13}$$

Defining the individual conditional probabilities in this way is satisfying for two reasons. First, it guarantees that when STB i is tuned to channel c, there is at least one individual i_k who is watching, and similarly that when STB i is *not* tuned to channel c, no individuals are watching. Second, it preserves the relative values of the r_{c,i_j} among the individuals for each channel.

The final step is to estimate each $\mathbb{P}(y_{c,i_j})$ (the marginal individual probabilities). Our assumptions entail that $\mathbb{P}(b_{c,i} = 1 | y_{c,i_j} = 1) = 1$ (the TV is guaranteed to be tuned to channel c if any individual is watching channel c). Bayes' rule yields:

$$\mathbb{P}(y_{c,i_j} = 1) = \mathbb{P}(y_{c,i_j} = 1 | b_{c,i} = 1) \mathbb{P}(b_{c,i} = 1) \tag{14}$$

We utilize the estimates for b_{c_i} as defined above in WHE: $\mathbb{P}(b_{c_i} = 1) = 1$ if b_i was tuned to channel c for at least 30 minutes per week in the training period, otherwise $\mathbb{P}(b_{c_i} = 1) = 0$.

Note the assumption implicitly embedded in our model. Rather than assume that each individual watches independently of all others as in SPHE, we assume instead that each individual watches

independently of others *given* the STB label. In section 2.2.2, we showed concrete interpretations of situations that would cause individuals to watch differently when they live with others, rather than alone. Also note that this model allows for more nuanced estimates. Namely, it is simple to compute joint marginal probabilities: the probability that multiple individuals are watching at the same time. This inference will be helpful later on, when we conduct some preliminary analyses of social viewership facilitated by this model.

2.4 An Example

To see the difference in predictions that each of these methods makes, consider an example STB in a hypothetical household containing two members: an adult man and an adult woman. We know the $b_{c,i}$ values for this STB; among the other channels, assume that the STB did watch the Oprah Winfrey Network (OWN), which is highly associated with female viewers in the single-person STBs; the National Football League channel (NFL), similarly highly associated with male viewers in the single-person STBs; but did not watch Animal Planet, which is roughly equally likely to be watched by females and males.[4] Tables 1 and 2 show the predictions that each model might make for the members of this household.

Because WHE makes equal predictions for all household members, it makes the (probably incorrect) estimate that the man watches OWN and the woman watches NFL. The SPHE relies on scores learned in the single-person household but does not adjust them to match what the specific focal multi-person household has actually watched; therefore, SPHE estimates that both household members will watch Animal Planet with non-zero probability, even though the household has historically never watched Animal Planet. It is therefore clear that ME is the only model that makes acceptable estimates: they account for the specific household's viewership patterns yet still differentiates among the household members.

3 Evaluation

Evaluation is not straightforward for this problem, due to the lack of ground-truth labels for which person is watching in the multi-person households. We have derived several properties that models

[4]See Appendix A for descriptions of the channels in our data set and the demographic groups that they are typically associated with.

Table 1: Viewership estimates for a woman residing in the hypothetical household.

		y_{c,i_j}		
Channel	Actual $b_{c,i}$	WHE	SPHE	ME
OWN	1	1	.2	1
NFL	1	1	.05	.2
Animal Planet	0	0	.1	0

Table 2: Viewership estimates for a man residing in the hypothetical household.

		y_{c,i_j}		
Channel	Actual $b_{c,i}$	WHE	SPHE	ME
OWN	1	1	.01	.025
NFL	1	1	.4	1
Animal Planet	0	0	.15	0

must necessarily possess if they are to be considered successful. These properties are all related to the core problem of understanding which person is watching TV; however, unlike the core problem, we have ground-truth labels that can be used to evaluate performance on these tasks.

This section presents these properties, as well as describing how we measure success (or lack thereof) on each one. These can be viewed as separate standards by which we will evaluate our success. Each of the heuristic estimators performs well at one task; however, they both have poor performance on at least one other task and can thus be ruled out as acceptable solutions. A successful model *must* perform well on all three components.

3.1 Household-level Predictive Ability

The first goal that we would like to accomplish is to maximize the generalization accuracy of our household-level predictions. We have defined b_i as a binary variable—either the household watches the channel or not. In our experimental set-up, each STB's viewership is divided into a set of training weeks and test weeks. The methods we describe in this paper produce a score for each household which represents its estimated probability of watching. We therefore measure the AUC (area under the ROC curve) for these scores, versus the actual value in the test period. AUC is equivalent to the Mann-Whitney-Wilcoxon statistic, which measures the probability that a classifier ranks any randomly chosen positive instance higher than any randomly chosen negative instance (Provost and Fawcett 2001). We compute the average AUC across all channels for each model that we evaluate.

Note that while it is a proxy for the actual goal of the modeling, evaluating on STB-level predictions does provide important information: one would expect better individual-person-level estimates to yield better STB-level estimates. As we describe in Section 6 below, a typical setup in similar problems is to train models on bags (collections of instances) in a set of training data, while holding out some of the bags as a test set (Dong 2006). Instead of comparing predicted vs. actual STB labels on a held-out set of STBs, our task is to compare on the same STBs, but on held-out weeks of data.

3.2 Single-person Similarity

The second task that successful models must accomplish is to generate predictions that are relatively similar to a benchmark demographic distribution. This benchmark is based on the relative amounts that different groups of adults in single-person STBs watch each channel. The set of single-person STBs is particularly useful here, because they provide a clean label—in these cases, we know which household member is the one watching the television. It is important to note that just as with the first evaluation task, this is a proxy for what we truly want to estimate. Models that do poorly at this task can be ruled out. Ceteris paribus, improving similarity to the single-person household benchmark gives evidence that we are doing a better job estimating individual probabilities.

For adults, we can divide the single-person population based on any set of demographic attributes. We start by using gender as our example. To measure the relative amounts that women and men, respectively, watch each channel, we compute the log odds ratio (LOR): the log of the ratio of the odds of a woman watching a channel to the odds of a man watching that channel. We also compute the log odds ratio for a different demographic binary variable: age 18-45 versus age 46+ (this could be extended to any feature that a user would like to differentiate upon). Denote the log odds ratio for gender, for channel c as $LOR_{G,c}$, and the log odds ratio for channel c for age as $LOR_{A,c}$.

The bars in Figure 2 show some channels ranked by metric $LOR_{G,c}$ for gender (we have ignored individuals without a female or male label, who make up a very small minority of the sample). Positive bars (toward the bottom of the plot) represent channels where the log odds of a woman watching are higher than those of a man watching. These channels include Oprah Winfrey Network, Lifetime, and Lifetime Movie Network, all of which are channels that explicitly target women with

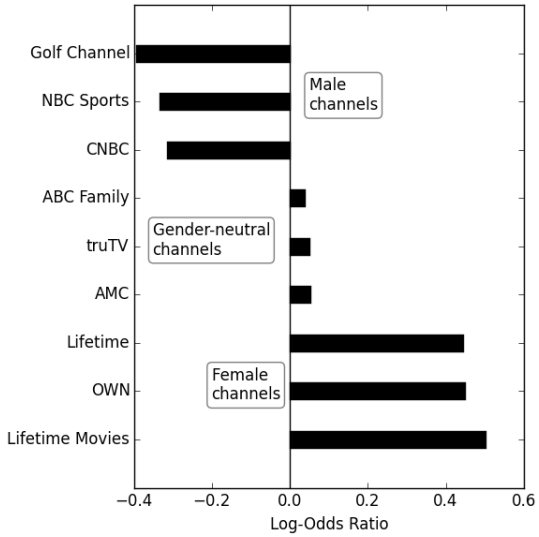

Figure 2: The bars represent the log odds ratio for women vs. men watching channel c, $LOR_{G,c}$ for a few selected channels. The LOR produces an order that is roughly consistent with what one might expect women or men to watch more. See Appendix A for descriptions of the channels and demographic groups they are typically associated with.

their programming. Negative bars (at the top of the plot) represent channels where men in single-person households are more likely to watch than women in single-person households. These channels include CNBC (a business news channel), the Golf Channel, and NBC Sports.

Figure 3 shows channels ranked in decreasing order of $LOR_{A,c}$. Channels that younger adults are more likely to watch include Nickelodeon, MTV, and Disney XD. VH1 broadcasts music and reality show programming, and Disney XD and Nickelodeon are for children and teenagers. Channels that older people are more likely to watch include Fox Business Network and Fox News (news networks) and AMC (which broadcasts classic movies and "prestige" TV shows like "Mad Men").

The purpose of this task is to ensure domain validity so that the predictions generated will be acceptable to stakeholders. We utilize similarity to the single-person log odds ratios across channels because one common practice is to only use the single-person households for demographic estimation. This is because such predictions are viewed by stakeholders as being more intuitive than those generated by WHE. In general, we will consider a model to be better if its aggregate assignment to individuals within multi-person households tracks closely to the gender and age log odds ratios as shown in Figures 2 and 3. Therefore, we compute the average relative difference to these single-person benchmark amounts across all of the channels for each model.

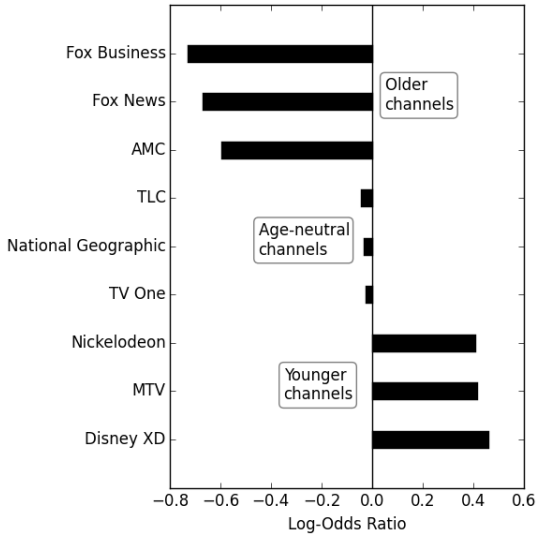

Figure 3: Similar to Figure 2, the bars represent the log odds ratio for younger vs. older people watching channel c, $OR_{A,c}$ for select channels, in descending order. Again, the channels are in an order that is roughly consistent with what one might expect older or younger people to watch more. See Appendix A for descriptions of the channels and demographic groups they are typically associated with.

NB: despite its utility for evaluation, we would expect the single-person STB distribution of viewership by gender (or age) to differ from the true distribution of viewership in multi-person STBs for the reasons mentioned in Section 2.2.2. However, this distribution is a useful proxy for domain validation via external stakeholders.

3.3 Total Audience

The final task that successful models must perform correctly considers the total expected predicted audience (across all demographic groups). While the actual audience in the test period is unobserved, it is possible to calculate the maximal possible audience that each channel could attain (as a reminder, N_i is the number of people associated with STB i):

$$\text{MaxAudience}_c = \sum_{i=1}^{N} N_i b_{c,i} \tag{15}$$

The maximum audience totals (as a percent of total people in the sample) are shown in Figure 4 (ESPN, a sports channel, has disproportionately high viewership). Each model's total expected audience for each channel is given by:

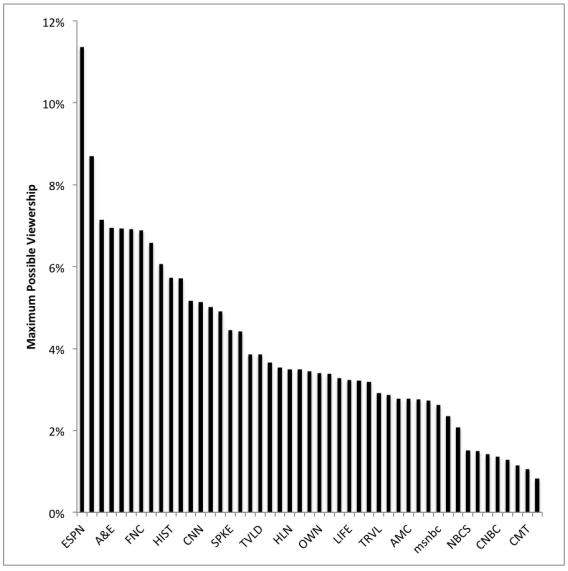

Figure 4: Each bar represents the total possible audience for each channel, and they are sorted in decreasing order of popularity.

$$\text{ExpectedAudience}_c = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \mathbb{P}(y_{c,i_j} = 1) \tag{16}$$

Each model's expected audience should certainly be less than the maximum possible audience for channel c: it is not possible that people associated with STBs that weren't tuned to channel c during the test period would watch. Furthermore, even though it is technically possible, it would not be reasonable to estimate that everyone in the audience watches TV all the time.

In summary, we want to develop models that will perform competitively on *all* of our evaluation tasks: predictive performance, similarity to demographic benchmarks, and reasonable total audience estimate. Pursuing only predictive performance may lead to incorrect attribution, since it is a proxy for what we actually want to evaluate; however, pursuing only similarity to demographic benchmarks may lead to misleading estimates for groups whose viewership differs significantly from that found in single-person households. In either case, the total estimated audience should be realistic—a successful model should not predict that more than the possible audience will watch any channel (or, similarly, an unreasonably small audience).

It is also important to note that these tasks only utilize the STB viewership data. Relying on external data would defeat some of the purposes of this research: it can be expensive and unreliable. However, other external data may be collected or purchased to enhance the credibility

of this evaluation suite. Exploring the value of external data for this predictive task would make an interesting future research topic.

4 Experiments

This section describes the experimental set-up and shows the results of computing estimates of the probability that each member of a multi-person household will watch each channel using the WHE, SPHE, and ME models on our data. We measured each model's performance under each evaluation technique.

4.1 Data

This section describes the details of the data setting in which we apply our framework. In particular, we describe the data sources, the formulation of target variables and feature vectors, and the way in which we set up a train/test split for evaluation.

Our data are from a major IPTV entertainment provider and contain anonymized, individual-level viewership for a large U.S. state for a time period of six months. These data are generated by about 215,000 Set-Top-Boxes (STBs), the device through which users access television programming. The STBs are distributed across about 75,000 households. We have information about the channel that each STB was tuned to, at what time, and for how long.[5] We also have a separate data set which contains demographic information regarding the individuals in each household. In order to not bias our results against the SPHE, we filter the data to comprise households containing only adults.[6]

There are a few assumptions that we make in order to shape the construction of training and test sets from the viewership data. We are interested in computing viewership probabilities in specific individual households, so first of all we assume that all households are present in our sample for the entire time period. Second, we want to ignore any seasonal effects beyond the weekly viewing patterns. This is somewhat necessary since our data only encompasses twenty-three weeks. Therefore, while any household's overall viewership may change over time in the data, in modeling

[5]There are several other data collection and processing issues laid out by Balakrishnan et al. (2012), from which we draw our setting.

[6]We do see that the results are qualitatively similar if we include households containing children.

we assume that individual probabilities of watching are constant with respect to time (this may be a source of error in our models). We place the first 15 weeks of data into the training set, the next four into a validation set which is used for tuning of various parameters, and leave the remaining 4 weeks as a test set.

The second component of the set-up is formulation of viewership per household and channel as a target variable. As mentioned above, we use binary labels, setting $b_{c,i}$ to 1 if the channel was watched more than 30 minutes per week on average over all of the weeks and 0 otherwise.[7] There are 47 channels in our data.[8]

4.2 Results: Predictive Ability

Figure 8 summarizes AUCs across channels resulting from WHE, SPHE, and ME. WHE, which amounts to using the b_i value from the training data as the prediction for the test data, has high AUC values (average across channels is .77). This implies that households are relatively consistent in their viewing patterns—knowing the probability that STB i will watch a channel in one time period gives a lot of information on the probability that it will watch the channel in a different time period.

SPHE performs substantially worse than WHE, with an average across the channels of .58. The reason performance is so low for this model is that demographics alone simply are not a very good predictor of individuals' behaviors. There are so many channels that a person from a particular demographic segment could watch, but each individual only chooses to watch certain channels. This is why the ME is so important: it takes the SPHE which is based only on demographics and then tunes it to the habits of each particular household.

The ME modeling formulation leads the STB-level $b_{c,i}$ predictions it generates to be identical to those of WHE. Incorporating STB history improves upon the SPHE predictions dramatically. Performing a t-test for difference in means shows that SPHE performs statistically significantly

[7] Separately, we also tested assigning 15 vs. 4 vs. 4 weeks randomly to the train/valid/test sets for each household, aggregating over STBs within each household so that we would be modeling using a complete picture of each household's viewership, and varied thresholds for binarizing the viewership. In all cases, the results were qualitatively very similar to what we present below.

[8] The same modeling and evaluation techniques could be applied to more fine-grained targets than channels; for instance, we could also divide into dayparts (ad insertions are typically purchased at a channel-daypart level). A daypart is a chunk of time within a week used for purchasing ad insertions, for instance, "Monday between 6PM and midnight."

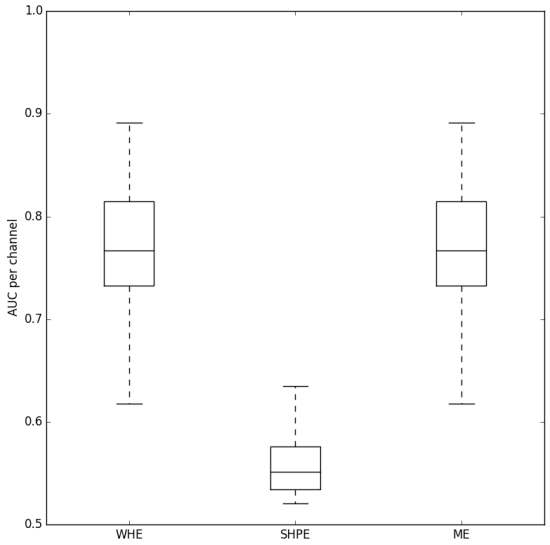

Figure 5: AUC for multi-person households in test weeks for WHE (predictions for all individuals are the household label in training weeks), SPHE (predictions for individuals are based on similar individuals' labels in single-person households), and ME. WHE and SPHE of course have excellent performance using this metric, but SPHE performs very poorly.

worse than both ME and WHE ($p < .01$).

Based on the above metric (household-level predictive ability), one might ask if we could just use WHE; however, note that WHE does not actually make **any** estimates that differentiate who is watching, which is the task that we have set out to accomplish!

4.3 Results: Demographic Benchmarks

Figure **??** plots the various models' log odds ratios using predictions from the various models as described in Section 3.2. In these figures, each point represents an estimate for one channel. The x-axis represents LOR computed for actual viewership in sinlge-person STBs among women vs. men and, separately, older vs. younger adults. The y-axis values are the odds-ratio values computed from the various models. The closer the models' estimates are to the single-person values, the better (so the $y = x$ line is there as a visual guide). To summarize the performance of each model on each demographic binary, we also measure the mean absolute percent error (MAPE) between the model's estimates and the single-person benchmark.

For gender, WHE assigns viewership roughly equally to each gender, across all channels, in multi-person households. This is evident in Figure 6, where WHE looks relatively flat across all

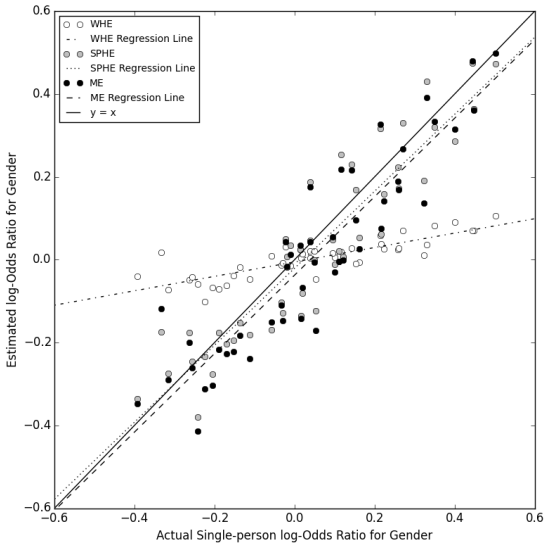

Figure 6: Actual single-person log odds ratios versus log odds ratios in multi-person households estimated by various models for gender. SPHE and ME are much better at separating women and men's viewership.

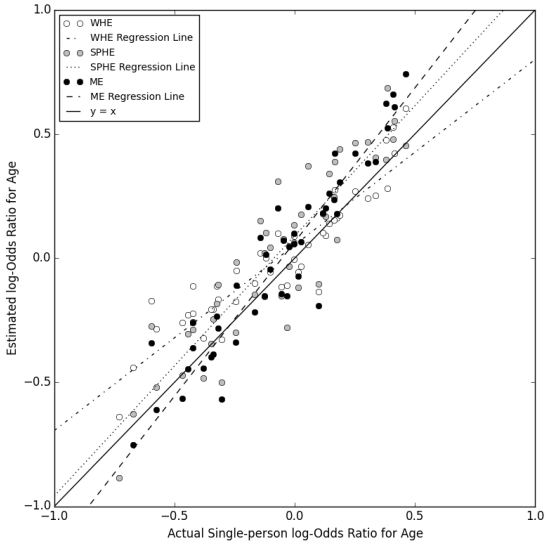

Figure 7: Actual single-person log odds ratios versus log odds ratios in multi-person households estimated by various models for age. All models perform similarly at separating younger from older viewership when all households are evaluated.

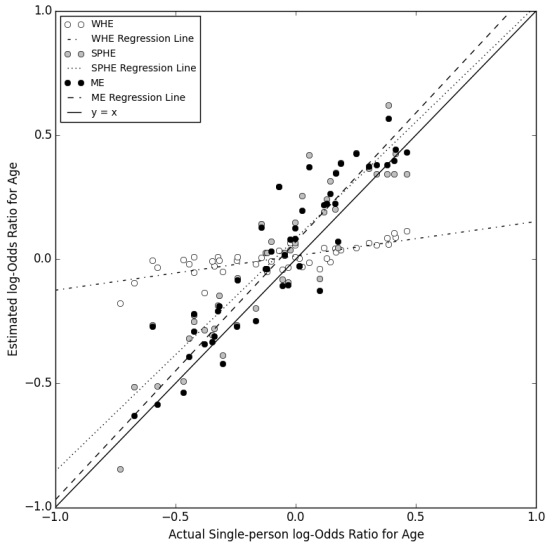

Figure 8: Actual single-person log odds ratios versus log odds ratios in multi-person households estimated by various models for age (only including households containing at least one younger and one older person). When only households containing a younger person AND an older person are included, WHE's weakness is revealed.

channels. This is because households with more than one adult are overwhelmingly more likely to contain a man and a woman than any other configuration. The MAPE for WHE versus the single-person estimates across all channels is 15.3%. SPHE and ME represent a vast improvement in terms of differentiation between genders. Both of these models are much closer to the actual single-person distribution of female-male viewership. The MAPEs for SPHE and ME, respectively, are 8.2% and 7.5%. A t-test for difference of means shows that we cannot reject the null hypothesis that SPHE and ME are not statistically significantly different in terms of their MAPEs, but that ME is significantly better than WHE ($p < .01$).

With respect to age (log odds ratio of a young person watching versus an older person watching), all of the estimators' assignments correspond fairly closely to the single-person distribution, and all are similar to one another (see Figure 7). WHE likely did a good job because adults are likely to live with other adults of similar age. That is, the vast majority of households containing "young" members contain only other young people, and vice versa.

If we filter, however, to only include households that contain at least one younger adult and at least one older adult, as in Figure 8, we can see that again, WHE does not distinguish between members of the two different groups. ME has MAPE 11.6%, SPHE has MAPE 12.8%, and WHE

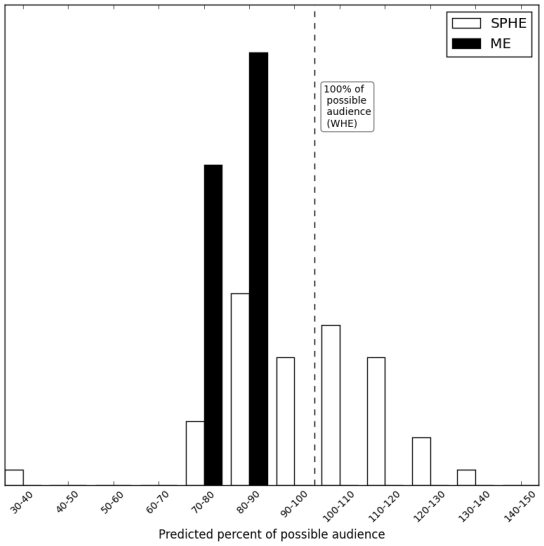

Figure 9: Total estimated audience as a percent of the possible audience for SPHE and ME (WHE estimates all of the possible audience is watching, and so is represented by the dotted vertical line). ME is the only model that provides reasonable estimates; SPHE estimates that many channels will have a total audience that is greater than is possible. WHE yields a ratio of exactly 1 for every channel; this is probably also not realistic.

has MAPE 22.6%. The difference between ME and SPHE is not significant at $p = .01$, but the difference between SPHE and WHE (and between ME and WHE) is ($p < .01$).

4.4 Results: Expected Audience Estimates

Figure 9 summarizes for each model the ratio

$$\frac{\text{ExpectedAudience}_c}{\text{MaximumAudience}_c} \tag{17}$$

for each of the 47 channels in our data. Recall from Section 3.3 that we don't know the true ratio (since we don't have ground truth labels for who is actually watching), but that this ratio should be strictly less than 1.

The vertical dotted line represents a ratio of 1; note that this is the ratio that WHE would attain (assuming the end users know the true STB label values $b_{c,i}$ in the test period). SPHE generates predictions of unrealistically high total audience for nearly half of the channels. SPHE estimates that 22 out of the 47 have a higher audience than is actually possible. Additionally, SPHE has a few estimates that seem unreasonably low. On the other hand, ME provides much more realistic

Table 3: Result summary

	WHE	SPHE	ME	WHE Best	SPHE Best	**ME Best**
AUC	.77 (.06)	.58 (.04)	.77 (.06)	✓		✓
Gender Ratio	.15 (.11)	.08 (.06)	.08 (.06)		✓	✓
Age Ratio	.11 (.09)	.14 (.09)	.13 (.10)	✓	✓	✓
Gender Ratio (Pairs Only)	.18 (.13)	.08 (.05)	.09 (.06)		✓	✓
Age Ratio (Pairs Only)	.23 (.16)	.13 (.09)	.12 (.09)		✓	✓
Audience Ratio <1 Channel Count	0	25	47			✓

audience estimates, ranging from 70-90% of the maximal possible audience. This highlights another reason why ME is valuable: it is the only model that provides a realistic estimate of total audience for each channel. It also illustrates a possible reason why the industry is reluctant to move away from the audience estimates provided by WHE: it consistently overestimates the viewership (and therefore the value of advertising) for any demographic group or channel.

4.5 Result Summary

In summary, WHE has high predictive power at the household level, but does not distinguish among individuals within households; SPHE does assign different probabilities within households, but its predictive power is low because it does not leverage the STB's history of viewing. Neither WHE nor SPHE provide reasonable predictions of what the total audience for each channel will be, when compared to the maximal possible audience. ME incorporates the strengths of each of the two baseline models and therefore is the only method in our experiments that succeeds at all of the tasks.

Table 3 summarizes the mean and standard deviation (in brackets) for each model's performance on each of six tasks. WHE succeeds (by either having the best performance or being statistically indistinguishable from the best) at only two tasks. SPHE succeeds at four tasks. Only ME shows successful performance on all six tasks.

5 Social Analyses

An important benefit of doing this work is that it facilitates a deeper understanding of how people watch television. This section describes a few examples of the sort of analysis that is possible if we have estimates of which individuals are watching TV.

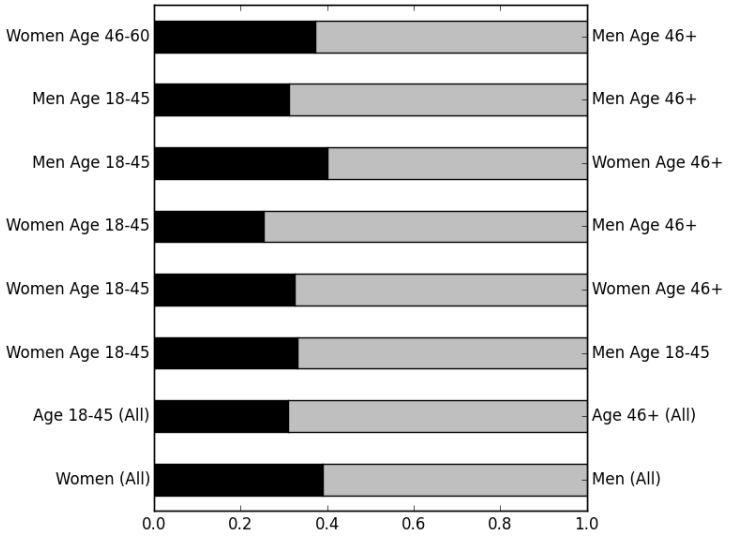

Figure 10: In STBs containing at least one representative of each group, which person gets assigned a higher coefficient?

5.1 Most Dominant Groups

Which demographic groups tend to be more dominant in households? One of the consequences of the Mixed Estimator is that it provides a coefficient corresponding to each individual in the STB through the vector β_i. These coefficients can roughly be interpreted as the degree to which that individual's prior viewing probability (based on the single-person STBs) correlates with the STB's overall viewership.

Using the households that contain various pairs of types of people (men and women, or younger people and older people), we count the number of STBs where a member of each group had the higher coefficient value. Figure 10 summarizes the number of "wins" for each demographic group in various types of pairs.

Interestingly, it appears that men's tastes dominate women's in households containing at least one of each. This concurs with results showing that wives' viewing patterns are more dependent on their husbands' than vice versa (Yang et al. 2006). Similarly, older people's tastes tend to dominate younger people's tastes.

Table 4: Differences for demographic groups in single-person households vs. when living with a member of a different demographic group.

	Watch Less	Watch More
Women with men	VH1	NBC Sports
	TV One	Golf
	BET	Fox Business
	Lifetime Movies	Speed
	Oprah Network	National Geographic
Men with women	VH1	Fox Business
	Style	Fox News
	E!	NBC Sports
	BET	TV Land
	Bravo	Golf
Age 18-45 with age 46+	Disney	AMC
	Nickelodeon	CNBC
	Cartoon Network	Syfy
	MTV	USA
	E!	Lifetime Movies
Age 46+ with age 18-45	Discovery	Syfy
	Fox Business	Family
	ESPN2	Lifetime
	Fox News	AMC
	E!	TV One

5.2 Living Partners' Effects on Viewing

How does living with different types of people alter people's viewing behaviors? We computed individuals' expected probabilities of watching in multi-person households using the ME estimates, conditioned on various household set-ups, and compared a few groups.

Table 4 shows the effect that living with someone of another demographic group has on individual's viewing probabilities. The first row shows the channels that women are most and least likely to watch when they live with at least one man, versus what they watch when living in household that do not contain a man. For instance, women who live with at least one man watch significantly less VH1 than they do in female-only households, and significantly more NBC Sports.

From the first two sections of Table 4, it is apparent that when they live in households together, both women and men watch less programming that is associated with celebrities, entertainment, fashion, and African-American culture. Conversely, both men and women watch more sports-oriented programming such as NBC Sports and the Golf Network, as well as more conservatively-oriented news programming such as Fox News and Fox Business. Younger adults who live with

Table 5: Which channels are most likely to be watched by all household members? Or alone? Descriptions of these channels are given in Appendix A.

Likely to be watched by all	**Likely to be watched alone**
Animal Planet	Fox Business Network
Food Network	Style
Biography	Lifetime Movie Network
truTV	VH1
Cartoon Network	Lifetime

older adults seem to watch less of youth-oriented channels such as Nickelodeon and MTV, and more movie channels. Older adults who live with younger adults watch less conservative news, and more entertainment-oriented channels. These hypotheses would require a more thorough analysis to fully understand, but show the intriguing directions that future research could take.

5.3 Group Viewing vs. Alone

We are interested in investigating both the social and individual dynamics of TV viewership. The description of ME shows that it is possible to compute joint probabilities, and thus to compute the probability that a person is watching TV alone, versus with other people.

We computed the joint probability of all household members watching versus the average probability of each household member watching alone. Table 7 describes the top 5 channels that are likely to be watched by all household members, versus the ones that are the most likely to be watched alone.

The channels that are most likely to be watched by all household members include Animal Planet, Food Network, Biography, truTV, and Cartoon network. These channels all seem to show content which is broadly palatable to people of all ages and genders, as well as being sufficiently inoffensive that they could be watched in groups (parents and children could watch them together, for instance). The channels most likely to be watched alone are Fox Business Network, Style, Lifetime Movie Network, VH1, and Lifetime. These are among the most polarizing channels by gender and age in the single-person households, and so it makes intuitive sense that they would be watched alone.

Similarly, we re-analyzed the results from Section 4.4 and found the channels watched by the highest percentage of individuals in each STB, given that the STB was tuned to those channels. The

highest-scoring channels on this metric included Animal Planet, truTV, TV Land, Food Network, and Bio. The lowest-scoring channels include VH1, Style, Lifetime Movies, Lifetime, and Fox Business Network. Thus (unsurprisingly), the two metrics were nearly in agreement on top and bottom-scoring channels.

One application for this type of work may be in recommender systems. A well-explored problem in that field is that of making recommendations to groups rather than to individuals (Jameson 2004). One challenge is that of understanding groups' preferences (Masthoff 2004), (Masthoff 2011). This analysis helps by estimating household members' implicit preferences, taking into account the household-level preferences.

6 Related Work

6.1 TV Audience Prediction

There is a long history of using Nielsen or other panel data to make inferences about television audiences. Example uses of such data include forecasting when audiences will watch (Liu 2010, Meyer and Hyndman 2006, Rust et al. 1992, Weber 2002); optimizing ad placement to maximize the desired audience (Abe 1996, Currim and Shoemaker 1990, Horen 1980, Rust and Eechambadi 1989); and understanding viewers' behaviors as social phenomena (Chaney et al. 2014, Yang et al. 2006).

Modern algorithms for understanding viewing behaviors rely on data from STB devices (Balakrishnan et al. 2012, Kitts and Au 2014) or DVRs (Spangler et al. 2003). Of course, with the rise of viewership on streaming devices and the increase in social media data, there are inference methods using these newer forms of technology (Hill and Benton 2012b,a).

6.2 Multi-Instance Learning

Our problem is closely related to the subfield of machine learning called Multi-Instance Learning (MIL). In MIL, individual instances are grouped into "bags." There are different feature vectors associated with each instance, but the instances themselves are unlabeled. Instead, the bags are labeled (Dietterich et al. 1997). In the framework for our problem of interest, households constitute the bags which are made up of individual viewers.

Many algorithms have been developed for solving the basic MIL problem. These include probabilistic approaches such as Diverse Density (Maron and Lozano-Pérez 1998) and Multiple Instance Logistic Regression (Xu and Frank 2004), algorithms similar to kNN that leverage similarity among instances such as Citation kNN (Wang and Zucker 2000), and heuristics that convert multi-instance problems into single-instance problems (Dong 2006). There are also other, more complicated algorithms for MIL problems related to ours; for instance, algorithms that are specifically designed to yield predictions for the individual instances that make up the bags (Liu et al. 2012), and algorithms that break the assumption that instance labels within bags are independently drawn (Zhou et al. 2009). Further, there are substantial differences between our setting and that typically found in MIL problems: bags are normally larger and less numerous than the STBs we have, and there is systematic heterogeneity in the demographic constitution of STBs.

However, although our problem is related to MIL, there are a few key differences that necessitate slightly different modeling. First, the setting is different: existing research assumes that predictions will need to generalize *across* bags. That is, the use case will be to predict the labels of bags not appearing in the training set. Our goal, instead, is to make predictions for the same STBs, but during a different time period. This additional structure allows us to use within-STB information for modeling. We have developed a method that is intended to generalize out-of-time, rather than across bags. Second, while we do evaluate STB-level predictions, we are actually more interested in making inferences about individual instances. Unfortunately, evaluation of the small subset of MIL algorithms that are designed for making individual inferences such as done by Liu et al. (2012) is usually facilitated by using either labeled or simulated data. In Section 3 we developed an alternative metric for evaluation that sidesteps the need for individual instance labels. Thus, existing MIL techniques are designed for a different setting, and should not be expected to be the best designs for our setting.

6.3 Domain Adaptation

A third way that our problem differs from many application areas in multi-instance learning is the presence of a special class of STBs: those containing only one instance. The domains frequently given as examples in MIL research do not refer to bags containing only one instance. Single-person households, however, are common. Further, these single-person households are quite informative

in that we know which household member is watching: in most cases, it is the sole resident. Thus, we can view the set of single-person households as a relatively cleanly labeled set of data. However, there are differences in viewing behavior between even demographically similar individuals living alone versus with other people (Mora 2010). This is an instance of a domain adaptation or transfer learning problem. We extend our understanding of behavior within single-person households to individuals within multi-person households.

Domain Adaptation refers to learning models using data from one domain with the intention of applying the resultant models to a separate but related domain (Daume III and Marcu 2006). One situation in which domain adaptation is advantageous is when labeled data do not exist in the target domain, but there do exist labeled data in a related domain (known as the "source" domain, or sometimes "out-of-domain" data). We can consider the target domain here to be the universe of multi-person households, for which we do not have labels of which person is watching. Therefore, the source domain here would be the single-person households.

Transfer learning is very closely related to domain adaptation and has been used previously to accomplish multi-instance learning. In (Raykar et al. 2008), the authors develop a generative solution for transferring knowledge across domains. In (Zhang and Si 2009), the authors formulate Multi-Instance Transfer Learning (MITL) as a non-convex optimization problem and develop a procedure for solving it. Finally, the authors of (Kotzias et al. 2014) develop an objective function for propagating labels among similar instances and enforcing a multi-instance relationship for the bags in the sentence/review sentiment problem.

The research in the existing literature assumes that the bags in the source domain will be labeled, but the instances are not. Our work differs in that we assume the existence of this special class of bags where we know the labels of the instances—the single-person households. Further, the methods developed in the existing literature don't leverage individual bags' histories to develop bag-specific models, as we have.

7 Discussion and Conclusions

This paper has described a novel method that combines modeling from a proxy population (single-person households) with adaptation to the target population to compute estimates of individual

viewing probabilities within multi-person households, when the ground truth is never observed. The method succeeds by leveraging the proxy population of STBs which only contain one person to distinguish viewership patterns among types of individuals as well as using STB history to improve predictive performance.

This method is very fast because the estimation for each household depends only on a logistic regression with a few dozen data points. The most time-consuming part is learning the single-person probability estimates; however, this part of the process can be done once and then updated infrequently. It does not take much time to learn the individual STB-level logistic regression models either when new STBs enter the sample, or as time passes for the existing STBs.

There are at least three key applications for being able to accurately estimate the probability of individuals within a household watching a particular show or channel. The first is "media planning," by which we mean here the practice of providing a set of recommended channels/times for paid ad insertions. Estimates of individuals' viewing can be aggregated and used in media planning. Second, there are other users for these predictions, such as the content providers (for use in alternative analytics, such as designing new programming to appeal to their current audience). Third, viewership probability estimates may eventually be useful for *addressable targeting*, which means individually targeting households to receive ads. Addressable targeting has long been promised by ad networks and has even been implemented by some (Vranica 2010); however, widespread adoption is still pending (Perlberg 2014).

The model also has additional applications outside the world of advertising. Television viewing was the most popular leisure activity in the US in 2015: the average American spends close to 3 hours per day watching TV, and on any given day, about 80% of the population watches at least some TV (U.S. Bureau of Labor Statistics 2015). In Section 5, we demonstrated examples of the type of analyses that could be done to facilitate a greater understanding of the way people watch TV.

Attempting to individually target devices that are shared by multiple people is an issue that is present to some extent in online settings as well and so has broad practical applications. For instance, multiple individuals may share a web browser, so online advertisers may not serve ads to the individual for whom they are meant. Further, online retailers make recommendations for additional products to purchase based on what the user has purchased before. If several users

share an account, disambiguating which one is doing the shopping could help to make better recommendations. Another related application is in making recommendations for what to watch for groups versus individuals (Chaney et al. 2014), (Jameson 2004), (Masthoff 2011), (Masthoff 2004).

A major limitation of this work is that it does not account for children in any way. Children are an especially difficult demographic: there are no households that consist only of children. Also, most available demographic data doesn't contain information about children because of legal restrictions. However, the framework that we have developed is sufficiently flexible that it can be modified to incorporate children. A simple heuristic way to incorporate kids into these models is to use the single-person probabilities for adults as they are in SPHE and ME, and incorporate a separately learned probability for kids. Preliminary experiments using this method show promising results.

Another limitation is that the Mixed Estimator does not incorporate potentially important information. For example, many households include more than one STB. The viewership of these STBs will clearly not be independent. Future work will include developing a Bayesian model that still has the same strengths as the ME heuristic, following the intuition that the single-person probabilities can be viewed as a loose prior, but accounts for further dependencies. Another desired property would be for the model to generalize well to previously unseen STBs, in addition to generalizing out-of-time for the same set of STBs.

In conclusion, the key contributions of this paper are: (1) the development of a method which leverages complementary strengths of existing heuristics; (2) the development of an evaluation suite which allows us to compare performance in the absence of ground-truth labels; (3) the evaluation of two state-of-the-art heuristic estimation methods for this problem as well as the novel method; and (4) the facilitation of behavioral analysis of TV watching decisions. It is important to note that we do not claim our new method is the best possible one; however, the developed framework is highly flexible, computationally efficient, and is an improvement over currently-used methods. Most importantly, this paper addresses a novel problem in television advertising that has not been explored in prior literature and has broad practical implications.

A Channel Descriptions

Table 6: Descriptions of 47 channels in our data set, summarized from individual Wikipedia pages (wik) and other online sources (pgm).

Channel	Description
A&E	Reality series, documentaries
ABC Family	Movies and series for and about families
AMC	Classic movies, drama series (Mad Men, Breaking Bad)
Animal Planet	Shows about animals
BET	African-American music, entertainment, and news
Biography	Stories about famous and historical people
Bravo	Arts, entertainment, and pop culture
Cartoon Network	Cartoons (for kids during the day, adults at night)
CNBC	Business news
CNN	News
Comedy Central	Comedy news, animation, variety; stand-up
Country Music Television	Country music reality and scripted series
Discovery Channel	Non-fiction entertainment
Discovery Science	Science programming
Disney XD	Youth-oriented animated and live-action shows
E!	Celebrities, entertainment, and Hollywood coverage
ESPN	Sports
ESPN2	Live and original sports shows
Food Network	Food shows: cooking, pop culture, and travel
Fox Business Network	Business and financial news
Fox News Channel	News
FX	General entertainment, movies, NASCAR
HGTV	Home building, decorating, gardening, and crafts
History Channel	Shows about history
HLN	Headline News
Lifetime	Entertainment and information for and about women
Lifetime Movie Network	Made-for TV and theatrical movies and mini-series for and about women
MSNBC	News
MTV: Music Television	Music and pop culture programming
National Geographic Channel	Adventure, exploration, culture, and natural science programming
NBCS	Live sports coverage programming about sports and outdoor
NFL Network	National Football League and football coverage
Nickelodeon	Kids' programming
OWN	Oprah Winfrey Network
Speed Channel	Motor sports and automotive
Spike TV	Horror, sci-fi, and fantasy
Style Network	Personal style and fashion
Syfy	Science fiction and fantasy
TBS	Sitcoms, reality shows, movies
The Golf Channel	24-hour golf coverage
The Travel Channel	Travel information and stories
TLC (The Learning Channel)	Stories about real people and experiences
truTV	Reality shows and true crime
TV Land	Classic and recent TV series
TV One	African-American lifestyle and entertainment
USA Network	Comedic and scripted dramas
VH1	Music and pop culture series, movies, and reality shows

Table 7: Channel Demographics, summarized from individual Wikipedia pages (wik) and other online sources (pgm).

Channel	Target Audience Notes	Single Person Male %
A&E	49% male; age 25-54; affluent viewers	44%
ABC Family	18-49, young adults	42%
AMC	59% male; 25-54	47%
Animal Planet	50% male, median age 43; pet owners	46%
BET	African-Americans age 12-54	34%
Biography	50% male, technologically advanced, age 18-54,	44%
Bravo	50% male, median age 45, LGBT	40%
Cartoon Network	70% kids and teens, 30% adults 18-49	43%
CNBC	60% male; age 25-54	42%
CNN	57% adults 25-54	59%
Comedy Central	66% male; age 18-34; upscale adults	44%
Country Music Television	66% male; median age 43.7	51%
Discovery Channel	50% male; age 18-49	51%
Discovery Science	50% male	56%
Disney XD	Boys age 8-14	44%
E!	50% male; age 18-54	41%
ESPN	72% male; age 18-54	54%
ESPN2	79% male	52%
Food Network	35% male; age 25-54	42%
Fox Business Network		57%
Fox News Channel	60% male; age 25-54	49%
FX	51% male; age 18-49	47%
HGTV	30% male; age 25-54	38%
History Channel	75% male; age 25-64	49%
HLN	Millenials	38%
Lifetime	23% male; age 18-49	35%
Lifetime Movie Network	28% male; age 18-49	34%
MSNBC	57% male; age 25-54	45%
MTV: Music Television	age 12-34; median age 21	43%
National Geographic Channel	55% male; age 25-54	51%
NBCS		52%
NFL Network	male-skewed	55%
Nickelodeon	70% kids age 2-11, 30% adults age 18-49	41%
OWN	mainly women; age 25-54	35%
Speed Channel	82% male	55%
Spike TV	62% male; age 18-49	48%
Style Network	29% male; age 18-49	38%
Syfy	55% male; age 25-54	46%
TBS	age 18-54; median age 36.5	47%
The Golf Channel	77% male	60%
The Travel Channel	55% male; median age 46.2	49%
TLC (The Learning Channel)	52% male	39%
truTV		44%
TV Land	45% male; age 25-54	38%
TV One	40% male; age 18-49	34%
USA Network	49% male; age 18-54	40%
VH1	42% male; median age 25.8	36%

References

PG media cable network profiles. `http://www.pgmedia.tv/news_profiles.html`. Accessed: 2016-10-20.

Wikipedia. `http://www.wikipedia.com`. Accessed: 2016-10-20.

Makoto Abe. Audience accumulation by television daypart allocation based on household-level viewing data. *Journal of Advertising*, 25(4):21–35, 1996.

Suhrid Balakrishnan, Sonik Chopra, Douglas Applegate, and Simon Urbanek. Computational television advertising. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 71–80. IEEE, 2012.

Bill Carter and Emily Steel. TV ratings by Nielsen had errors for months. *Wall Street Journal*, October 2014. October 10, 2014.

Allison JB Chaney, Mike Gartrell, Jake M Hofman, John Guiver, Noam Koenigstein, Pushmeet Kohli, and Ulrich Paquet. A large-scale exploration of group viewing patterns. In *Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video*, pages 31–38. ACM, 2014.

Imran S Currim and Robert W Shoemaker. Is television advertising being placed to reach product users? *Marketing Letters*, 1(3):209–220, 1990.

Hal Daume III and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126, 2006.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997.

Lin Dong. *A comparison of multi-instance learning algorithms*. PhD thesis, The University of Waikato, 2006.

Shawndra Hill and Adrian Benton. Talkographics: using what viewers say online to calculate audience affinity networks for social tv-based recommendations. *Available at SSRN 2273381*, 2012a.

Shawndra Hill and Adrian Donald Benton. Social tv: Linking tv content to buzz and sales. 2012b.

Jeffrey H Horen. Scheduling of network television programs. *Management Science*, 26(4):354–370, 1980.

Anthony Jameson. More than the sum of its members: challenges for group recommender systems. In *Proceedings of the working conference on Advanced visual interfaces*, pages 48–54. ACM, 2004.

Brendan Kitts and Dyng Au. A comparison of algorithms for tv ad targeting. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 296–305. IEEE, 2014.

Dimitrios Kotzias, Misha Denil, Phil Blunsom, and Nando de Freitas. Deep multi-instance transfer learning. *arXiv preprint arXiv:1411.3128*, 2014.

Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. Key instance detection in multi-instance learning. 2012.

Xiaohong Liu. Empirically modeling of audience behavior in the television industry. 2010.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.

Judith Masthoff. Group modeling: Selecting a sequence of television items to suit a group of viewers. In *Personalized Digital Television*, pages 93–141. Springer, 2004.

Judith Masthoff. Group recommender systems: Combining individual models. In *Recommender systems handbook*, pages 677–702. Springer, 2011.

Denny Meyer and Rob J Hyndman. The accuracy of television network rating forecasts: The effects of data aggregation and alternative models. *Model Assisted Statistics and Applications*, 1(3):147–155, 2006.

Jose Domingo Mora. *Understanding the social structure of television audiences: Three essays*. PhD thesis, Business Administration: Faculty of Business Administration, 2010.

Margaret Morrison and Dean M Krugman. A look at mass and computer mediated technologies: Understanding the roles of television and computers in the home. *Journal of Broadcasting & Electronic Media*, 45(1):135–161, 2001.

Steven Perlberg. Targeted ads? TV can do that now too. *Wall Street Journal*, November 2014. November 20, 2014.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine learning*, 42(3):203–231, 2001.

Vikas C Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R Bharat Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th international conference on Machine learning*, pages 808–815. ACM, 2008.

Roland T Rust and Naras V Eechambadi. Scheduling network television programs: A heuristic audience flow approach to maximizing audience share. *Journal of Advertising*, 18(2):11–18, 1989.

Roland T Rust, Wagner A Kamakura, and Mark I Alpert. Viewer preference segmentation and viewing choice models for network television. *Journal of Advertising*, 21(1):1–18, 1992.

William E Spangler, Mordechai Gal-Or, and Jerrold H May. Using data mining to profile tv viewers. *Communications of the ACM*, 46(12):66–72, 2003.

Michael Stack. Personal communication, November 2014.

Emily Steel. The chief of Viacom says Nielsen is outdated. *Wall Street Journal*, November 2014. November 13, 2014.

U.S. Bureau of Labor Statistics. American time use survey, 2015. URL `http://www.bls.gov/news.release/atus.t01.htm`. [Online; accessed 2016-09-05].

Suzanne Vranica. Targeted TV ads set for takeoff. *Wall Street Journal*, December 2010. December 10, 2010.

Jun Wang and Jean-Daniel Zucker. Solving multiple-instance problem: A lazy learning approach. 2000.

René Weber. Methods to forecast television viewing patterns for target audiences. *Communication Research in Europe and Abroad Challenges of the First Decade. Berlin: DeGruyter*, 2002.

Xin Xu and Eibe Frank. Logistic regression and boosting for labeled bags of instances. In *Advances in knowledge discovery and data mining*, pages 272–281. Springer, 2004.

Sha Yang, Vishal Narayan, and Henry Assael. Estimating the interdependence of television program viewership between spouses: A Bayesian simultaneous equation model. *Marketing Science*, 25(4):336–349, 2006.

Dan Zhang and Luo Si. Multiple instance transfer learning. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 406–411. IEEE, 2009.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009.