CTV Reach Measurement Using Panel Data

Emil Martayan, Han Yan, Jon Vaver, Rachel Fan, Wei Liu, Uday Chaudhary

Google Inc.

Abstract

Connected TV (CTV) devices blend characteristics of digital desktop and mobile devices – such as the option to log in and the ability to access a broad range of online content – and linear TV – such as a living room experience that can be shared by multiple members of a household. This blended viewing experience requires the development of measurement methods that are adapted to this novel environment. For other devices, ad measurement and planning have an established history of being guided by the ground truth of panels composed of people who share their device behavior. A CTV panel-only measurement solution for reach is not practical due to the panel size that would be needed to accurately measure smaller digital campaigns. Instead, we generalize an existing approach used to measure reach for other devices that combines panel data with other data sources (e.g., ad server logs, publisher-provided self-reported demographic data, survey data) to account for co-viewing. This paper describes data from a CTV panel and shows how this data can be used to effectively measure the aggregate co-viewing rate and fit demographic models that account for co-viewing behavior. Special considerations include data filtering, weighting at the panelist and household levels to ensure representativeness, and measurement uncertainty.

1 Introduction

Connected TV (CTV) refers to a TV set connected to the internet. The internet connection may be provided by a Smart TV device, a digital media player (e.g., Fire TV Stick or Chromecast), or a gaming console. High-quality measurement for CTV is critical. CTV advertising spend in the US was expected to total 24.6 billion USD in 2023 and is expected to rise to 34.3 billion USD by 2025 [3]. Additionally, among advertisers who increased CTV spending in 2023, 37% reallocated budgets from linear TV [9].

CTV is a different type of digital device with unique challenges that require device-specific modeling. It differs from other digital devices (e.g., desktop and mobile devices) in that the watchtime sessions are typically longer, the content is consumed in a lean-back environment, the same device can be shared by multiple people, and there is a higher likelihood that multiple people are viewing the content together. This living room co-viewing behavior is prevalent on TV screens when people consume live linear TV content. However, in the CTV setting, users have the ability to log in or consume content anonymously, have access to more content in a broader range of formats, are served real-time content suggestions, and experience more friction in switching between content providers (c.f., channel surfing on linear TV).

More fundamentally, CTV measurement is needed for effective and efficient ad serving, accurate ad measurement for reporting, and reach planning (i.e., which demographic groups can be reached and how to reach them). These needs are addressed with the capability to determine the number of people watching and their demographics.

Well-established models have been developed to measure the reach of digital campaigns [7, 6, 8]. These models use data from multiple sources to compute audience reach metrics, including ad server logs, publisher-provided self-reported demographic data, census data, and panel data. As a consequence, it is sometimes referred to as a "hybrid approach" to reach measurement. This same approach also works for measuring CTV reach, if it is generalized to account for the prevalence of co-viewing on CTV.

More than one approach has been used to understand and characterize co-viewing behavior on CTV. In a previous paper, Fan et al. demonstrated how survey data can be used to identify variations in the co-viewing rate with respect to key variables, such as viewer demographics (e.g., gender and age), time of day, and video genre [4].

The survey used to collect co-viewing data is ongoing and consists of a single question: "including yourself, how many people are watching this TV right now?". There are five answers to choose from: "1", "2", "3", "4 or more", or "Prefer not to answer". There is also an option to skip the survey. The survey question appears before a YouTube video shown on a CTV screen by replacing the pre-roll ad that would otherwise be served before the video. For each survey response, the following are known from the ad impression log data: the demographic label of the logged-in user, the time at which the response was logged, and the genre of the video following the survey. Users cannot receive the survey more than once every 35 days, so the survey responses can be assumed to be independent observations. This survey is highly scalable across the vast YouTube CTV user base, allowing for cost-effective and time-efficient data collection on an international scale. This scale allows for a large sample size and a reliable understanding of how key variables influence the co-viewing rate. Nevertheless, the survey is a single-question format that is subject to potential non-response bias and does not provide visibility into ground truth viewer demographics or how individual viewing behavior varies across time and across different CTV apps. So, this survey is not a complete solution to characterizing co-viewing for CTV measurement.

In this paper, we describe how high-quality panel data can be used to supplement surveys in providing CTV measurement. A panel is a group of people who are recruited to be representative of a target population and consent to share behavioral data that would not otherwise be available (e.g., who is watching, how long they watch, how frequently they watch, etc.). The data is collected in a continuous manner, meaning that we obtain repeated measurements for each panelist over time. Compared to survey data, the panel data is far more comprehensive and rich. It provides viewing behavior over time and across apps, and panelist activity can be merged with demographic data to provide measurement of co-viewing demographics. However, panel recruitment can take several months or years and this limits the scale at which panel data can be collected.

The scalability of the survey data and the richness of the panel data are highly complementary. Combining the scalability of survey data with the richness of panel data provides a set of tools that makes it possible to reliably understand and measure CTV behavior. Panel data is used to estimate an overall co-viewing rate and understand demographic viewing behavior (i.e., who is watching with whom). On the other hand, survey data is used to execute more granular co-viewing rate modeling, allowing for relative co-viewing rate adjustments by country, demographics, time of day, and video genre.

In §2 we describe key elements of the US CTV panel, such as recruitment, metering, demographic composition, and weighting to ensure demographic representativeness. In §3 we delineate the ground truth data provided by the panel. In §4 we describe the modeling components needed to measure CTV using panel data.

2 About the Panel

The panel described below was recruited to provide visibility into CTV viewing behavior. The ground truth data generated by this panel plays a key role in CTV measurement.

2.1 Recruitment

The data used in the analyses described below comes from a panel that is probabilistically recruited using address-based sampling to ensure that it is representative of the US population. Address-based sampling frames are the best sampling frames available for national US household surveys [5]. During the enrollment process, prospective panelists are asked to complete a survey to describe themselves and their household and to assess eligibility. Those who have an internet connection and own a CTV device are deemed eligible for panel membership. Panelists also enroll desktop and mobile devices, which makes this a full cross-screen panel and opens the possibility for cross-device/cross-household measurement.

2.2 Metering

Panelists install digital and TV meters and are provided with a network device. The network device is used to detect CTV devices, collect YouTube content and ads, and collect signals used for user prompting. The TV meter supports a custom app and is the interface for panelists to pair their CTV devices and to attribute user-level viewership. The user interface includes a tile for each registered panelist in the home, as well as a "GUEST" tile. Users proactively indicate they are watching the TV screen, or reactively when prompted, by simply clicking on the tiles. Clicking again indicates that they are finished watching.

2.3 Demographic Composition and Weighting

During registration, panel members provide important demographic information about themselves and their household. Demographic data enables comparisons and calibration to US CTV population data. As part of the privacy policy, panel membership is restricted to those who are 13 years or older.

To ensure demographic representativeness, weights are generated to correct for any deviation from benchmarks. Weights are calculated at the panelist and household levels using different weighting algorithms and calibrated against different target benchmarks, such as the US internet population and the US CTV population. For CTV measurement, we use the empirical calibration algorithm, which is based on the generalized regression estimator (GREG), to match the panel against the US CTV population on relevant dimensions [10]. There are many calibration methods available. However, we choose the empirical calibration algorithm because it performs well at aligning the panel to the population, while having the lowest variance among competing methods in our experience [7]. The weighting algorithm works by minimizing the L2 norm of the weights vector subject to matching the marginal distributions between the sample and the target population for the demographic variables of interest. First, household-level weights are calculated using attributes relevant to CTV measurement, such as household size and household income. Next, panelist-level weights are calculated using the household-level weights as baseline weights. These weights are calculated using relevant individual-level attributes, such as gender and age. It can be shown that minimizing the norm of the weights can be expressed as maximizing the effective sample size.[1] This formulation is useful in that it is more suggestive of the bias-variance tradeoff at play: weights help to decrease the bias in our measurement at the cost of increasing the variance. To straddle this bias-variance tradeoff, we may post-process the weights and therefore not match the target distributions exactly.

In Figure 1, we showcase an application of the household-level calibration weights. We present the marginal household size and household income distributions for the population benchmark and the panel, after applying household-level weights. The weights ensure that the weighted panel distributions closely match the population benchmark distributions. Any minor differences in the distributions may be attributable to post-processing of the weights.

3 Measurement Data

Viewing activity on CTV is decomposed into constituent playback sessions. A new playback session is created any time a new video plays on the TV screen or a user indicates a change in the attributed viewers behind the screen. The latter point means that a single video may be split across multiple playbacks if the attribution data changes. When joined together, the logged playback sessions data and attribution data indicate which viewers watched a playback session. These viewers may include panelists or guests. In addition to the list of viewers, we collect the start time and end time of each playback session, which makes it possible to calculate session duration. This is particularly useful for producing watchtime distributions and for calculating watchtime-weighted statistics.

Because attribution data is collected across the length of a playback, there is no particular position in the playback session at which we measure co-viewing. In contrast, in the survey measurement setting, the measurement position corresponds to that of a preroll ad, since survey questions are served immediately before the start of a video playback. Nevertheless, it is possible to differentiate between ad and non-ad content in the playback sessions logs.

In addition to playback data, a separate demographics data table is maintained that includes key demographic attributes for each panelist and household. The variables most germane to the CTV measurement use cases include gender, age, household size, and household income, though other demographic attributes are collected as well. This table also includes panelist-level and household-level calibration weights. Because the size and composition of the panel change over time as panelists are churned and others are recruited to participate, it is important to maintain up-to-date calibration weights to correct for any deviations from population benchmarks.

Table 1 provides an example of the format of the CTV panel data. The table contains three distinct playback sessions, and each playback is duplicated in the table for each attributed viewer. Additionally, the table contains demographic attributes for the attributed panelists. Note that these columns represent a small subset of the available signals and are presented to show the basic format of the data.

Because multiple playback sessions are collected from the same households, playback sessions are not independent within a household. They are assumed to be independent across households, since the panel is a very small fraction of the US CTV population.

Cross-app measurement is also possible due to the installed network device. The network device collects important signals from paired CTV devices within the household, including the domain names corresponding to the content appearing on the screen. The processed data can be used to analyze co-

[1] Kish's effective sample size is defined as $n_{\text{eff}} = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}$, where w_i is the weight of sample i.

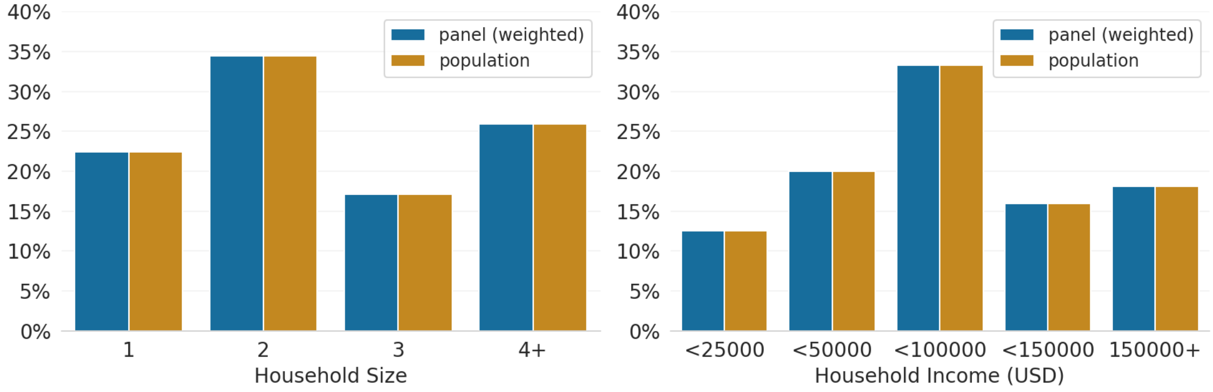


Figure 1: Marginal household size distribution (left) and household income distribution (right) for the population benchmark and the panel, after applying household-level calibration weights. The population benchmark is a Current Population Survey (CPS) household benchmark for the US CTV population. The application of calibration weights ensures that the weighted panel distributions align closely with their population analogs.

viewing behavior across non-YouTube apps, such as Netflix, Hulu, and Amazon Prime. Although, because of the differences in the data collection process, there is less accompanying information in this setting.

4 CTV Measurement

CTV has the same need for reach measurement as other brand advertising campaigns. This includes the ability to estimate the number of unique people exposed to a campaign and the distribution of these people across a specified set of demographic (gender and age) groups. However, unlike the measurement of more traditional digital brand advertising campaigns, CTV measurement needs to account for the common scenario in which multiple people are exposed to the same ad impression. That is, it needs to take co-viewing into consideration.

This section describes the modeling components needed to measure CTV using panel data.

4.1 Unique Reach

The most straightforward approach for measuring the unique reach of an online audience is via direct panel-based measurement. This approach requires a very large panel to measure just the very largest campaigns. Appendix A contains a related idealized sizing analysis showing the panel size needed to achieve reliable reach measurement within a specified error tolerance. Even for the largest digital campaigns, a large US panel comprised of 100,000 people would fail to provide reliable measurement of campaign reach. A different approach is required to measure reach across the full range of digital campaign sizes.

Generally speaking, existing methods for measuring the reach of digital campaigns for desktop and mobile devices can be applied to digital campaigns for CTV. These methods use data from multiple sources to compute audience reach metrics including ad server logs, publisher-provided self-reported demographic data, census data, and panel data. This so-called “hybrid approach” takes advantage of cross-campaign learning to fit a reach curve that can accurately estimate the reach of campaigns of any size.[2] Appendix B contains a related idealized sizing analysis that is analogous to the one in Appendix A. It shows that a “hybrid approach” provides accurate reach measurement that does not degrade as campaigns become smaller. This property allows even modestly sized panels to provide a basis for accurate measurement of reach.

The “hybrid approach” is described in the following modeling and methodology papers:

- [7] presents a methodology for measuring the reach and frequency of online ad campaigns by audience attributes for a single device or cookie type. The method produces corrected cookie and impression counts by demographic attributes and a model to map the number of cookies to the number of people reached.

[2] This reach curve can also be used to measure the reach of advertising campaigns in countries that do not have a panel. Note that this application of the model still uses the country and campaign specific signals from the ad server logs, publisher-provided self-reported demographic data, and population data as input. This avoids the costly and time-intensive process of growing and maintaining a panel in every country that has a demand for reach measurement.

playback ID	household ID	panelist ID	cookie demo	panelist demo	duration
1	H1	A	F25-34	M25-34	D1
1	H1	B	F25-34	F25-34	D1
1	H1	GUEST	F25-34	NULL	D1
2	H1	A	M25-34	M25-34	D2
2	H1	B	M25-34	F25-34	D2
3	H2	D	F45-54	F45-54	D3
3	H2	E	F45-54	F35-44	D3

Table 1: Example format of the CTV playback data. This example includes two households providing a total of three playback sessions. Each playback session is duplicated in the table for each attributed viewer. Each playback is characterized by the household ID, the list of attributed viewers (including at most one guest), the demographic labels of each panelist, and the duration of the playback. Note that the cookie demo is a playback-level attribute, meaning that all viewers associated with a single playback will have the same cookie demo label. This demographic label is predicted by the ad serving system.

- [6] extends the methodology to the cross-device setting and allows reach reporting at the level of device and cookie type. The main development in the paper is the concept of an Activity Distribution Function (ADF), which describes the probability that a person generates cookies across a set of cookie types. It also demonstrates that an ADF based on a mixture of Dirac delta functions can model an arbitrary multiple-device reach curve.

- [8] introduces the concept of Virtual People. The reach and demographic correction models described in [7] and [6] are replaced by a model that assigns a virtual person identifier to each ad event. It follows that the reach of an online audience can be estimated via a simple count of unique virtual people assigned to the corresponding set of events. Each virtual person has accompanying demographic attributes (e.g., gender and age), meaning that the demographic composition of an audience can be obtained by counting the demographic labels of the virtual people assigned to a corresponding set of events.

Audience measurement for digital CTV campaigns poses challenges not addressed previously because of the prevalence of co-viewing on CTV devices. In particular, if it is believed that multiple people are watching the TV screen, then more than one virtual person identifier needs be assigned to a single event. As a result, models are needed to estimate the number of people watching the TV screen at the impression level and to estimate the demographic attributes of all viewers. Models that make use of CTV panel data to address these needs are described below.

4.2 Average Co-viewing Rate

The co-viewing rate is a statistic providing a measure of the expected number of people watching the TV screen. It can be measured at the country level, the impression level, or any level in between. The description below pertains to the measurement of the average co-viewing rate across the entire panel (i.e., US level). Also note that the co-viewing rate can either include or exclude situations in which nobody is watching. The calculations below exclude situations in which no one is watching the TV screen because these situations are modeled separately.

Describing the co-viewing rate calculation requires some notation. Suppose we have n households and N panelists in the panel. For each household $i \in \{1, ..., n\}$, let S_i denote the set of playback sessions that household i watched. Next, for each playback $j \in S_i$, let d_j be the duration (in seconds) of playback j and let c_j be the number of attributed viewers watching playback j.

The smallest possible value of c_j is one because at least one person has to be present to provide attribution data. By extension, this means the calculated co-viewing rate will be at least one. Additionally, because the TV meter includes a "GUEST" tile, our measurement of the co-viewing rate is not limited to panelists watching the screen; it includes panelists and guests.[3] The calculation of the co-viewing rate can be modified to exclude guests from the calculation, depending on the intended use case.

4.2.1 Weighting

We can improve the estimate of the co-viewing rate by applying calibration weights. For example, if the sample contains a larger proportion of multi-inhabitant households relative to the US CTV pop-

[3] Strictly speaking, this statement is not true because there can only be at most one guest per playback.

ulation, and these households tend to have a higher co-viewing rate, then the simple estimate of the co-viewing rate will be an overestimate of the true co-viewing rate. This potential for bias is mitigated by giving a smaller weight to multi-inhabitant households and a larger weight to single-inhabitant households so that the weighted sample is more representative of the US CTV population. To generate these weights, we apply the empirical calibration algorithm [10] with quadratic loss and calibrate against the distribution of households in the US CTV population across household size and household income. We opt for household-level weights rather than panelist-level weights because we intuitively expect household-level attributes such as household size to have the most significant impact on co-viewing behavior. This algorithm outputs a weight w_i for each household $i \in \{1, ..., n\}$.

As such, we can define the duration-weighted co-viewing rate as follows:[4]

$$m = \frac{\sum_{i=1}^{n} w_i \sum_{j \in S_i} d_j c_j}{\sum_{i=1}^{n} w_i \sum_{j \in S_i} d_j}.$$

Additionally, let $x_i = \sum_{j \in S_i} d_j c_j$ denote the total co-viewer watchtime for household i and let $t_i = \sum_{j \in S_i} d_j$ denote the total watchtime for household i. These quantities give a summary of co-viewing behavior at the household level and are useful for winsorization and uncertainty estimation.

Winsorization is used to attenuate the impact that any single household can have on the co-viewing rate calculation. Let γ_q denote the q-th quantile of the distribution of household-level total watchtime values.[5] We clip each t_i to be at most γ_q; for each $i \in \{1, ..., n\}$, define

- $\tilde{t}_i = min(t_i, \gamma_q) = t_i \cdot min(1, \gamma_q / t_i)$.
- $\tilde{x}_i = x_i \cdot (\tilde{t}_i / t_i) = x_i \cdot min(1, \gamma_q / t_i)$.

The winsorized co-viewing rate is calculated by replacing each x_i and t_i with $\tilde{x}_i$ and $\tilde{t}_i$, respectively. Notice that this can be reframed as a reweighting of the data, in which each household i is given a weight of $r_i = min(1, \gamma_q / t_i)$. In our experiments, we find that even a small amount of trimming can decrease the standard error of the co-viewing rate. We set $q = 0.95$ to apply trimming to the households with total watchtime above the 95th percentile of the total watchtime distribution.

4.2.2 Measurement Uncertainty

To accurately estimate the standard error of the co-viewing rate, we need to be mindful of the correlation structure of the data caused by having repeated observations within households. We assume that playbacks are dependent within a household and independent across households. Using the delta method, we calculate an approximation of the variance of the co-viewing rate (see Appendix C for the derivation):

$$\text{Var}(m) \approx \frac{1}{n_{eff}} \left(\frac{\sigma_x^2}{\mu_t^2} - \frac{2\mu_x \rho \sigma_x \sigma_t}{\mu_t^3} + \frac{\mu_x^2 \sigma_t^2}{\mu_t^4} \right),$$

where $n_{eff} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}$ denotes Kish's effective sample size, μ_x and σ_x denote the mean and standard deviation of the total co-viewer watchtime distribution, μ_t and σ_t denote the mean and standard deviation of the total watchtime distribution, and ρ denotes the correlation between the total co-viewer watchtime distribution and the total watchtime distribution. We can estimate the variance by applying the corresponding sample estimates of these population parameters. The standard error is calculated by taking the square root of the variance estimate. The main advantage provided by the delta method is a closed-form approximation for the variance; however, since this estimate relies on a Taylor series approximation, it may not always be accurate.

To validate the analytical result above, we also implement a bootstrap resampling approach to estimate the standard error of the co-viewing rate. In particular, we use a block bootstrap that samples households with replacement from the original data.[6] The simulation procedure is the following:

```
Algorithm 1: Block bootstrap
Set the number of bootstrap replicates $B$ to
  some large number (e.g., $B = 5,000$).
for $b \in \{1, ..., B\}$ do
    Sample $n$ households with replacement
      from the original data. When a
      household is selected, include all of its
      playback sessions in the bootstrap
      sample.
    Calculate the co-viewing rate $m_b$ using
      the current bootstrap sample.
end
Calculate the empirical standard deviation of
  the $B$ co-viewing rate values $\{m_1, ..., m_B\}$.
```

[4] The session-weighted co-viewing rate can be calculated by simply setting all of the d_j's equal to one.

[5] For the session-weighted co-viewing rate, we apply winsorization to the household-level playback session count distribution (this is accomplished by setting the d_j's equal to one).

[6] We also implement a stratified block bootstrap that samples households with replacement within each combination of household size and income. This ensures that all of the bootstrap samples have the same demographic composition. However, because we obtain similar results with both approaches, we opt for a simple block bootstrap.

We construct (Gaussian) 95% confidence intervals using the delta method standard error and the bootstrap standard error. We find that both methods generally produce similar confidence intervals, with the bootstrap providing slightly tighter intervals. As such, we feel confident in both methods and choose to present the bootstrap confidence intervals in the results section below.

4.2.3 Measurement Results

Figure 2 shows the US co-viewing rate for YouTube on CTV measured at the daily level (calculated over a 28-day rolling window) along with 95% pointwise confidence intervals. The co-viewing rate is relatively stable along with the size of the confidence interval. There are some minor fluctuations in the co-viewing rate, which may be attributable to small organic changes in co-viewing behavior or to measurement noise. Also note that the error margin is small relative to the co-viewing rate. This means that we are able to confidently measure the average co-viewing rate for the US using the available panel data.

4.2.4 Applications

The average co-viewing rate is a useful statistic to understand the overall co-viewing patterns in the panel and to monitor changes over time. Furthermore, the average co-viewing rate is applied to estimate co-viewing rates at various levels of granularity. These applications include:

- Generating country-level co-viewing rates.[7]
- Monitoring for periodic or systematic changes in co-viewing behavior across time.
- Comparing co-viewing rates across CTV apps (e.g., YouTube, Netflix, Hulu).
- As a baseline for generating impression-level co-viewing rates using event attributes, such as the demographic label of the cookie, time of day and day of week, and video genre.

The last item is particularly relevant for reach measurement as outlined in the next section.

4.3 Impression-Level Co-viewing Rate

More granular co-viewing rate modeling is required to estimate the number of viewers watching the TV screen at the ad impression level. It is more efficient to combine panel and survey data to meet this modeling challenge. The application of survey data for this purpose is described in [4].

In this approach, impression-level co-viewing rates are generated by modifying the average co-viewing rate using impression-level attributes: demographic (gender and age) label of the cookie, time of day and day of week, and video genre. Millions of co-viewing survey responses are used to generate a *relative co-viewing factor* (RCF) for each attribute.[8]

Let E_d denote the RCF for demographic group d, let E_t denote the RCF for time of day slice t, and let E_g denote the RCF for genre g. The co-viewing rate for an impression with attributes (d, t, g) is given by

$$m_{dtg} = (m - 1) \cdot E_d \cdot E_t \cdot E_g + 1,$$

where m denotes the overall co-viewing rate. The co-viewing rate m_{dtg} can then be used to determine the number of viewers for the given event via random sampling. That is, if $m_{dtg} = 1.3$, for example, then the number of viewers is 1 with probability 0.7 and 2 with probability 0.3.

Combining large-volume survey data with high-quality panel data in this way makes it possible to generate impression-level co-viewing estimates across countries without the impractical and cost-prohibitive need to build a panel in every country.

4.4 Co-viewing Demographics

A model is needed to estimate the demographics of CTV viewers. Co-viewing is much more common on CTV devices than other digital devices and this poses challenges for accurately estimating viewer demographics. In this section we describe a machine learning approach to estimate CTV viewer demographics at the impression level using the CTV panel data. The goal is to demonstrate the utility of this data for fitting an effective demographic model. So, while there are many potential models and nuances that could be considered, we choose a simple model to illustrate the general approach.

[7] Non-panel countries rely on co-viewing rate estimates derived from survey data rather than panel data. However, the observed ratio of panel and survey-based co-viewing rates in panel countries can be applied to mitigate the potential response bias of co-viewing rates measured using surveys in non-panel countries.

[8] The relative co-viewing factor for data slice j is defined as $E_j = \frac{m_j - 1}{m - 1}$, where m is the overall co-viewing rate and m_j is the (calibrated) co-viewing rate for slice j.

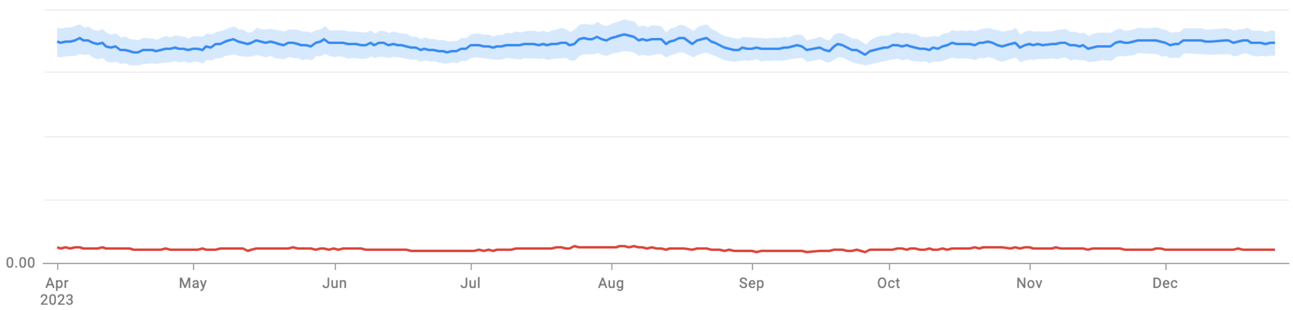


Figure 2: Time series of the daily US average co-viewing rate from 1 April 2023 through 25 December 2023. The blue line represents the co-viewing rate and the shaded blue band represents the 95% pointwise confidence interval. The red line is the error margin (i.e., the confidence interval half-width).

4.4.1 Weighting

The estimation of co-viewer demographics requires bespoke calibration weights distinct from those used in other CTV measurement use cases. In particular, we implement an algorithm that outputs a weight for each panelist, where these weights are a function of each panelist's demographic information, as well as the demographics of anyone else living in their household. The motivation for this procedure is the fact that, while a sample of individuals may be generally representative of some target population, the household-level groupings of those individuals may not be representative. For example, we may have an otherwise representative sample of young males (e.g., males 18-34) that is not representative of their living arrangements (e.g., living alone, living with a roommate, or living with parents).

Because household composition is likely correlated with co-viewing pattern – after all, living with someone provides an opportunity to view CTV content with them – it is imperative to ensure the household composition of the panel matches that of the target population. That is, we want the demographic composition of panel households to match the distribution of households in the CTV population. The population benchmark we use is a CPS Computer and Internet Use benchmark for the US, which provides person-level demographics grouped by household. This benchmark data source is not specific to the CTV universe.

Accounting for all possible demographic compositions at the household level is intractable, so we simplify the problem by considering pairs of demographic groups. In particular, we seek weights to match the distribution across these demographic pairs in the panel with the distribution from a population benchmark. The weights are designed to minimize the difference between the rates of the weighted sample and the population benchmark.

The following terms are used to describe the optimization problem that is solved to find these weights:

- N: the number of panelists.
- K: the number of demographic groups.
- D: the $N \times K$ demo assignment matrix, where $D_{ij} = 1$ if panelist i has demo j (and 0 otherwise).
- A: the $N \times K$ co-inhabitant matrix, where A_{ij} gives the number of panelists of demo j with whom panelist i lives.
- B: the $K \times K$ target matrix,[9] where B_{ij} gives the population-level probability of a co-inhabitant having demo j, given that an inhabitant has demo i.[10]
- w: the N-vector of panelist weights.

The matrices D and A are constructed using the panel data, while B is constructed using the population benchmark data. Let d_i be the i-th column of D and a_j be the j-th column of A. Notice that $(D^\top A)_{ij} = d_i^\top a_j$ represents the number of times a panelist with demo i and a panelist with demo j are observed living in the same household. Similarly, $(D^\top diag(w)A)_{ij} = (d_i \odot w)^\top a_j$ represents the weighted count.[11] The goal is to obtain a weighted

[9] In practice, we extend the procedure to account for single-inhabitant households, i.e., we include a column in the target matrix representing the probability of a single-inhabitant household (and rescale the remaining entries of the matrix). For the sake of conciseness, we omit a detailed description of this extension.

[10] Put another way, across all pairs of inhabitants (x, y) where x has demo i, this is the fraction of times y has demo j.

[11] Since d_i is a binary vector, the elementwise product of d_i and w extracts the weights of panelists of demo i, i.e., $(d_i \odot w)_\ell = w_\ell$ if panelist ℓ has demo i (and 0 otherwise).

sample matrix $D^\top diag(w)A$ that is as close as possible to the target matrix B. A regularization term is included in the objective function to ensure that the resulting weights are not too extreme. The weights w are the solution to the following convex optimization problem:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \left\| D^\top diag(w)A - B \right\|_F^2 + \lambda \left\| w - \bar{w}\mathbf{1}_N \right\|_2^2 \\ \text{subject to} \quad & D^\top diag(w)A\mathbf{1}_K = \mathbf{1}_K \\ & w \geq \mathbf{0}_N \end{aligned}$$

The regularization parameter λ provides control of the bias-variance tradeoff: a smaller value of λ puts more weight on minimizing the deviation between the weighted sample and the target, while a larger value of λ puts more weight on minimizing the (scaled) variance of the weights. Furthermore, the penalty term ensures that the problem has a unique solution, if one exists, where panelists with the same demo and household composition are assigned the same weight, which is a desirable property. Additionally, the constraints force the resulting weighted sample matrix to be row-stochastic. We use the CVXPY library to solve the optimization problem [2].

In other optimization-based formulations of calibration weighting, it is often customary to include a constraint in the optimization problem to force the weights to sum to one [10, 1]. With this constraint, minimizing the variance of the weights is equivalent to minimizing the L2 norm of the weights. In our formulation, we do not include a constraint forcing the weights to sum to one. Instead, we constrain the weighted sample matrix to be row-stochastic, since the target matrix B is row-stochastic.

We ran tests to decide whether to include a penalty on the variance or the L2 norm of the weights. These tests indicate that for each value of λ, the formulation that includes a penalty on the variance of the weights outperforms the formulation with a penalty on the L2 norm of the weights in terms of effective sample size and Frobenius distance between the weighted panel matrix and the census target matrix. As such, we opt for a penalty on the variance of the weights.

Figure 3 shows the heatmaps of the census target matrix B and the weighted panel matrix. For this result, we consider panelist age groups, but this methodology can be applied to combinations of gender and age, for example. Each row represents a probability distribution over the demographic groups. The conclusion is that the weights closely align the panel to the census target matrix.

The regularization parameter λ was set to 20 for this analysis, as this specification reasonably balances bias and variance. Figure 4 shows the Frobenius distance between the census target matrix and the weighted panel matrix for various values of λ (blue curve). Notice that the Frobenius distance decreases as λ becomes smaller, as expected. As a baseline, we show the Frobenius distance between the census target matrix and the unweighted panel matrix (gray dotted line). Additionally, the orange dotted line shows the Frobenius distance between the census target matrix and the weighted panel matrix after applying the panelist-level weights described in Section 2.3. The gap between the orange line and the blue curve represents the added benefit of applying the weights generated by this bespoke weighting procedure.

4.4.2 Model Description

The goal of the demographic model is to estimate the gender and age group of every viewer exposed to an ad impression. More specifically, it is to predict the probability distribution across the set of demographic groups for each viewer. In this application, we define a demographic distribution as a categorical distribution over the following demographic (gender and age) groups: {Female, Male}×{13-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+}. This impression-level information can be aggregated to understand audience composition at the campaign level.

CTV panel data provides a representative set of ground truth data for viewer demographics at the ad impression level and various impression-level features that can be used to support demographic predictions. We use the following features in the modeling example described below: the demographic label of the cookie, the corresponding quality label[12] of this demographic label, day of the week, time of day, and video genre. Some CTV playback sessions have multiple genres. In these cases, one of the genres is randomly assigned to the impression. Any videos that do not have a genre label are assigned to the "no genre" category.

We use these features and the ground truth demographic labels from the CTV panelists to train random forest classifiers to predict viewer demographics at the impression level. Training a random forest classifier is not the only approach for demographic modeling, and other signals can be considered to improve model performance. Such efforts are not described here since our current goal is simply to demonstrate the utility of the CTV panel in fitting a demographic model.

[12] The quality label of the demographic label of a cookie comes from the output of another model. The model assigns a quantitative score to the demographic label of each cookie to represent the level of confidence we have in the demographics.

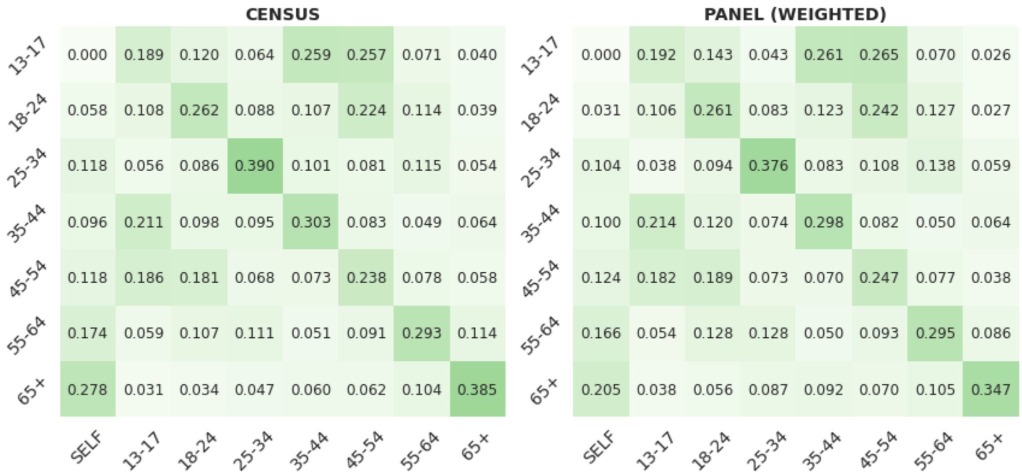


Figure 3: Heatmaps of the census target matrix B (left) and the weighted panel matrix (right), after applying the weights generated using a regularization parameter of $\lambda = 20$. Note that each row represents a probability distribution across the demographic groups (i.e., each row has non-negative entries and sums to one). In each row i, the "SELF" entry gives the probability of a single-inhabitant household across all households with at least one inhabitant of demo i. Notice that for $\lambda = 20$ the weighted panel matrix aligns closely, but not exactly, with the census matrix on the left.

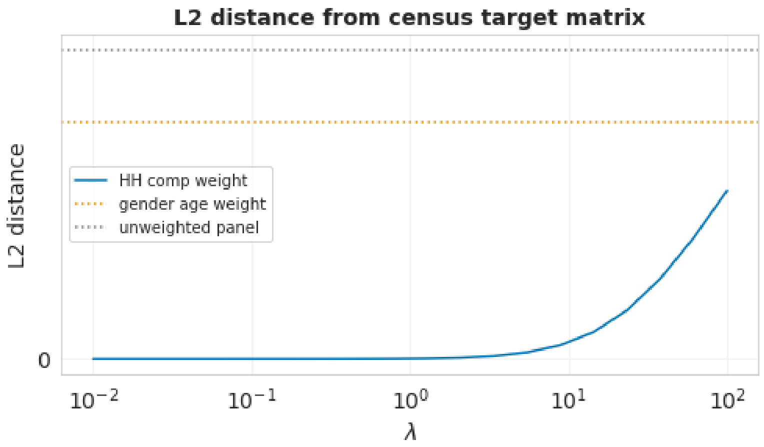


Figure 4: Performance of the weights for different values of the regularization parameter λ. The blue curve shows the Frobenius distance between the census target matrix and the weighted panel matrix as λ varies. The gray dotted line shows the Frobenius distance between the census target matrix and the unweighted panel matrix. Additionally, to illustrate the added value of this bespoke weighting procedure, we show the Frobenius distance betweeen the census target matrix and the weighted panel matrix after we apply the panelist-level weights described in Section 2.3 (orange dotted line).

Separate demographic models are needed for signed-in and signed-out cookie spaces as the quality of the signals and user behavior can differ across these two spaces. Here we describe the model and results for the signed-in space.

CTV impressions can be classified into three categories: no viewer present, one viewer present, or multiple viewers present. Demographic modeling applies to the latter two cases. One approach is to apply the co-viewing model described in Section 4.2 to predict the number of viewers and then fit separate demographic models for the single-viewer and multiple-viewer cases. Here, we take an alternative approach by first defining a primary viewer and potential co-viewers for an ad impression. One random forest classifier is fit to predict the probability distribution across demographic groups for the primary viewer and a second random forest classifier is used to predict the corresponding probability distribution for any remaining co-viewers. In this alternative approach, the first random forest classifier for the primary viewer is shared by all impressions that have at least one viewer present (i.e., the latter two cases listed above), while the second random forest classifier is only applicable to the third case in which at least two people are present.

In the single-viewer case, the sole person watching the TV is naturally deemed the primary viewer. For the multiple-viewer case, we define the primary viewer as the panelist whose age and gender most closely match the gender and age of the demographic label of the cookie. The priority queue for this demographic matching is the following: both gender and age match (match score = 2) > only gender or only age matches (match score = 1) > neither is a match (match score 0). Ties among the highest scoring panelists are broken at random to select one panelist as the primary viewer. Any remaining panelists who are not selected as the primary viewer are treated as co-viewers.

The random forest classifier for the primary viewer case consists of 1,000 trees that can each grow to a maximum depth of 50. Each split in the tree is determined based on a random subset of the features. We set the number of features to 36 (this is the square root of the number of features after converting the original categorical features using one-hot encoding) when looking for the best split and use the Gini impurity to measure the quality of a split. The signals are weaker for the co-viewer model, so we rely on more complex trees to extract as much information as possible. The random forest for the co-viewer case consists of 1,000 trees, but each tree can grow to a depth of 100 and each tree split is based on 100 features selected at random. When training both random forest classifiers, we weight each panelist by the weight we computed in Section 4.4.1, so that our training data is more representative of the general CTV population and their co-inhabitant patterns.

4.4.3 Model Application

We train two random forest classifiers using the approach described in Section 4.4.2 using CTV impressions in the signed-in space from about 45,000 campaigns. We apply the two random forest classifiers to 111 larger out-of-sample campaigns to estimate impression-level demographic distributions. These testing campaigns are selected based on the number of unique panelists reached in each campaign to ensure that we have an accurate measure of ground truth campaign-level demographic distributions.

For this illustrative model, the general procedure for estimating the impression-level demographic distribution of a campaign is the following:

1. Collect campaign impressions and associated signals for prediction.
2. Apply the random forest classifier for the primary viewer case to obtain the estimated demographic distribution of the primary viewer for each impression.
3. For each impression, apply the co-viewing rate model to probabilistically determine if a co-viewer is present.
4. If a co-viewer is present, apply the random forest classifier for the co-viewer case to estimate the demographic distribution of the co-viewer.
5. Aggregate the demographic distributions of the primary viewers and co-viewers across all impressions to obtain an estimated demographic distribution for the campaign.

For the analysis described in this section, we use the ground truth number of viewers rather than determine this probabilistically as specified in step 3. This leads to a fair performance evaluation of the random forest classifiers that is not coupled with the performance of the co-viewing rate model.

For each campaign, we compare the estimated demographic distribution with the true demographic distribution as determined from the ground truth of the CTV panel using two metrics: the shuffle distance and Pearson correlation between these two distributions. The shuffle distance between two demographic distributions d_1 and d_2 is the following:

$$\text{shuffle}(d_1, d_2) = \frac{1}{2} \left\| d_1 - d_2 \right\|_1 ,$$

where the i-th entry of demographic distribution vector d represents the probability associated with the i-th demographic group. The shuffle distance provides a measure of the total absolute difference across all categories between two categorical distributions and has been used previously in measuring online audiences [6].

The following individual campaign-level result is helpful for interpreting shuffle distance magnitudes. In Figure 5, the blue bar plot shows probability mass differences between the true demographic distribution of a particular campaign and the overall demographic distribution of all CTV traffic. These two distributions are significantly different from one another and the corresponding shuffle distance between them is 0.64.[13] In contrast, the orange curve shows the probability differences between the estimated campaign demographic distribution using the random forest classifiers described above and the true demographic distribution of the campaign. These distributions are very similar to one another and the shuffle distance between them is 0.1. These results indicate that the random forest models effectively identify the demographic distribution for this demographically-skewed campaign.

At a more aggregate level, among the 111 out-of-sample test campaigns, the median shuffle distance is 0.09, the 95th percentile is 0.14, and the maximum is 0.22 (see Table 2). Additionally, the Pearson correlations are high for all campaigns. The minimum correlation is 0.86, the median correlation is 0.99, and for 95% of campaigns the correlation is above 0.92. The blue histograms in Figure 6 show the distributions of shuffle distances and Pearson correlations among the 111 campaigns.

As an additional evaluation of the random forest classifiers this campaign-level analysis was repeated for the subset of campaigns that have demographic distributions that are more distant from the overall demographic distribution of all CTV traffic. These campaigns have demographic distributions with a shuffle distance of more than 0.3 from the overall demographic distribution of all CTV traffic. 29 campaigns met this criteria with an average shuffle distance of 0.41 and a maximum shuffle distance of 0.64.

Figure 6 shows histograms of shuffle distance and Pearson correlation among the 29 skewed campaigns (red bars) along with histograms of the 111 test campaigns (blue bars). Compared to the 111 testing campaigns, the shuffle distances among the skewed campaigns cover a narrower range of values. Pearson correlations among the skewed campaigns span the whole range of correlations among all test campaigns, but are less concentrated near 1. More quantitatively, the 29 skewed campaigns show a slight increase in the average shuffle distance and a slight decrease in the average correlation (see Table 3). The median shuffle distance is 0.11 and the median correlation is 0.96. The maximum shuffle distance is 0.16 and the minimum correlation is 0.86. Overall, the demographic models are robust in predicting campaign demographic distributions for a variety of campaigns, including those with demographic compositions quite different from the global demographic distribution for all CTV traffic.

We also verify that the demographic models can consistently provide accurate estimates at the demographic group level. For each demographic group, we compare the vector of true campaign demographic probabilities with the vector of estimated campaign demographic probabilities generated by the demographic model. For a fair comparison of correlations and estimation quality among different demographic groups, we combine Age 13-17 and Age 18-24 in each gender group into a single age group, i.e., F13-24 and M13-24, hence each correlation measures the predictive performance over a roughly 10-year age range. The correlation between these two vectors is an indication of model performance for a specific demographic group. In all demographic groups, correlations are above 0.9, indicating that our demographic model consistently estimates demographic weights well across demographic groups. See Table 4 for a list of correlations in each demographic group. Restricting to the more skewed campaigns, there is a slight decrease in correlation from 0.93 to 0.87 in F65+, while for all other demographic groups correlations are still above 0.9 and are very similar to those in Table 4.

5 Summary

CTV devices have higher levels of sharing and co-viewing than other digital devices. This characteristic requires special considerations for reach modeling. One such consideration is the use of CTV panel data as a tool for observing and measuring co-viewing behavior. Although CTV panel data does not provide a practical means for directly measuring the reach of CTV campaigns, it is a key source of information that should be used in conjunction with survey data, ad server logs, publisher-provided self-reported demographic data, and census data.

CTV panel data is especially useful for measuring the aggregate co-viewing rate and for fitting demo-

[13] Such differences in demographic distributions can be caused by variations in campaign design, such as the specification of demographic or publisher targeting.

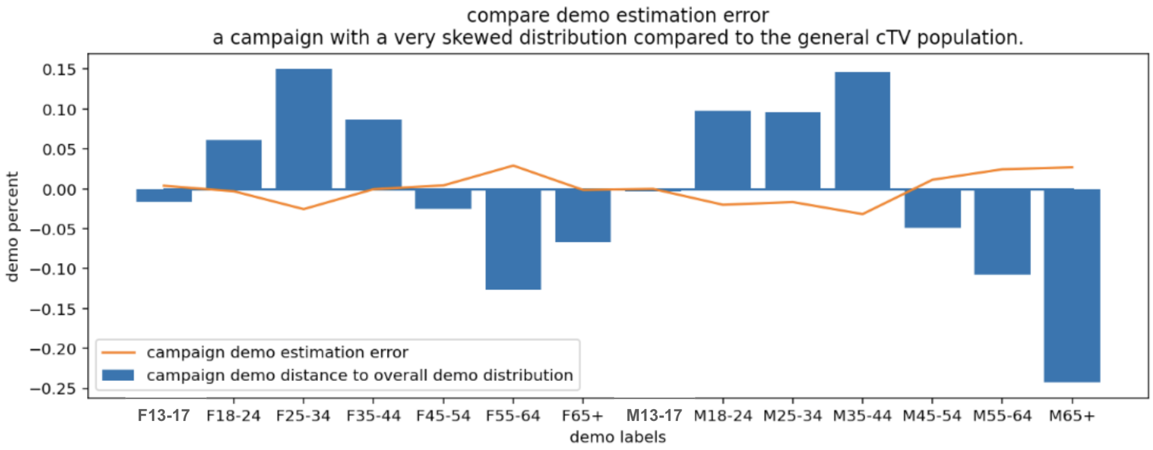


Figure 5: An illustrative example of the predictive performance of the random forest classifiers on a campaign with a highly skewed demographic distribution. The blue bar plot shows the probability differences between the true demographic distribution of the campaign and the overall average demographic distribution for CTV traffic. The shuffle distance between these two distributions is 0.64. The orange curve shows the weight differences between the estimated campaign demographic distribution using the random forest classifiers and the true demographic distribution of the campaign. The shuffle distance between these two distributions is 0.1.

Metric	Minimum	Median	cutoff for 95% of campaigns	Max
Shuffle Distance	0.02	0.09	0.14	0.22
Correlation	0.86	0.99	0.92	0.999

Table 2: Summary statistics of shuffle distances and Pearson correlations among the 111 testing campaigns. For shuffle distance, the "cutoff for 95% of campaigns" is the 95th percentile of shuffle distances. For Pearson correlation, the "cutoff for 95% of campaigns" is the correlation at the 5th percentile since it is favorable to have a higher correlation.

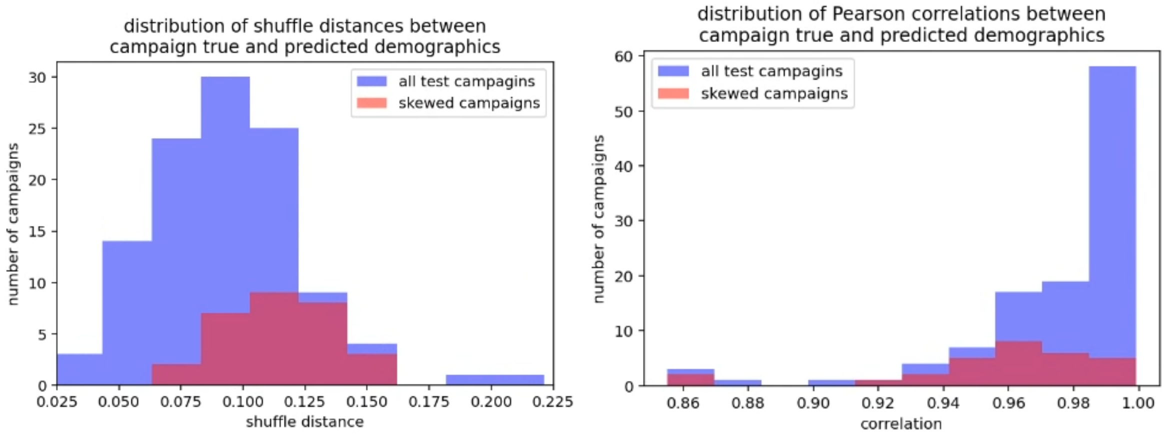


Figure 6: Histograms of shuffle distances (left) and correlations (right) between campaign true demographic distribution and campaign estimated demographic distribution among testing campaigns. Histograms in blue include all 111 out-of-sample testing campaigns, while histograms in red are a subset of 29 campaigns with demographic distributions that are more distant from the global demographic distribution of all CTV traffic.

Metric	Minimum	Median	cutoff for 95% of campaigns	Max
Shuffle Distance	0.07	0.11	0.15	0.16
Correlation	0.86	0.96	0.88	0.995

Table 3: Summary statistics of shuffle distances and Pearson correlations among the 29 campaigns whose demographic distributions are most distant from the overall demographic distribution of the CTV traffic. For shuffle distance, the "cutoff for 95% of campaigns" is the 95th percentile of shuffle distances. For Pearson correlation, the "cutoff for 95% of campaigns" is the correlation at the 5th percentile since it is favorable to have a higher correlation.

Demographic Group	Correlation	Demographic Group	Correlation
F13-24	0.92	M13-24	0.95
F25-34	0.91	M25-34	0.94
F35-44	0.97	M35-44	0.97
F45-54	0.92	M45-54	0.97
F55-64	0.95	M55-64	0.97
F65+	0.93	M65+	0.98

Table 4: Correlation for each demographic group. For each demographic group, we compute the correlation between the vector of campaign true demographic probabilities and the vector of campaign estimated demographic probabilities.

graphic models that account for co-viewing. The aggregate co-viewing rate can be measured accurately and is relatively stable across time. The generation of impression-level co-viewing rates requires the volume and country-level coverage that survey data provides.

The measurement of co-viewing demographics requires the use of panelist weighting that matches household composition with census benchmarks. Without this weighting, the demographic model may not properly account for the opportunity that one demographic group has to co-view with another demographic group. When such considerations are taken into account, CTV panel data can be used to reliably estimate the demographic composition of reach at the campaign level, even for campaigns with demographic compositions that are skewed.

Acknowledgements

We are grateful to our colleagues at Google for many invaluable discussions, comments, and suggestions. In particular, we would like to acknowledge Thomas Vaughan, Eileen Park, Evgeny Skvortsov, Heidi Liao, Eduardo Ruiz, Shaun Lysen, Anand Sengodan, Doug Eastman, Mary Yang, Ankit DP, and Melinda Salaman for their insights and contributions to this work.

References

[1] Shane Barratt, Guillermo Angeris, and Stephen Boyd. Optimal representative sample weighting. *Statistics and Computing*, 31(19), 2021.

[2] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[3] eMarketer. Connected tv (ctv) advertising spending in the united states from 2019 to 2027 (in billion u.s. dollars) [graph]. `https://www.statista.com/statistics/1048897/connected-tv-ad-spend-usa/`, February 2024.

[4] Rachel Fan, Wei Liu, and Jon Vaver. Ctv co-viewing rate estimation using online surveys. Technical report, Google Inc., 2021.

[5] Rachel Harter, Michael P Battaglia, Trent D Buskirk, Don A Dillman, Ned English, Mansour Fahimi, Martin R Frankel, Timothy Kennel, Joseph P McMichael, Cameron Brook McPhee, et al. Address-based sampling. Technical report, AAPOR, 2016.

[6] Jim Koehler, Evgeny Skvortsov, Sheng Ma, and Song Liu. Measuring cross-device online audiences. Technical report, Google Inc., 2016.

[7] Jim Koehler, Evgeny Skvortsov, and Wiesner Vos. A method for measuring online audiences. Technical report, Google Inc., 2013.

[8] Evgeny Skvortsov and Jim Koehler. Virtual people: Actionable reach modeling. Technical report, Google Inc., 2019.

[9] IAB (U.S.). Sources for connected tv (ctv) and over-the-top (ott) media

budgets according to marketers in the united states in 2023 [graph]. https://www.statista.com/statistics/1227143/shift-media-budget-ott-ctv-usa/, February 2024.

[10] Xiaojing Wang, Jingang Miao, and Yunting Sun. A python library for empirical calibration. *arXiv preprint arXiv:1906.11920v2*, 2019.

Appendix A

Direct Panel Measurement

In this section, we describe a simple model of the number of people reached by an ad campaign. This simplified model illustrates the relationship between panel size, campaign size, and measurement error. Let n_{pop} denote the size of the population, n_{panel} the size of the panel, and n_{camp} the size of the campaign (more specifically, the expected campaign reach).

Assume that each panelist is reached by the campaign independently with probability $p = \frac{n_{camp}}{n_{pop}}$; that is, each panelist $i \in \{1, ..., n_{panel}\}$ has a random variable $X_i \sim \text{Ber}(p)$ representing whether or not they were reached by the campaign. Consequently, notice that

$$X = \sum_{i=1}^{n_{panel}} X_i \sim \text{Bin}(n_{panel}, p)$$

is a random variable quantifying the total number of panelists reached by the campaign. The rate (binomial proportion) at which panelists are reached is given by $\frac{X}{n_{panel}}$, which can be used to project the panel reach to the population. The population reach is given as $X_{pop} = \frac{n_{pop}}{n_{panel}} X$. Now we need to calculate the variance of this random variable. Notice that

$$\begin{aligned}\text{Var}(X_{pop}) &= \text{Var}\left(\frac{n_{pop}}{n_{panel}} X\right) \\ &= \frac{n_{pop}^2}{n_{panel}^2} \text{Var}(X) \\ &= \frac{n_{pop}^2}{n_{panel}^2} n_{panel}\, p\,(1-p) \\ &= \frac{n_{camp}(n_{pop} - n_{camp})}{n_{panel}}.\end{aligned}$$

Consequently, the standard error is

$$SE = \sqrt{\text{Var}(X_{pop})} = \sqrt{\frac{n_{camp}(n_{pop} - n_{camp})}{n_{panel}}}.$$

Finally, the $(1-\alpha) \times 100\%$ confidence interval width is given by $2z_{1-\alpha/2}SE$, where z_q denotes the q-th quantile of the standard normal. For ease of comparison, we investigate the error on a relative scale, so we calculate the confidence interval width divided by the size of the campaign. Figure B.1a shows that even a large panel comprised of 100,000 panelists fails to provide viable measurement for even the largest digital campaigns.

Appendix B

Hybrid Measurement

In this section, we present an illustrative model for estimating the number of people reached by an ad campaign. The distinction between this model and the one described in Appendix A is that this model leverages logs data in addition to panel data to fit a reach curve. We demonstrate that this hybrid measurement approach, unlike panel-based direct measurement, provides reliable reach estimates for the full range of digital campaign sizes.

Again, let n_{pop} denote the size of the population, n_{panel} the size of the panel, and n_{camp} the size of the campaign (i.e., the expected campaign reach). We assume the data comes from an *Exponential Bow* model [6], which is characterized by the following cookie-to-user function:

$$r(c; \kappa) = \frac{n_{pop}\kappa c}{n_{pop} + \kappa c},$$

where κ represents the slope at the origin and should be close to 1. Additionally, note that this implies the following inverse (i.e., user-to-cookie) function:

$$c(r; \kappa) = \frac{n_{pop} r}{\kappa(n_{pop} - r)}.$$

Under this model, the probability that the j-th person in the population has at least one cookie, given a campaign with c cookies, is given by

$$P(c_j > 0 | c) = \frac{1}{1 + n_{pop}/\kappa c}.$$

Notice that the expected number of people reached by the campaign can be calculated by summing these probabilites across all individuals in the population:

$$\sum_{j=1}^{n_{pop}} \frac{1}{1 + n_{pop}/\kappa c} = \frac{n_{pop}}{1 + n_{pop}/\kappa c} = \frac{n_{pop}\kappa c}{n_{pop} + \kappa c},$$

which is exactly $r(c; \kappa)$, as desired. Finally, let $c^* = c(n_{camp}; \kappa)$ denote the number of cookies associated with a campaign of size n_{camp}. We fit a model and use it to predict the number of people reached by a campaign with c^* cookies exposed. The simulation procedure is the following:

```
Algorithm 2: Hybrid model
  // Generate campaign data for
     training
  for $i \in \{1, ..., n_{train}\}$ do
      Sample $c_i \sim \text{Exp}(\lambda)$.
      Sample $r_i \sim \text{Bin}(n_{pop}, \frac{1}{1+n_{pop}/\kappa c_i})$.
  end
  // Fit reach curve multiple times
  for $b \in \{1, ..., B\}$ do
      // Find panel coverage for
         training data
      for $i \in \{1, ..., n_{train}\}$ do
          Sample $\tilde{c}_i^b \sim \text{Bin}(c_i, \frac{n_{panel}}{n_{pop}})$.
          Sample $\tilde{r}_i^b \sim \text{Bin}(r_i, \frac{n_{panel}}{n_{pop}})$.
      end
      Fit a value of $\hat{\kappa}^b$ to the panel data, i.e.,
        $\hat{\kappa}^b = \text{argmin}_\kappa \sum_{i=1}^{n_{train}} \left(\tilde{r}_i^b - \frac{n_{panel}\kappa\tilde{c}_i^b}{n_{panel}+\kappa\tilde{c}_i^b}\right)^2$
      Use the fitted model to predict the
        number of people reached $\hat{r}^b = r(c^*; \hat{\kappa}^b)$.
  end
  Calculate the empirical standard deviation of
    the $B$ reach values $\{\hat{r}^1, ..., \hat{r}^B\}$.
```

The panel size n_{panel} will influence the panel coverage for each campaign and will therefore influence the κ value we fit to the data. With a larger panel, we expect a more reliable estimate of κ and therefore a more reliable reach estimate.[14] The $(1-\alpha) \times 100\%$ confidence interval width is given by $2z_{1-\alpha/2}SE$, where z_q denotes the q-th quantile of the standard normal. We divide the confidence interval width by the size of the campaign to obtain the relative error. Figure B.1b shows that even a small panel comprised of 1,000 panelists provides reach measurement within the 20% error tolerance. Additionally, for each panel size, notice that the reach estimation error is quite stable across the range of campaign sizes (c.f., the larger error for smaller campaigns in the direct panel-based measurement setting).

Appendix C

Derivation of Co-viewing Rate Variance

We will use the delta method to obtain an approximation of the variance of the co-viewing rate m. For simplicity, let $\bar{x} = \sum_i x_i w_i$ and $\bar{t} = \sum_i t_i w_i$. We can express the co-viewing rate m as

$$m = f\left(\begin{bmatrix}\bar{x} \\ \bar{t}\end{bmatrix}\right) = \frac{\bar{x}}{\bar{t}}.$$

Additionally, let $m_x = \mathbb{E}[\bar{x}] = \mu_x \sum_i w_i$ and $m_t = \mathbb{E}[\bar{t}] = \mu_t \sum_i w_i$ denote the expected value of $\bar{x}$ and $\bar{t}$, respectively. The gradient of the function f evaluated at the mean is then given by

$$v = \nabla f\left(\begin{bmatrix}m_x \\ m_t\end{bmatrix}\right) = \begin{bmatrix}\frac{1}{m_x} \\ -\frac{m_x}{m_t^2}\end{bmatrix}.$$

Next, we need the covariance matrix of $\begin{bmatrix}\bar{x} \\ \bar{t}\end{bmatrix}$, which we will denote by Σ. In particular, we need to calculate $\text{Var}(\bar{x})$, $\text{Var}(\bar{t})$, and $\text{Cov}(\bar{x}, \bar{t})$. Notice that

$$\text{Var}(\bar{x}) = \text{Var}\left(\sum_i x_i w_i\right) = \sigma_x^2 \sum_i w_i^2.$$

Similarly, we have that

$$\text{Var}(\bar{t}) = \text{Var}\left(\sum_i t_i w_i\right) = \sigma_t^2 \sum_i w_i^2.$$

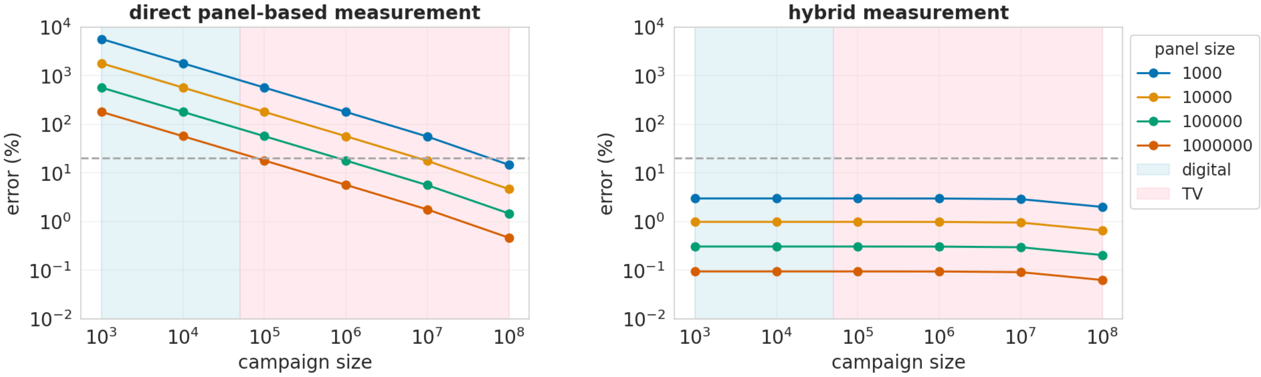
Now for the covariance term, assuming dependence between co-viewer count and watchtime within a household but independence across households, notice that

$$\begin{aligned}\text{Cov}(\bar{x}, \bar{t}) &= \text{Cov}\left(\sum_i x_i w_i, \sum_j t_j w_j\right) \\ &= \sum_i \sum_j w_i w_j \text{Cov}(x_i, t_j) \\ &= \sum_i w_i^2 \text{Cov}(x_i, t_i) \\ &= \rho\sigma_x\sigma_t \sum_i w_i^2.\end{aligned}$$

We now have everything we need. By the delta method, we have that $\text{Var}(m) \approx v^\top \Sigma v$;

$$\begin{aligned}\text{Var}(m) &\approx \begin{bmatrix}\frac{1}{m_t} \\ -\frac{m_x}{m_t^2}\end{bmatrix}^\top \begin{bmatrix}\sigma_x^2 \sum_i w_i^2 & \rho\sigma_x\sigma_t \sum_i w_i^2 \\ \rho\sigma_x\sigma_t \sum_i w_i^2 & \sigma_t^2 \sum_i w_i^2\end{bmatrix} \begin{bmatrix}\frac{1}{m_t} \\ -\frac{m_x}{m_t^2}\end{bmatrix} \\ &= \frac{\sigma_x^2 \sum_i w_i^2}{\mu_t^2 (\sum_i w_i)^2} - \frac{2\mu_x\rho\sigma_x\sigma_t \sum_i w_i^2}{\mu_t^3 (\sum_i w_i)^2} + \frac{\mu_x^2\sigma_t^2 \sum_i w_i^2}{\mu_t^4 (\sum_i w_i)^2} \\ &= \frac{1}{n_{eff}}\left(\frac{\sigma_x^2}{\mu_t^2} - \frac{2\mu_x\rho\sigma_x\sigma_t}{\mu_t^3} + \frac{\mu_x^2\sigma_t^2}{\mu_t^4}\right).\end{aligned}$$

[14] Certainly, a model that produces a biased estimate of κ can still achieve small measurement error across trials. However, we confirm that our fitted models fit the data well and closely recover the true κ value.



(a) Reach estimation error via panel measurement.

(b) Reach estimation error via hybrid measurement.

Figure B.1: Relationship between campaign size, panel size, and reach measurement error. The error is defined as the width of the 90% (Gaussian) confidence interval of the estimated population reach relative to the campaign size (×100%). The gray dashed line represents a maximum admissible measurement error of 20%. The shaded areas in blue and pink show typical campaign sizes for digital and linear TV ads, respectively. The population size is assumed to be 300 million people. (left) Notice that for even the largest digital campaigns, a large panel fails to provide viable measurement of campaign reach via direct panel-based measurement. (right) However, we see that even a small panel comprised of 1,000 panelists provides reach estimates below the 20% error tolerance via the hybrid measurement approach. The simulation parameters for this set of results are the following: $\kappa = 1$, $n_{train} = 500$, and $1/\lambda = 20,000,000$.