Conditional Independence Approximation to Cross-media Incremental Reach

Xinyang Shen, Yunwen Yang, Jim Koehler, Lu Zhang

Google Inc.

Feb 22, 2019

Abstract

An independence assumption has been used as a solution to calculate cross-media ads reach when cross-media panel data are unavailable. In this paper, we introduce *restricted conditional independence assumption* (RCIA) and apply it to calculate the digital video incremental reach on top of TV. By conducting large-scale simulation studies based on the cross-media universes in a single source panel, we show that RCIA based incremental reach is close to the standard calculation based on the cross-media universes. Moreover, substantial variance reduction is achieved from RCIA with single-media panel data compared to the standard calculation based on the cross-media universes. We use distance correlation to measure the strength of dependence between ad impressions of TV campaigns and digital video. The magnutudes of distance correlation between these two types of ad impressions are small, and only $\sim 2\%$ TV campaigns show statistically significant dependence. As a summary, because of its high accuracy, low variance, simplicity and low cost, we recommend that the RCIA approximation as a preferred solution than the standard cross-media universe calculation.

1 Introduction

With the growth of the Internet in the last 20 years, online video advertising has become an effective media platforms to deliver impactful reach to targeted audiences. Research from Nielsen (2017) in Q2 2017 found that 36% of the US population watch video on computers and 69% watch video on smartphones. Using multiple media together is believed to create synergy in building brand equity, see Lim et al. (2015). Another major reason of using multiple media channels is to increase both reach and frequency to targeted audiences. As a result, incremental reach of digital video advertising on top of TV reach is an important metric that marketers use widely for both media planning and post-campaign reporting purposes, see examples in Jin et al. (2013).

The preferred instrument to measure online incremental reach is through a cross-media universe in a single source panel (SSP), which is a group of panelists that are probabilistically recruited

to represent certain populations of interest. Google has contracted several SSP to track TV and YouTube monetizable ad impressions for participating panelists. The online and television ads viewing data of those panelists are captured along with their demographic information such as age, gender, income group, etc. The TV ad impressions are historical ads that panelists have watched. All TV ad impressions are labeled with specific TV campaign information such as company, brand, specific campaign name, and channel information. The YouTube monetizable ad impressions include both actual YouTube ads shown to panelists and potential ad spots where ads are allowed to serve to panelists if purchased by advertisers. With cross-media universes in an SSP, we can calculate the reach from one or multiple devices by directly counting the number of panelists. One big problem of such SSP is that it can be expensive to recruit and maintain. In reality, such SSP on TV and digital platforms are unavailable in many countries. In addition, the size of such cross-media universe in an SSP is typically limited especially within some gender and age groups. On the other hand, single-media panels can be more easily recruited and its size is generally several times bigger than a cross-media one. Chen et al. (2014) developed a data enrichment approach that combines small cross-media panels and large single-media panels for incremental reach estimation. Another problem of using cross-media universes in an SSP to calculate incremental reach is that its calculation does not include the incremental reach from the internet only population, and thus tends to under estimate the incremental reach.

Due to population size and budget concerns, many countries don't have a cross-media panel available. For those countries, it is a generally accepted practice to assume an independence assumption to calculate reach across different media channels. The independence assumption, also known as Sainsburry formula in Caffyn and Sagovsky (1963), assumes there is no particular correlation between exposures to one media channel or another. Assume that we have two channels with reaches represented as proportions of population: r_1 and r_2, then the combined reach r_c is defined as:

$$r_c = 1 - (1 - r_1)(1 - r_2) = r_1 + r_2 - r_1 r_2, \tag{1}$$

where $r_1 r_2$ calculates the joint reach, the percentage of the population reached by both channels, under independence. A modified version of Sainsburry formula varies from the above simple independence assumption by adding a small adjustment ρ to account for correlation between exposures to different media channels:

$$r_c = r_1 + r_2 - \rho r_1 r_2, \tag{2}$$

Following Chen et al. (2014), we adopt the same notion of *conditional independence assumption*, or CIA. That is, TV and digital video ad impressions are conditionally independent given the subject's demographics. An obvious limitation of the above independence assumption is that it does not distinguish people with only subset of devices from people with all devices. Therefore, we propose to further restrict the CIA assumption to the people who own both TV and Internet devices (computer, mobile, tablet). In other words, we exclude people who have either TV only or Internet device only from CIA. We call this as *restricted conditional independence assumption*, or

RCIA. In this paper, we investigate the performance of the independence based incremental reach of digital video ads by comparing it to the standard incremental reach calculation based on the cross-media universes in an SSP. Furthermore, we use distance correlation by Szekely et al. (2007) and Szekely and Rizzo (2009) and its associated statistical testing procedure proposed in Szekely and Rizzo (2013) to validate the independence assumption in RCIA.

Note that under-reporting is a common and serious problem in collecting YouTube ad impressions through panelists, see Goerg et al. (2015). We apply the probabilistic model based imputation methodology in Goerg et al. (2015) to impute back YouTube ad impressions for all panelists. The imputation model imputes the unobserved YouTube ad impressions based on observed YouTube ad impressions only, and consequently, imputed YouTube ad impressions show weaker dependence with TV ad impressions compared to observed TV ad impressions. Because all panels suffer from under-reporting and need to be imputed for product use, our comparison of RCIA and the standard cross-media universe based calculation uses imputed YouTube ad impressions. Instead, the statistical correlation measurement uses observed YouTube ad impressions.

The remainder of this paper is organized as follows. In Section 2, we describe the incremental reach calculation using multiple cross-media universes in an SSP with RCIA. In Section 3 we construct simulation studies that cover a large range of potential scenarios based on an SSP data to show that the difference between RCIA based and standard incremental reach calculation is small, ignorable in practice. RCIA also exhibits an important benefit in variance reduction, which is consistent with the observation in Chen et al. (2014). Section 4 uses distance correlation to measure dependence strength. Statistical testing on conditional independence of TV campaigns and digital video ad impressions show no evidence to reject the independence assumption.

2 Restricted Conditional Independence Model

An SSP usually consists of multiple panel universes: cross-media panel universes and single-media panel universes. Each panel universe is constructed to be a probabilistic sampling of a specific device (single or multiple) population with panelists' weights properly calibrated to the corresponding population. The weight of a panelist is interpreted as the number of people the panelist represents in the corresponding population universe, and thus, one panelist could appear in multiple universes and has different weights in different universes. In this work, we use the SSP provided by INTAGE[1] called i-SSP (INTAGE Single Source Panel). It provides the following panel universes:

- TV universe
- Desktop universe
- Mobile universe

[1]Intage web site: https://www.intage.co.jp

- Desktop_Mobile (intersection) universe
- Internet_TV (intersection) universe
- Internet universe
- Desktop_Mobile_TV (intersection) universe

By joining Internet_TV universe with Desktop universe and Mobile universe, respectively, we can construct the Desktop_TV universe and Mobile_TV universe. The following definitions and notions are introduced to define the RCIA based incremental reach calculation. Let N denote the population size of a particular universe, $R()$ denote the number of reached people, and $E()$ denote the incremental reach of digital video as the absolute people count on top of TV reach. By default, these numbers are calculated for a specific demographic only. For simplicity, we omit the notion of demographic in the formulas below. In addition, we use D for desktop impressions, M for mobile impression, and T for TV impressions. We use t, d, m in subscript of $R()$ and $E()$ to represent the metric defined in TV, Desktop, and Mobile universes respectively. For example, $R_{dmt}(D+M \geq k)$ means the $k+$ reach of digital video ads in Desktop_Mobile_TV universe, which represents the population that owns all three devices; $E(D \geq k)$ means the $k+$ incremental reach of desktop video ads; $E(M \geq k)$ means the $k+$ incremental reach of mobile video ads. Define I as the indicator function where $I_{j=0} = 1$ when $j = 0$, and $I_{j=0} = 0$ for $j > 0$.

Given the RCIA, the joint reach of desktop and TV devices is:

$$\begin{aligned} R(D=i,T=j) &= R_{dt}(D=i,T=j) + I_{j=0}R_{d\backslash t}(D=i) + I_{i=0}R_{t\backslash d}(T=j) \\ &= \frac{R_{dt}(D=i)R_{dt}(T=j)}{N_{dt}} + I_{j=0}R_{d\backslash t}(D=i) + I_{i=0}R_{t\backslash d}(T=j) \\ &= \frac{R_{dt}(D=i)R_{dt}(T=j)}{N_{dt}} + I_{j=0}[R(D=i) - R_{dt}(D=i)] \\ &\quad + I_{i=0}[R(T=j) - R_{dt}(T=j)], \end{aligned} \tag{3}$$

where $d\backslash t$ denotes the desktop population that does not own TV device and $t\backslash d$ means the TV population that does not own desktop device. We assume that all panel weightings are properly calibrated across demographic groups of interest such that they match population benchmarks exactly. Similarly, the joint reach of mobile and TV devices with the RCIA is:

$$\begin{aligned} R(M=i,T=j) = \frac{R_{mt}(M=i)R_{mt}(T=j)}{N_{mt}} &+ I_{j=0}[R(M=i) - R_{mt}(M=i)] \\ &+ I_{i=0}[R(T=j) - R_{mt}(T=j)]. \end{aligned}$$

Define vt as the universe of population that owns TV device and at least one of desktop and mobile

devices. The joint reach of digital and TV devices is:

$$R(D+M=i,T=j)=\frac{R_{vt}(D+M=i)R_{vt}(T=j)}{N_{vt}}+I_{j=0}[R(D+M=i)-R_{vt}(D+M=i)]$$
$$+I_{i=0}[R(T=j)-R_{vt}(T=j)].$$

If an SSP contains panel universe for vt, then we would use it directly to calculate the formula above. However, the partial device participation issue in the recruitment of cross-media universe makes it hard to construct a high-quality panel for this specific audience. Many panelists only participate in the study for one tracking device even though they may own multiple devices and they do not provide information on what devices they own. An alternative is to calculate the $R_{vt}()$ using panel universes that represent the intersections of various device populations.

$$R_{vt}(T=j)=R_{dt}(T=j)+R_{mt}(T=j)-R_{dmt}(T=j)$$
$$R_{vt}(D+M=i)=R_{dt}(D=i)-R_{dmt}(D=i)+R_{mt}(M=i)-R_{dmt}(M=i)$$
$$+R_{dmt}(D+M=i)$$

With the RCIA, the incremental reach of desktop, mobile, and both devices can be defined as follows:

$$\begin{aligned}E(D\geq k)&=R(D+T\geq k,T<k)\\&=R_{dt}(D+T\geq k,T<k)+R_{d/t}(D\geq k)\\&=\frac{\sum_{i=1}^{k}R_{dt}(D\geq i)R_{dt}(T=k-i)}{N_{dt}}+R(D\geq k)-R_{dt}(D\geq k),\end{aligned}\tag{4}$$

$$E(M\geq k)=\frac{\sum_{i=1}^{k}R_{mt}(M\geq i)R_{mt}(T=k-i)}{N_{mt}}+R(M\geq k)-R_{mt}(M\geq k),\tag{5}$$

$$\begin{aligned}E(D+M\geq k)&=\frac{\sum_{i=1}^{k}R_{vt}(D+M\geq i)R_{vt}(T=k-i)}{N_{vt}}\\&+R(D+M\geq k)-R_{vt}(D+M\geq k).\end{aligned}\tag{6}$$

The incremental reach calculation as specified in (4), (5) and (6) still relies on the cross-media universes, such as Destkop_TV universe. In many real applications, we may only have single-media panels and thus, the incremental reach calculation as specified in (4), (5) and (6) is not applicable. We need to further make the invariant assumption, i.e., that the percentage of reach from ad impressions through a device is the same between the cross-media population and the single-media population owning that device. With both RCIA and invariant assumption, we have the following

joint/incremental reach calculations which rely on the single-media panels only:

$$\begin{aligned}
R(D=i,T=j) &= \frac{N_{dt}}{N_d N_t} R(D=i)R(T=j) + I_{j=0}(1-\frac{N_{dt}}{N_d})R(D=i) \\
&\quad + I_{i=0}(1-\frac{N_{dt}}{N_t})R(T=j), \\
R(M=i,T=j) &= \frac{N_{mt}}{N_m N_t} R(M=i)R(T=j) + I_{j=0}(1-\frac{N_{mt}}{N_m})R(M=i) \\
&\quad + I_{i=0}(1-\frac{N_{mt}}{N_t})R(T=j), \\
R(D+M=i,T=j) &= \frac{N_{vt}}{N_v N_t} R(D+M=i)R(T=j) + I_{j=0}(1-\frac{N_{vt}}{N_v})R(D+M=i) \\
&\quad + I_{i=0}(1-\frac{N_{vt}}{N_t})R(T=j),
\end{aligned}$$

$$E(D \geq k) = \frac{N_{dt}}{N_d N_t} \sum_{i=1}^{k} R(D \geq i)R(T=k-i) + (1-\frac{N_{dt}}{N_d})R(D \geq k), \tag{7}$$

$$E(M \geq k) = \frac{N_{mt}}{N_m N_t} \sum_{i=1}^{k} R(M \geq i)R(T=k-i) + (1-\frac{N_{mt}}{N_m})R(M \geq k), \tag{8}$$

$$E(D+M \geq k) = \frac{N_{vt}}{N_v N_t} \sum_{i=1}^{k} R(D+M \geq i)R(T=k-i) + (1-\frac{N_{vt}}{N_v})R(D+M \geq k). \tag{9}$$

Note that the invariant assumption is much weaker compared to the independence assumption. An intuitive explanation of the invariant assumption with a given device is that the histogram of device inventory impressions per person is the same between the population with this device only and the rest of population who has this device and some other devices. Moreover, a single-media panel is usually much bigger compared to a cross-media one. Estimating the device reach in a cross-media universe by its counter part in a single-media panel would introduce some bias but meanwhile could gain a big variance reduction, which is demonstrated in the comparision studies in Section 3.3.

3 Comparison Studies

We compare the RCIA based incremental reach formulas introduced in Section 2 with the standard calculation based on a cross-media universe in an SSP using two aspects: bias and variance reduction. When evaluating the deviation of RCIA based calculation from the standard one based on a cross-media universe, we use the RCIA only solutions, i.e., equations (4) to (6). For variance reduction evaluation, we use the combination of RCIA and the invariant assumption, i.e., equations (7) to (9). Although RCIA plus the invariant assumption is more applicable to evaluate real situations where cross-media universe data are not available, it would be hard to separate out the deviation introduced by the invariant assumption from the one introduced by the independence

assumption. Due to limited sample size and recruitment difficulty of cross-media universe, single device reach from cross-media universe may not be suitable to be used as the ground-truth for comparison against single-media panels.

The comparison study is based on INTAGE[2] Japan SSP with data collected between 5/1/2017 to 5/31/2017 on YouTube and TV ad impressions:

- INTAGE[3] Japan SSP panel covers 18+ populations in Tokyo, Osaka and Nagoya, i.e., three Japan TV regions. It has $\sim 5,000$ active TV panelists, $\sim 5,000$ mobile panelists, $\sim 18,000$ desktop panelists, $\sim 3,000$ TV + desktop panelists, $\sim 2,600$ desktop + mobile panelists, and $\sim 1,000$ TV + desktop + mobile panelists,

We also conduct the comparison on other panels and obtain similar findings. Because of license issues, we only report the results using Japan INTAGE panels.

3.1 Simulation

Simulated TV campaigns are typically used in media planning stage where we optimize budget allocation with a new TV campaign. Historical TV campaigns are used when we want to optimize cross-media reach based on an existing TV campaign, and when reporting on historical campaigns.

We simulate ad campaigns with combinations of six different factors:

1. YouTube ad format: homepage/masthead[4] and watchpage[5]. See Jin et al. (2013) for details.
2. k+ frequency: $k = 1, \ldots, 10$,
3. 25 demographic buckets:
   - No demographic targeting, target everyone
   - 24 demographic buckets from the combination of gender and age groups:
     - ⋆ Gender: female, male, and both.
     - ⋆ Age groups: $18-24$, $18-44$, $25-34$, $25-54$, $35-44$, $45-54$, 55+, and 18+.
4. TV campaign type: historical TV campaigns and simulated TV campaigns.
5. Purchased TV ad impressions:
   - Historical TV campaign: we picked 100 historical TV campaigns uniformly across all historical TV campaigns that reached at least 100 TV panelists overall.

[2]Intage web site: https://www.intage.jp
[3]Intage web site: https://www.intage.jp
[4]Display ad on the YouTube homepage, purchased on a cost-per day basis
[5]In-display ads alongside videos watched on YouTube

- Simulated TV campaign: Use 1% of the total TV inventory impressions and generate the 15 evenly distributed quantile points as the purchased TV impressions, including 0 TV impressions.

6. Purchased desktop and mobile impressions as two separate input types. The selection of these two types of impressions also depends on the TV campaign type:

   - Historical TV campaign:

     The allocated desktop and mobile watch page impressions are chosen as the same proportion of total desktop and mobile inventory impressions, 25 proportion points in total: $0, 0.4^{15}, \ldots, 0.4^{6}, 0.01, 0.02, \ldots, 0.1,\ 0.28, 0.46, 0.64, 0.82$.

   - Simulated TV campaign:

     ⋆ Watchpage:

       We use a Cartesian product of these 21 proportions on desktop and mobile separately to create 21^2 combinations of simulated desktop and mobile purchased impressions: $0, 0.4^{15}, \ldots, 0.4^{6}, 0.01, 0.02, \ldots, 0.1$

     ⋆ Homepage:

       We use a Cartesian product of these 15 proportions on desktop and mobile separately to create 15^2 combinations of simulated desktop and mobile purchased impressions: 0, 1, ..., 14 days of averaged daily homepage inventory impressions.

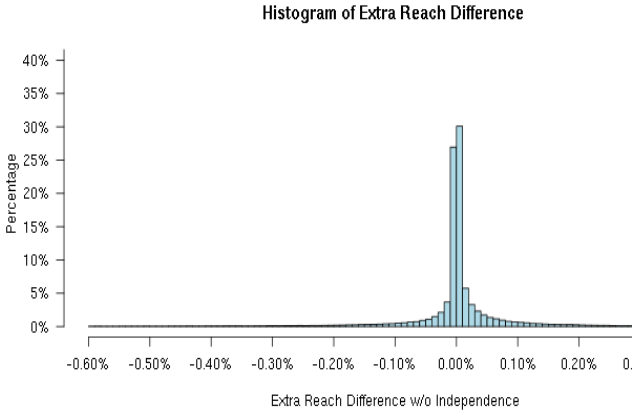
The combination of the above input parameters lead to a huge number of simulation scenarios. When we use simulated TV campaign type, the simulation study has $10 \times 25 \times 15 \times 15 \times 15 = 0.9M$ scenarios for homepage and another $10 \times 25 \times 21 \times 21 \times 15 = 1.7M$ scenarios for watchpage.

3.2 Comparison Results

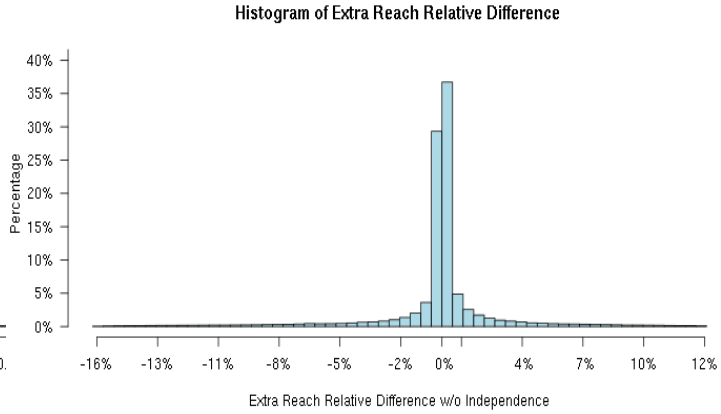
We calculate both absolute difference and relative difference of incremental reaches between the RCIA approximation and the standard calculation based on a cross-media universe in an SSP. Let E_{ind} represent incremental reach with RCIA and E represent the standard cross-media based incremental reach. The difference is simply $E_{ind} - E$ and the relative difference is $1 - \frac{E}{E_{ind}}$.

Figures 1 and 2 are the histograms of differences and relative differences with simulated TV campaigns and historical TV campaigns, respectively, based on Japan INTAGE panel data. Each chart is truncated to include only 95% of all campaigns in simulation. We can see that the differences of most simulation scenarios are small enough to be ignorable in practice and also symmetric around zero. This shows that the RCIA does not introduce systematic bias. Compared to simple differences, the relative differences have larger values but are still concentrate in a neighborhood of zero. In fact, most large relative differences are caused by its denominator, YouTube incremental reach being small, as illustrated in Figure 3, which shows the 95-th quantile of the relative differences broken down by incremental reach bins. For example, for the scenarios with incremental reach less

than 2%, the 95-th quantile of relative differences are 50%, i.e., 5% of scenarios have relative differences greater than 50%. For scenarios with incremental reach larger than 10%, the 95-th quantile drops below 10%, i.e., less than 5% of scenarios have relative difference greater than 10%.

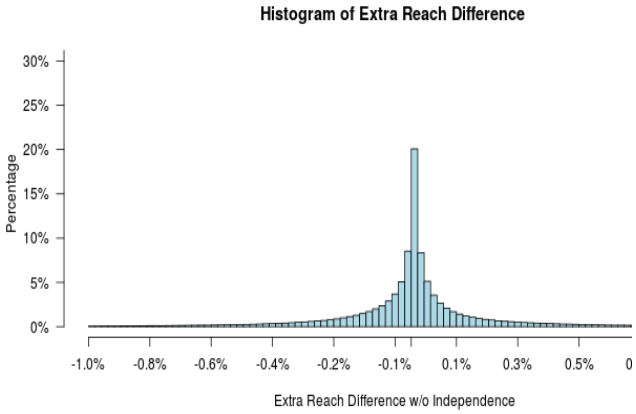


(a) Absolute Difference of YouTube Incremental Reach

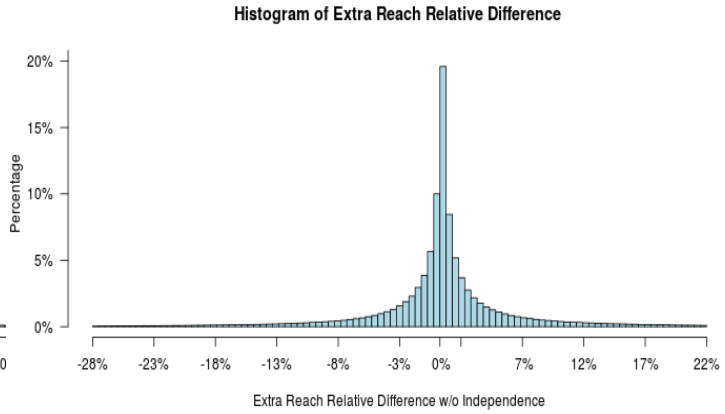


(b) Relative difference of YouTube Incremental Reach

Figure 1: Distribution of Incremental Reach Differences Based on Japan INTAGE Panel with Simulated TV Campaigns, Truncated to 95% of All Campaigns



(a) Absolute Difference of YouTube Incremental Reach



(b) Relative difference of YouTube Incremental Reach

Figure 2: Distribution of Incremental Reach Differences Based on Japan INTAGE Panel with Historical TV Campaigns, Truncated to 95% of All Campaigns

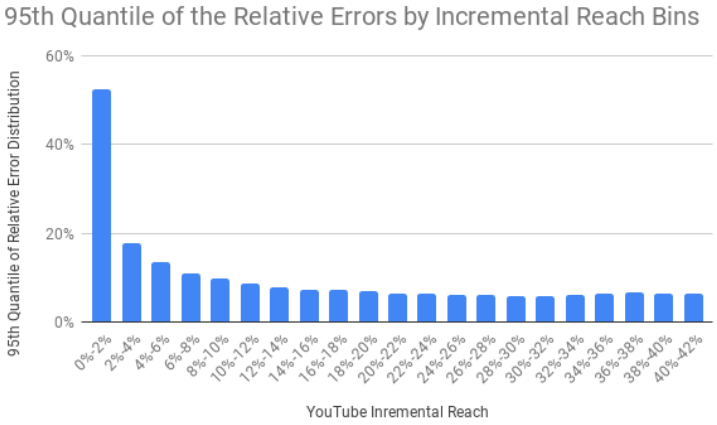


Figure 3: Distribution of 95th Quantile of Relative Differences By YouTube Incremental Reach Bins, Japan INTAGE Panel with Historical TV Campaigns

3.3 Variance of YouTube Incremental Reach

In this section, we compare the variance of the YouTube incremental reach calculations based on: 1) the combination of RCIA and invariant assumption; and 2) a cross-media universe. Different from the previous comparison on accuracy of RCIA, we use RCIA with single-media panels and invariant assumption for variance evaluation.

The variances are calculated by bootstrapping panelists within each non-overlapping demographic bucket. A total of 500 bootstrap samples and a subset of simulation scenarios from Section 3.1 are compared:

- $K+$: 1 and 3.
- Simulated TV campaigns only
- TV impressions: 3 settings at 20th, 50th, 80th quantiles of TV inventory impressions.
- YT impressions: 6 settings per device: 80th, 40th, 10th, 4th, 0.8th, 0th of total impressions.

are used in the evaluation.

The variance of YouTube incremental reach is measured by the coefficient of variation, the ratio of the standard deviation over the mean. Table 1 shows the coefficients of variation for the two methods of incremental reaches calculation. It is clear that RCIA using single-media panels can greatly reduce the variance of incremental reach compared to the cross-media universe based one. The variance reduction is partly achieved by the substantially larger panel size of single-media panels, i.e, 18000 panelists in single-media panels, and 5000 panelists in cross-media universes. When there is no demographic targeting, i.e., the case indicated by the last row in the table,

the coefficient of variation is merely 3% because the size of panelists is the sum of that across all individual demographic buckets. Such big variance reduction makes RCIA using single-media panels a strong competitor to the cross-media panel based calculation. It would be prohibitively expensive to recruit a cross-media panel to obtain similar variance as the RCIA using single media panels achieves.

Demographic Bucket	# of Scenarios	CV of YouTube Incremental Reach	
		RCIA	Cross-media
F[18, 24]	748	0.3%	29.9%
F[25, 34]	706	0.4%	15.3%
F[35, 44]	707	0.5%	16.6%
F[45, 54]	629	0.5%	21.7%
F[55+]	543	0.3%	32.5%
M[18, 24]	871	1.1%	21.6%
M[25, 34]	880	0.4%	11.2%
M[35, 44]	796	0.3%	11.0%
M[45, 54]	770	0.5%	12.2%
M[55+]	718	0.3%	17.3%
NONE	1,158	0.0%	3.2%

Table 1: Coefficient of Variations of YouTube Incremental Reach with Japan INTAGE Panel

4 Statistical Dependence of Video Ad Impressions

In this section, we use a statistical dependence measurement to investigate if there is statistically significant dependence between TV and YouTube ad impressions, i.e., to validate the conditional independence assumption in RCIA. Different from the comparison studies in Section 3, observed YouTube ad impressions are used in analysis. Intuitively, a person would have less time to watch YouTube if he already spend a lot of time watching TV. This suggests a negative correlation between TV and YouTube watch time. However, for reach planning and reporting purposes, we are interested in the relationship of TV and YouTube ad impressions, rather than watch time.

The performance of various dependence measurements has been studied extensively in the literature, such as Pearson's correlation, rank correlation, Cramer's V, mutual information, distance correlation and maximum information coefficient, etc. Because YouTube and TV ad impressions have high variance, low frequency per panelist and high concentrations at zero, many traditional dependence measurements become not suitable. In our study of statistical dependence, we use distance correlation by Szekely et al. (2007) and Szekely and Rizzo (2009) and its associated non-paramteric t-test of independence by Szekely and Rizzo (2013). Szekely et al. (2007) shows that distance correlation is always within in $[0, 1]$ and is zero only if the two random variables are independent. Moreover, distance correlation measures both linear and nonlinear correlation between two random variables. More technical details of distance correlation are in Appendix. Because we conduct statistical tests across all TV campaigns simultaneously, we apply the false discovery rate (FDR) method from Benjamini and Hochberg (1995) to control for the expected proportion of type

I errors.

Figure 4 shows the histograms of distance correlations of TV and YouTube ad impressions. Most of distance correlations are less than 10%. Independence testing results shown in Table 2 indicate that less than 5%, of TV campaigns are rejected for independence with significance level 0.05 after FDR adjustment. These results provide justification to use RCIA as a valid approximation method.

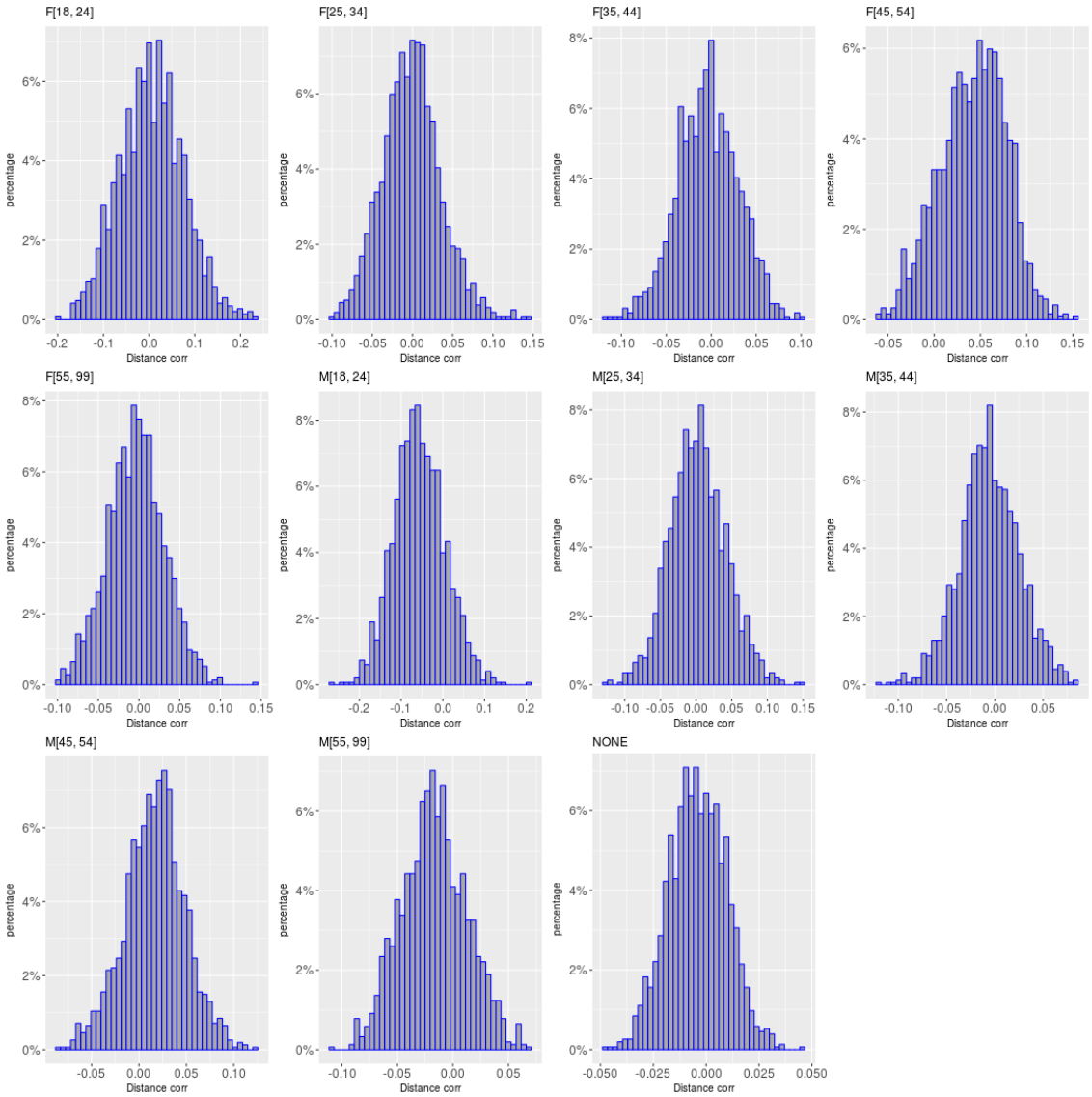


Figure 4: Distribution of Distance Correlations from Japan INTAGE Panel

5 Conclusion

In this work we show how to calculate YouTube incremental reach with the *restricted conditional independence assumption* (RCIA). Large scale simulation scenarios based on an SSP are generated

Demo buckets	# of Campaigns	Rejection Rate
F[18, 24]	1,451	2.8%
F[25, 34]	1,537	0.2%
F[35, 44]	1,538	0.5%
F[45, 54]	1,538	0.0%
F[55+]	1,538	4.8%
M[18, 24]	1,480	0.3%
M[25, 34]	1,538	2.0%
M[35, 44]	1,538	0.8%
M[45, 54]	1,538	0.5%
M[55+]	1,538	3.3%
NONE	1,538	3.7%

Table 2: Rejection Rates of Independence Testing on Distance Correlations with 0.05 Significance Level

to show that RCIA based incremental reach is very close to the standard calculation based on the cross-media universes in an SSP. The incremetal reach calculation with RCIA only still relies on cross-media universes. Thus, when only single-media panels are available which is common in practice, the invariant assumption is needed for the RCIA to work. The invariant assumption is a weaker assumption compared to independence assumption. A big advantage of RCIA plus invariant assumption based approximation is its low variance because it uses single-media panels which has much larger sample size compared to a cross-media one. Difficulty in cost of recruiting cross-media panelists are widely known in the ads industry. Furthermore, statistical independence testing indicates that only a small percentage of campaigns have statistically signinificant dependence between TV and YouTube ad impressions. While these results focus on incremental reach, they are applicable to all other types of cross-media metrics.

RCIA and invariant assumption based cross-media reach calculation should be strongly preferred than the standard calculation based on a cross-media universe in an SSP due to its high accuracy, low variance, simplicity, and low cost. More studies, especially those in more countries, would help to further validate our recommendation. We like to point out that this work focuses on the independence of TV and YouTube ad impressions and its conclusion may not be applicable to general video content viewing.

6 Appendix

6.1 Distance Correlation

Distance correlation is a relatively new measurement of dependence, developed in Szekely et al. (2007) and Szekely and Rizzo (2009). It uses the distance between data points as part of its calculation.

Let $(x_i, y_i), i = 1, \ldots, n$ be a statistical sample from pairs of random variables (X, Y). First compute

distance matrices A and B as:

$$A_{i,j} = \|X_i - X_j\|, \quad i,j = 1,\dots,n,$$
$$B_{i,j} = \|Y_i - Y_j\|, \quad i,j = 1,\dots,n,$$

where $\|\cdot\|$ denotes Euclidean norm. Then we take all doubly center distances $\bar{A}$ and $\bar{B}$ as

$$\bar{A}_{i,j} = A_{i,j} - \bar{A}_{i\cdot} - \bar{A}_{\cdot j} + \bar{A}_{\cdot\cdot},$$
$$\bar{B}_{i,j} = B_{i,j} - \bar{B}_{i\cdot} - \bar{B}_{\cdot j} + \bar{B}_{\cdot\cdot},$$

where $\bar{A}_{i\cdot}$ is the mean of the i-th row, $\bar{A}_{\cdot j}$ is the mean of the j-th column, and $\bar{A}_{\cdot\cdot}$ is the total mean of the distance matrix A. The distance covariance is defined as the square root of:

$$V_{xy}^2 = \frac{\sum_{i,j=1}^{n} \bar{A}_{ij}\bar{B}_{ij}}{n^2}. \tag{10}$$

The distance correlation is the square root of:

$$R_d^2 = \frac{V_{xy}^2}{V_{xx}V_{yy}}. \tag{11}$$

The population version of distance variance and distance correlation can be defined similarly as above, see Szekely et al. (2007) for details. Szekely et al. (2007) shows that $0 \leq R_d \leq 1$ and $R_d = 0$ only if the two random variables are independent. Because of these nice properties, we use distance correlation in our work here as the dependence measurement. The disadvantage of distance correlation is its large memory requirement. An approximation is to take a smaller but still large enough samples for large data sets. Szekely and Rizzo (2013) developed a non-parametric t-test of independence. The distribution of the test statistics is approximately distributed as standard normal when $n \geq 10$.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300.

Caffyn, J. M. and Sagovsky, M. (1963). Net audiences of british newspapers: A comparison of the agostini and sainsbury methods. *Journal of Advertising Research*, pages 21–25.

Chen, A., Koehler, J., Owen, A., Remy, N., and Shi, M. (2014). Data enrichment for incremental reach estimation. Technical report, Google Inc. https://ai.google.com/pubs/pub42246.

Goerg, G. M., Jin, Y., Remy, N., and Koehler, J. (2015). How Many Millennials Visit YouTube? Estimating Unobserved Events From Incomplete Panel Data Conditioned on Demographic Covariates. Technical report, Google Inc. https://ai.google.com/pubs/pub43451.

Jin, Y., Koehler, J., Goerg, G. M., and Remy, N. (2013). The optimal mix of tv and online ads to maximize reach. Technical report, Google Inc. https://ai.google.com/pubs/pub41669.

Lim, J. S., Ri, S. Y., Egan, B. D., and Biocca, F. A. (2015). The cross-platform synergies of digital video advertising: Implications for cross-media comapaigns in television, internet and mobile tv. *Computers in Human Behavior*, 48:463–472.

Nielsen (2017). Total audience report, q2 2017. Technical report, The Nielsen Company.

Szekely, G. J. and Rizzo, M. L. (2009). Brownian distance correlation. *The Annals of Statistics*, 3(4):1236–1265.

Szekely, G. J. and Rizzo, M. L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis*, 117:193–213.

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing independence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.