A Method for Measuring Online Audiences

Jim Koehler, Evgeny Skvortsov, Wiesner Vos

Google Inc.

Abstract

We present a method for measuring the reach and frequency of online ad campaigns by audience attributes. This method uses a combination of data sources, including ad server logs, publisher provided user data (PPD), census data, and a representative online panel. It adjusts for known problems with cookie data and potential non-representative and inaccurate PPD. It generalizes for multiple publishers and for targeting based on the PPD. The method includes the conversion of adjusted cookie counts to unique audience counts. The benefit of our method is that we get both reduced variance from server logs and reduced bias from the panel. Simulation results and a case study are presented.

1 Introduction

Advertisers would like to understand the attributes of the audience their ads are reaching. The definition of an audience commonly used for TV advertisers is the Gross Rating Point (GRP) [1], which is based on the reach and frequency of the audience by age and gender. Having these metrics for both online and TV ads would allow marketers to understand the aggregate performance of their marketing campaign in reaching their desired audience. Existing digital reporting practice, which measures cookies rather than people, makes it difficult for advertisers to know how the web, TV, and other platforms work together. Other audience breakdowns by a wider set of audience attributes such as ethnicity, education, income, and audience expressed interests are also possible.

This paper addresses a method to measure GRPs using a combination of data from several different sources to compute audience reach metrics: US census data, ad server logs from the ad serving network, publisher-provided self-reported demographic data, and a representative online panel. The number of people exposed to a campaign is inferred from the number of unique cookies exposed to these campaigns. For a subset of these cookies, demographic information is available from publisher provided data (PPD). These demographic labels may be incorrect for some of the cookies and the cookies with labels may not be representative of all cookies. Models are developed in Section 3 to adjust for possible bad and biased labels using panel data. A user is typically represented by multiple cookies, some of which may be shared with other users on the same device. Accurately inferring the number of users behind a given number of cookies as described in Section 4. These models are trained and evaluated using the online calibration panel data for which the true cookie-to-user relationships are known.

A probability-recruited online panel provides reliable data against which the GRP metrics are calibrated and verified based on statistical methodology. This panel should be aligned to the US

online population using data from the US Current Population Survey (CPS) [2] on key demographic variables such as age, gender, household income, and education level using demographic weights. The panel plays a key role in adjusting for demographic bias and cookie-sharing effects in PPD, in inferring models to accurately estimate the number of users behind aggregated cookie counts, and in evaluating the accuracy of the method.

Another approach to measure GRPs online would be through direct panel-based measurement. Direct measurement using a panel would require a very large panel to cover more than just the very largest campaigns and Web properties. It is therefore expensive and limited in coverage. Our proposed cookie-based method, calibrated using a smaller high-quality panel, is able to cover a larger part of the long tail of the web. The benefit of our method is that for our reach and frequency estimates, we get both reduced variance from server logs and reduced bias from the panel.

Section 2 of this paper discusses in detail the various data sources required for this method. Sections 3 and 4 develop the demographic correction models and cookie-to-user models, respectively. Sections 5 and 6 show simulations and illustrations of the method.

2 Data Sources

2.1 Server logs data and publisher provided data

Typically, online campaign performance is reported on the basis of the ad serving network's cookie, as most impressions are served to users who have a cookie. Ad server logs are a rich source of information in terms of audience reach and frequency of ad exposure. It also provides audience breakdown in terms of sites and real-time or near real-time audience data. However, cookies present significant technical challenges to overcome when used for audience measurement.

The first challenge is that a cookie does not identify a person, but a combination of a user account, a computer, and a web browser. Thus, anyone who uses multiple accounts, computers, or browsers has multiple cookies. Cookies do not differentiate between multiple users who share the same user account, computer, and browser [3]. In order words, a user can have multiple cookies and a cookie can have multiple users.

A second challenge is that not all cookies have demographic information attached to them, and when there is demographic information available it may be of questionable quality and biased. In order to obtain the demographic composition for an audience, at least a subset of the cookies for that audience need to have an age and gender label. Publisher provided data (PPD) can provide age and gender information, and potentially other audience attributes, for logged-in users via the ad request. This exchange of information is done anonymously so as to protect the user's identity. However, the composition of PPD is known to be biased toward the audience a publisher attracts, which does not necessarily reflect the US population. For example, demographics from publishers with a large number of users can be skewed to a younger audience. The quality of declared demographic information relies on the truthfulness of users and also the extent to which cookies are shared between multiple users.

The third challenge is that cookie deletion (or cookie churn) can also lead to inaccuracies in audience measurement, such as the overstatement of reach and understatement of frequency. It also impacts site-specific measurement by potentially leading to an overstatement of unique visitors and understatement of repeat visitors [4]. Finally, there is also inconsistent support for cookies across devices. Some mobile devices do not implement cookies for example. In this paper we introduce a

statistical method using online panel data to correct for these issues in aggregate.

2.2 Panel data

This method requires access to a high quality representative online panel. Typically a panel is recruited using geographic address based sampling (ABS) using the U. S. Postal Service Computerized Delivery Sequence File (CDSF) to match the US online population. This might include over-sampling to account for difficult to recruit targets. Measurement technology can be leveraged to monitor all online activity for individual panelists. For the purposes of this method, the panel is treated as the "truth". That is, we want our measurement system to mimic that reported by the panel but to have the ability to measure campaigns too small to assess using the panel.

Panel data present several challenges and limitations. The first of these is the tracking of out of home and/or work usage. The agreement to track panelist's devices does not extent to their employer-owned devices. Another important factor affecting the quality of individual level data is panelist compliance. The use of non-registered devices and the sharing of individually registered devices will distort panelist usage data. The cost of recruiting a high quality online panel is high, which limits panel size and therefore the ability to measure smaller campaigns.

Panel attrition is another issue which makes it challenging to maintain the representativeness of a panel despite probability-based recruitment and robust panel management processes. Bias can therefore occur in panels due to undercoverage of subsets of the population. This results in the demographic composition of the panel being biased with respect to the population. Demographic variables such as age, gender, household income, and education level are known to affect online behaviour. There is a growing base of statistical research on bias-correction of panels. In order to reduce bias, the data can be adjusted to correct those errors. Weighting adjustments [5] aim to reduce the variance of the estimates while adjusting for demographic differences between a sample and the population. This helps to adjust for the effects of panel attrition, that may cause panels to become less representative of the population over time. It is important to note that weighting based on demographic variables alone is not guaranteed to eliminate selection bias as panel bias may be related to factors other than demographics [6].

Several calibration methods exist such as generalized regression estimators (GREG) [7], RIM-weighting [8] and post-stratification. The CPS data, described in the following section, allows for calibrating the panel to the US online population on most key demographics. These methods leverage weighting methods in the R "survey" package [9]. GREG performs well at aligning a panel to the population, while it has the lowest variance among competing methods in our experience.

2.3 Population benchmarks

The Current Population Survey (CPS) [2], sponsored jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), is the primary source of labor force statistics for the population of the United States. The CPS is the source of numerous economic statistics, including the national unemployment rate, and provides data on a wide range of issues relating to employment and earnings. The CPS also collects extensive demographic data that complement and enhance our understanding of various population, economic and labor related measures, among many different population groups, in the states and in substate areas. The School Enrollment and Internet Use Supplement to the CPS [10] released in July 2011, contains a detailed breakdown of the US Internet population by demography and geography. This is the most detailed and comprehensive public

dataset on the US Internet population that we are aware of. It gives a detailed breakdown of the joint distribution for this population on key demographics such as age, gender, household income, education level and ethnicity. It also gives a detailed breakdown by geography to a sub-state level. It is worth noting that the definition of an Internet user we derive from this data is broad. Anyone who answered "yes" to either of the following questions would be regarded as an Internet user: Do you access the Internet at any location outside the home? At home, do you access the Internet? The current CPS data set surveys individuals age three and over.

3 Demographic Correction Models

The estimate of a campaign's audience consists of taking the total number of ad impressions, unique cookies exposed to the campaign, a subset of cookies which have PPD demographic labels, and then breaking down the impressions and uniques cookies into the demographic groups. The cookies in each group can be converted to unique users using the method from Section 4. The impressions from each group, when divided by that group's population number and multipled by 100, estimate the GRPs for that group. Finally, these impressions divided by the number of exposed users from that group, estimates the average frequency for that group. This section develops models to break either impressions or cookies into the demographic groups. We begin by a simple example and then add complexity before presenting the general model. Lastly, we describe how to evaluate the performance of the models.

3.1 Simple Example

Consider a simple example using D demographic groups and one publisher. Suppose that the PPD is unbiased in representing the population. Our only concern is possibly poor quality of the PPD labels. We can quantify this quality by an unknown D x D misclassification matrix P where $[P]_{ij} = Pr(\text{label } i|\text{truth } j)$. Hence the diagonals of P represent the fraction of the labels that are correct for each demographic group and represent a level of confidence of a correct label for that group. If a campaign had reached cookies (or impression) with the demographic breakdown represented by vector q_{truth}, then on average the PPD will estimate this as

$$q_{\text{PPD}} \approx P \cdot q_{\text{truth}}$$

and if we know P then we could estimate the q_{truth} by inverting P [1]

$$\hat{q}_{\text{truth}} = P^{-1} \cdot q_{\text{PPD}}$$

Estimates of P can be obtained perhaps by some subset of higher quality PPD or by leveraging the panelist cookie data. Unfortunately, these type of estimates are not great and any estimation errors can cause large changes in the inverse of P. Further, the assumption that the misclassification matrix is same for all sites/campaigns is most likely incorrect. Hence, we need a more robust method of estimating a correction matrix.

A better approach is to create a training set of campaigns and/or site visit data that is measured by both the panel and PPD. Let this data consist of N_{train} campaigns/sites each large enough to be confidently measured by the panel. For campaign (or site) i, let y_i be the proportion of panelist

[1] P may not be invertible but we ignore this uninteresting case.

cookies (or impressions) for each of the demographic groups (hence y_i is a vector of length D) and let x_i be a D-length vector of similar metric from the PPD. We model the relationship as

$$y_i = Ax_i + \epsilon_i \tag{1}$$

where A is D x D left-stochastic matrix [2]. That is, it re-distributes the PPD demographic proportions to better represent the actual population proportions for those exposed to the campaign or who visited the site. The non-negative restriction can be relaxed but the estimated y elements may then be negative. This can be fixed by setting all entries of y to $\max(y, 0)$ and then renormalizing: $y_{\text{new}} = y/|y|$. This model can be fit via least squares using a constrained optimization routine if there are constraints or by least squares or penalized least squares if there are no constraints.

3.2 General Model

The model presented above handles misclassification problems but not possible representation issues of the PPD. Most PPD do not represent the online population as the demographics are collected when users sign-up for their account and not all demographic groups are equally attracted to a given publisher or equally willing to share this information. One approach to fix the representation issue is to use propensity methods [11] to weight the PPD cookies to better represent the general cookie population. These weights could be based on age, gender, site visited, and browser and/or device. The cookie weights would also propogate to the impressions to provide impression weights. This approach is tricky to implement correctly and beyond the scope of this paper. However, if adequate weights are determined, then the model specified in (1) could be used after weights are applied to the cookies or propogated to impressions. Our approach is to include the ability to weight the PPD inside the correction matrix. That is, we modify our model to

$$y_i = Ax_i/|Ax_i| + \epsilon_i \tag{2}$$

where A is no longer a left-stochastic matrix although the entries should be non-negative. The diagonals of A now are increased to upweight underrepresented demographic groups and decreased to downweight overrepresented groups. However, since the model normalizes the estimate, the actual entries of A do not have intuitive meaning since rescaling A will produce the same fit.

This model is problematic if the publisher providing the PPD is also a publisher participating in the campaign or one of the sites visited. It is also a problem if the PPD is used as the targeting criteria for a targeted campaign. The model makes the PPD look like the online population and hence overcorrects the estimate for that publisher or those targeted cookies. The PPD should be representative of the publisher's site although the quality of the labels is still a concern. A general model to handle this is

$$y_i = (1 - \alpha_i)Ax_i/|Ax_i| + \alpha_i Bx_i + \epsilon_i \tag{3}$$

where α_i represents the fraction of cookies (impressions) for the ith campaign either served on the publisher's site or via cookie targeting using the PPD. Hence, $1 - \alpha_i$ represents the fraction of unlabeled cookies (or unlabeled impressions) for that campaign. If the PPD labels are perfect for the publisher's site, then $B = I$. But as in (1), if PPD has misclassification issues then B should be a left-stochastic matrix. We do not expect PPD to be biased for estimating activity on that

[2] A left-stochastic matrix is a square matrix with non-negative entries and columns that sum to one.

publisher's site or for cookies targeted using that PPD. It might be possible to estimate B from panel cookies that have PPD labels directly rather than via a model fit. Matrix A should be fitted using either least squares or penalized least squares.

3.3 Model for Multiple PPD

Model (3) can be extended to handle multiple PPD. Suppose we have multiple PPD sources and for the i-th campaign we can group cookies (impressions) into $M+1$ disjoint groups. The first group contains all unlabeled cookies (impressions) for the campaign. The other M groups represent sets of PPD that contain labels for that group. For example, if we have two PPDs, then M might be three with the first group consisting of cookies (impressions) with labels only from the first publisher, the second group with labels only from the second publisher, and the third group having labels from both publishers. When there are disagreements between publishers, a rule will be applied to either use only one publisher and hence that cookie (impression) will be in one of the first two groups or to probabilistically allocate that cookie to the "shared" group.

Now let the distribution of cookies (impressions) for the i^{th} campaign be represented by α_{im} for the $m = 1, \ldots, M$ groups and let $\alpha_{i\cdot} = \sum_{m=1}^{M} \alpha_{im}$. Hence we have $1 - \alpha_{i\cdot}$ unlabeled cookies (impressions). Further, for each of the M groups we have x_{im} which are the demographic proportion estimates as before. To model the unlabeled cookies we use all of the labeled data so now $x_i = (x'_{i1}, x'_{i2}, \ldots, x'_{iM})'$. The general model is then

$$y_i = (1 - \alpha_{i\cdot})Ax_i/|Ax_i| + \sum_{m=1}^{M} \alpha_{im} B_m x_{im} + \epsilon_i \tag{4}$$

where A is a D x DM matrix and there are M B_m left-stochastic matrices each of size D x D. Clearly as M increases, the demands on the training data increase as the number of coefficients increase linearly with M and hence a larger training dataset (M_{train}) is required or substantial constraints or least squares penalties need to be applied.

3.4 Model Evaluation

A representative and weighted panel as described in Section 2.2 should provide relatively unbiased estimates for the online population, and is arguably the best available benchmark against which to estimate the relative error of adjusted cookie estimates. Although unbiased, the limited size of the panel means that panel based benchmarks have high variance. When we directly compare our audience estimates against direct panel based estimates, even for the largest campaigns, the true measurement error attributed to our methodology is obscured by the variability attributed to the panel. This requires the removal of the panel variance from the error estimation, as outlined below.

We are trying to estimate the relative error of the adjusted cookie estimates against the (unknown) truth. Let p_P and p_C be the estimated proportions of the population belonging to a specific demographic bucket from the panel and cookies, respectively. If we believe that the panel provides an unbiased estimate, then we have $E[p_P]$ as an estimate of the truth. The error we want to estimate can therefore be expressed as $E[(p_C - E[p_P])^2]$. In practice we observe p_C and p_P. When we condition on the actual demographic proportions $E[p_P]$, it seems reasonable to assume that variations around this estimate are independent—or stated otherwise, to assume conditional independence between p_P and p_C. Then it follows that

$$\begin{aligned} Var(p_C - p_P) &= Var((p_C - E[p_P]) + (E[p_P] - p_P])) \\ &= Var(p_C - E[p_P]) + Var(p_P - E[p_P]) \\ &= \sigma_1^2 + \sigma_2^2 \end{aligned} \tag{5}$$

The term σ_1^2 in (5) represents the relative error we want to estimate. This can be viewed as relative error of the cookie estimate to the panel estimate, if we had a really large panel. The second term is the variance of the panel estimate, which we can estimate from the binomial distribution as $\frac{p(1-p)}{n}$ by plugging in the panel estimate p_P for p. So typically, n is the number of panelists who reached a site or were exposed to a campaign and p_P is the ratio of subset of exposed panelist matching that demographic label to n. When we subtract our estimate of the second component above from the left-hand side of Equation 5, we obtain an unbiased estimate of the $Var(p_C - E[p_P])$. This estimate could be negative for some campaigns, although we observe negative estimates for a very small proportion of campaigns in practice. Averaging over many campaigns should give us a better indication of the expected cookie deviation for the "truth" that is not obscured by the panel variability. By using a Gaussian approximation, we can estimate average precision across campaigns at a $(1-\alpha) \times 100\%$ confidence level as $\frac{z_{1-\alpha/2} \times \hat{\sigma_1}}{p_C}$.

4 Mapping Cookies to Users

It is widely known that the number of cookies that are exposed to a campaign may be much larger than the number of real-world users. There are many factors that cause this discepancy. For example, a user may have accessed a particular site from different computers. Or a user may have cleaned their browser cookies and revisited the page. ComScore estimates [4] that using cookies directly as a measurement of audience volume may inflate the numbers up to 2.5 times. We have investigated the patterns of cookie deletion behavior and determined the parameters of a model that transforms cookie counts to people counts.

Let L be some web location, it can be a campaign, a site, a URL, etc. Let T be some timespan. We denote by C_{LT} the number of cookies that have been observed at L in the duration of T. We denote by P_{LT} the number of people that have visited L in the duration of T. Let $\mathcal{I}$ be the whole internet and consequently $P_{\mathcal{I}T}$ be the total number of internet users during timespan T and $C_{\mathcal{I}T}$ the total number of cookies generated in the duration of T.

The probability of a randomly picked cookie visiting a web location L in timespan T can be calculated as

$$P(\text{cookie visits } L \text{ in timespan } T) = \frac{C_{LT}}{C_{\mathcal{I}T}},$$

and analogously for a random person this probability is computed as

$$P(\text{person visits } L \text{ in timespan } T) = \frac{P_{LT}}{P_{\mathcal{I}T}}.$$

Recall that for an event with probability p, the odds are defined as the number $\frac{p}{1-p}$. Thus we have

$$Odds(\text{cookie visits } L \text{ in timespan } T) = \frac{C_{LT}/C_{\mathcal{I}T}}{1 - C_{LT}/C_{\mathcal{I}T}} = \frac{C_{LT}}{C_{\mathcal{I}T} - C_{LT}},$$

and

$$Odds(\text{person visits } L \text{ in timespan } T) = \frac{P_{LT}}{P_{\mathcal{I}T} - P_{LT}}.$$

Our investigations of panel data have indicated that the odds of a panelist and a cookie belonging to a particular web audience are approximately linear:

$$\frac{P_{LT}}{P_{\mathcal{I}T} - P_{LT}} = \gamma_T \cdot \frac{C_{LT}}{C_{\mathcal{I}T} - C_{LT}}. \tag{6}$$

To determine the size of the real-world audience we can rearrange this relationship:

$$P_{LT} = \frac{\gamma_T C_{LT} \cdot P_{\mathcal{I}T}}{C_{\mathcal{I}T} + C_{LT}(\gamma_T - 1)}. \tag{7}$$

Note that all counts that occur in the right side can be obtained for each time-interval T and web location L from logs and panel data:

- C_{LT} the number of cookies seen at the web location L in time-interval T is obtained from the logs,
- $C_{\mathcal{I}T}$ the total number of cookies generated by all users in the interval T is obtained from the logs,
- $P_{\mathcal{I}T}$ the total number of internet users active during time interval T can be estimated using panel and internet census data.

In order to determine an appropriate γ_T we use the following algorithm:

1. For a set of web locations indexed by L (e.g. websites or campaigns) with each web location having the number of panelists visting, N_L, exceeding some threshhold, we count the number of cookies and people pair
$$p_L = (C_L, P_L).$$
2. For each pair p_L calculate parameter γ_L that fits the data perfectly. This is easily calculated by re-arranging Equation 6
$$\gamma_L = \frac{P_L C_{\mathcal{I}} - P_L C_L}{C_L P_{\mathcal{I}} - P_L C_L}$$
3. Set γ to be a weighted median of γ_L where the weight of web location L is proportional to N_L.

We have also used panel data to explore how γ_T depends on T. It turns out that with reasonably high accuracy for any T we have

$$\gamma_T = \Delta \cdot \frac{C_{\mathcal{I}T}}{P_{\mathcal{I}T}},$$

where Δ is a constant. Δ varies from country to country, but is never far from 1.0. Therefore for countries for which we do not have panel data we can apply the technology using Δ_0 which is the median of Δ over all countries for which we have panel data. That is we have

$$\gamma_T = \frac{C_{\mathcal{I}T}\Delta_0}{P_{\mathcal{I}T}}.$$

An illustration of the people and cookie relationship is presented in Section 6.1.

5 Simulation

This section present the results of a simulation study to understand the behavior of the correction models from Section 3. Two different simulations are shown with medium and severe misclassification matrices as discussed below. For both simulations, we used a 5,000 person representative panel and simulated 200 "training" campaigns and 1,000 "testing" campaigns that each reached at least 100 panelists. In practice, we use the cookies and impressions measured from the panelists as the calibration data. However, due to the difficulty of modeling and hence simulating cookie data, we only simulated panelists as reached or not reached for each campaign. Hence our simulation investigates the behavior of the demographic correction models and not the cookie-to-user model.

The overall reach of these campaigns were simulated using a beta distribution:

$$R \sim Beta(a = 0.6, b = 13.0)$$

where R is the proportion of the total population reached by the campaign. We considered 14 demographic age/gender groups with age groups: $\leq$ 17, 18 - 24, 25 - 34, 35 - 44, 45 - 54, 55 - 64, and 65+. The demographic proportions were simulated for each of the campaigns by simulating the gender breakdown as

$$q^{(G)}_{\text{truth}} \sim Beta(a = 6.0, b = 6.0)$$

and the age distribution as

$$q^{(A)}_{\text{truth}} \sim Dirichlet(\boldsymbol{\alpha} = 15 \times (0.10, 0.17, 0.25, 0.15, 0.16, 0.11, 0.06)')$$

and then creating q_{truth} by multiplying the respective elements of $q^{(G)}_{\text{truth}}$ and $q^{(A)}_{\text{truth}}$. We determined if a panelist j was reached by campaign i by using a Bernoulli trial with probability $R_i \times q_{\text{truth},i}(d(j))/w(d(j))$ where $d(j)$ indicates which demographic group panelists j belongs to and $w(\cdot)$ are the demographic population benchmarks discussed in Section 2.3. Any campaign that failed to reach at least 100 panelists was dropped and replaced by a campaign with sufficient reach.

We calculated the PPD measured demographic proportions by using misclassifications and weighting matrices specified for each simulation. That is, we calculated for each campaign i

$$q_{\text{PPD},i} = W \cdot P \cdot q_{\text{truth},i}$$

where P is the simulation specific misclassification matrix (see Section 3.1) and W is a weighting matrix. For each simulation, we weighted males twice that of females to mimic a publisher that has twice as many males as females registered users. Hence, W is a 14×14 diagonal matrix with 4/3 for male entries and 2/3 for female entries.

5.1 Medium Correction

The first simulation mimics a misclassification matrix where the significant misclassifications come from gender and adjacent age groups. The misclassification matrix, P_1 is shown in Table 1 and has diagonal elements (correct classification rates) ranging from 0.64 - 0.75. For this simulation, we inject additional noise by perturbing P_1 for each campaign i by $P_{1i} = P_1 + U_i$ where U_i is a

matrix with each element Uniform(-0.02, +0.02). We set negative entries of P_i to zero and then renormalize the columns.

Appendices that include all figures and tables are included at the end of the paper. Figure 1 shows the results for the Male 18–24 group for 200 of the test campaigns. Figure 1(a) shows the unadjusted PPD cookies and shows a tendency to over-predict (under-predict) small (large) campaigns for this demographic group. On average, the root mean squared error (RMSE) for the unadjusted cookies is 0.0193. Figure 1(b) shows the results for unconstrained regression (Model 1). For this model, any negative estimates were shifted to zero and then the $\hat{q}_i$ was renormalized. This model performed the best with no apparent bias in the fit and an average RMSE of 0.0080. Figure 1(c) shows the results of the normalized correction model (Model 2) fit using non-negative least squares. This model does a partial correction as the RMSE has been reduced from the unadjusted cookies to 0.0186 but still has significant bias issues. For reference we include the panel based measurements in Figure 1(d) - it has RMSE of 0.0164. The results for the panel are misleading as we require each campaign to reach at least 100 panelists while the other methods could be used to measure smaller campaigns. By construction of the simulation, the size of the campaigns do not affect the non-panel results.

We also show the results for Female 45 - 54 in Figure 2. For this demographic bucket, the unadjusted PPD cookies generally under predict and have larger RMSE of 0.0297. This is not surprising as the PPD has twice as many men as women. Once again, the unconstrained regression performed well with no apparent bias and RMSE of 0.0094. The normalized correction model fit using non-negative least squares does only a partial adjustment and has RMSE of 0.0214 while the panel RMSE is 0.016. The RMSE results for all demographic groups are shown in Table 2.

5.2 Severe Correction

For this simulation we create a misclassification that mimics the impact of cookie sharing problems - that is, gender and generational misclassifications. We generated P_2 by

$$P_2 = I_{14} + 0.3 \times \begin{pmatrix} 0_7 & I_7 \\ I_7 & 0_7 \end{pmatrix} + 0.15 \times \begin{pmatrix} C & C \\ C & C \end{pmatrix} + U$$

where I_k and 0_k are the identity and zero matrices of size k,

$$C = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and U is a matrix with elements from Uniform(0, 0.04). The second and third terms create gender and generational misclassification. Lastly, we renormalized the columns of P_2. For this simulation we did not perturb P_2 by campaign. The generated P_2 is given in Table 3. The diagonal entries are lower than from the first simulation ranging from 0.46 - 0.57.

Again we show the results only for Males 18–24 and Females 45–54 and these are shown in Figures 3 and 4, respectively. The tendencies for this simulation are similar to the first simulation, although

the magnitudes are different. For the unadjusted cookies we have RMSE of 0.0286 and 0.0464 for our two demographic groups while the unconstrained regression (1) does an even better job of correcting these errors. Now the RMSE are 0.0048 and 0.0038. The normalized regression makes partial corrections with RMSE of 0.0259 and 0.0271. The panel results are similar as before with RMSE of 0.0158 for both demographic groups. All RMSE results are given in Table 4.

We have performed other simulations with different misclassification matrices, more severe weighting, and with more severe campaign specific perturbations of the misclassification matrix. The results typically are ranked the same as presented above although the magnitude of the RMSE are different. As the misclassification matrix, the weights, or campaign specific perturbations increase, all methods have larger RMSE. Across all the simulations, the unconstrained regression model consistently had no trend bias although the tightness of the scatter around the identity line changes. The normalized regression model had improved performance if the non-negatitivity constraint was removed and any negative reach estimates were shifted to 0.

6 Illustrations

6.1 Converting Cookie Counts to People

The cookie-to-user model performance was evaluated using panel data on domains, campaigns, and synthetic campaigns data. In particular, Figure 5 shows a scatterplot where the points are US-based audiences of domains. The abscissa is the number of panel cookies observed at the domain and the ordinate is the number of people reached. It is evident that the number of people reached can be obtained as a function of the number of cookies reached. Our experiments have demonstrated that the trend of the scatterplot remains constant for different timespans and countries. In Figure 6, points represent domains and random synthetic media plans. Each synthetic media plan, M, is the union of a set of domains D_i. That is, a cookie or a person is considered to be reached by media plan M if it visited at least one of the domains D_i. This figure demonstrates that audiences of domains and the union of domains have the same trend, and a single model can be used to predict people audience for campaigns that run on one or multiple sites.

6.2 Case Study

The Google Nexus campaign was run from August 30, 2012 to September 30th, 2012. The campaign served more than 500 million impressions across more than 20 sites that included pre-dominantly news and technology sites. There was no demographic targeting on the campaign. The campaign was large enough to be measured reliably using panel data. In Figure 7, we compare 3 different estimates of the overall audience breakdown for this campaign. The first is a direct panel-based estimate for the weighted panel (shown in green), while the second is estimated using our GRP method (shown in red). The third estimate is obtained from counts of unadjusted YouTube cookies with demographic labels (shown in blue). It is clear that the GRP estimates closely mimics the direct panel based estimates, which is the aim of the method as it is calibrated to the panel. The effect of the panel-based calibration is clear from the difference between the GRP estimates and raw cookie-based YouTube estimate. Not only is the YouTube estimate here based on cookie counts and not users, but there is a clear bias towards younger audiences, especially young males, relative to the calibrated panel estimates.

7 Conclusion

We have presented a method for measuring the reach and frequency of online ad campaigns by auduence attributes. This method combines ad server logs, publisher provided user data, US census data, and a representative panel to produce corrected cookie and impression counts by these audience attributes. The method corrects for cookie issues such as deletion and sharing, and for PPD issues such as non-representativeness and poor quality of labels. It also generalizes for multiple PPD and for targeted campaigns. The method uses a model that converts cookie counts to users counts.

We presented a simulation study which demonstrates that the method has the ability to correct for non-representative and inaccurate PPD. Our simulation results show that the unconstrained regression model consistently gave the best results. However, real data have nuances not considered in the simulation and we advise to investigate the performance of multiple models on real world data in practice. We also presented the performance of the methods on real data. We demonstrated that the cookie to user model matches real data well, including media plans defined as domains and collections of domains. Finally, we showed the measurement of an online campaign and showed that the correction method matches the panel measured results.

References

[1] American Marketing Association Dictionary.
http://www.marketingpower.com/_layouts/Dictionary.aspx

[2] US Census Bureau and U.S. Bureau of Labor Statistics. Current Population Survey. *http://www.census.gov/cps/* or *http://www.bls.gov/cps/*

[3] A. M. Hormozi. Cookies and privacy. *EDPACS*. Vol.32, Iss. 9; pp. 1-13, 2005.

[4] Comscore. Cookie Deletion Whitepaper, 2007.
http://www.comscore.com/Insights/Presentations_and_Whitepapers/2007/Cookie_Deletion_Whitepaper.

[5] J. Bethlehem. Selection bias in web surveys. *International Statistical Review*, 78(2), 161-188, 2010.

[6] G. Loosveldt and N. Sonck. An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, 2(2), 93-105, 2011.

[7] J. C. Deville and C. E Sarndal. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87: 376-382, 1992.

[8] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4), 427-444, 1940.

[9] T. Lumley. Analysis of Complex Survey Samples *Journal of Statistical Software*, 9(1), 1-19, 2004.

[10] US Census Bureau and U.S. Bureau of Labor Statistics. Current Population Survey: Computer and Internet Use File, July 2011.
http://www.census.gov/apsd/techdoc/cps/cpsjul11.pdf.

[11] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrica*, 70(1), 41-55, 1983.

Tables

REFERENCES

	M17-	M18-24	M25-34	M35-44	M45-54	M55-64	M65+	F17-	F18-24	F25-34	F35-44	F45-54	F55-64	F65+
M17-	0.750	0.095	0.028	0.019	0.019	0.019	0.019	0.079	0.010	0.003	0.002	0.002	0.002	0.002
M18-24	0.095	0.674	0.095	0.028	0.019	0.019	0.019	0.010	0.071	0.010	0.003	0.002	0.002	0.002
M25-34	0.028	0.095	0.665	0.095	0.028	0.019	0.019	0.003	0.010	0.070	0.010	0.003	0.002	0.002
M35-44	0.019	0.028	0.095	0.665	0.095	0.028	0.019	0.002	0.003	0.010	0.070	0.010	0.003	0.002
M45-54	0.019	0.019	0.028	0.095	0.665	0.095	0.028	0.002	0.002	0.003	0.010	0.070	0.010	0.003
M55-64	0.019	0.019	0.019	0.028	0.095	0.674	0.095	0.002	0.002	0.002	0.003	0.010	0.071	0.010
M65+	0.019	0.019	0.019	0.019	0.028	0.095	0.750	0.002	0.002	0.002	0.002	0.003	0.010	0.079
F17-	0.040	0.005	0.002	0.001	0.001	0.001	0.001	0.711	0.090	0.027	0.018	0.018	0.018	0.018
F18-24	0.005	0.036	0.005	0.002	0.001	0.001	0.001	0.090	0.639	0.090	0.027	0.018	0.018	0.018
F25-34	0.002	0.005	0.035	0.005	0.002	0.001	0.001	0.027	0.090	0.630	0.090	0.027	0.018	0.018
F35-44	0.001	0.002	0.005	0.035	0.005	0.002	0.001	0.018	0.027	0.090	0.630	0.090	0.027	0.018
F45-54	0.001	0.001	0.002	0.005	0.035	0.005	0.002	0.018	0.018	0.027	0.090	0.630	0.090	0.027
F55-64	0.001	0.001	0.001	0.002	0.005	0.036	0.005	0.018	0.018	0.018	0.027	0.090	0.639	0.090
F65+	0.001	0.001	0.001	0.001	0.002	0.005	0.040	0.018	0.018	0.018	0.018	0.027	0.090	0.711

Table 1: Misclassification matrix for medium correction simulation

	Unadjusted Cookies	Unconstrained Regression	Normalized Regression	Panel
M17-	0.0179	0.0062	0.0150	0.0128
M18-24	0.0193	0.0080	0.0186	0.0164
M25-34	0.0202	0.0081	0.0199	0.0197
M35-44	0.0216	0.0082	0.0189	0.0159
M45-54	0.0182	0.0072	0.0195	0.0159
M55-64	0.0173	0.0071	0.0163	0.0136
M65+	0.0171	0.0053	0.0130	0.0101
F17-	0.0163	0.0079	0.0146	0.0133
F18-24	0.0286	0.0092	0.0194	0.0163
F25-34	0.0427	0.0108	0.0215	0.0184
F35-44	0.0240	0.0094	0.0192	0.0161
F45-54	0.0297	0.0094	0.0214	0.0160
F55-64	0.0219	0.0081	0.0168	0.0139
F65+	0.0138	0.0079	0.0185	0.0109

Table 2: Root Mean Squared Error for medium correction simulation

	M17-	M18-24	M25-34	M35-44	M45-54	M55-64	M65+	F17-	F18-24	F25-34	F35-44	F45-54	F55-64	F65+
M17-	0.466	0.010	0.006	0.086	0.084	0.017	0.013	0.154	0.002	0.011	0.090	0.080	0.015	0.021
M18-24	0.005	0.469	0.016	0.010	0.074	0.072	0.007	0.007	0.156	0.006	0.003	0.076	0.078	0.002
M25-34	0.014	0.012	0.458	0.017	0.011	0.085	0.101	0.015	0.003	0.147	0.005	0.006	0.070	0.085
M35-44	0.071	0.018	0.014	0.545	0.013	0.003	0.004	0.072	0.002	0.001	0.184	0.011	0.012	0.000
M45-54	0.085	0.074	0.005	0.003	0.470	0.011	0.012	0.082	0.083	0.013	0.006	0.148	0.017	0.007
M55-64	0.005	0.074	0.067	0.014	0.007	0.460	0.011	0.000	0.072	0.088	0.003	0.004	0.149	0.004
M65+	0.016	0.015	0.079	0.012	0.006	0.008	0.542	0.012	0.003	0.076	0.007	0.013	0.012	0.167
F17-	0.151	0.001	0.016	0.090	0.072	0.012	0.020	0.464	0.001	0.008	0.106	0.073	0.010	0.008
F18-24	0.006	0.145	0.015	0.008	0.077	0.078	0.003	0.003	0.486	0.010	0.008	0.073	0.078	0.009
F25-34	0.013	0.004	0.148	0.006	0.004	0.077	0.086	0.015	0.019	0.478	0.003	0.002	0.083	0.098
F35-44	0.074	0.009	0.010	0.165	0.009	0.018	0.009	0.074	0.002	0.009	0.565	0.018	0.001	0.018
F45-54	0.074	0.080	0.009	0.015	0.146	0.005	0.013	0.082	0.072	0.011	0.004	0.488	0.010	0.001
F55-64	0.018	0.080	0.074	0.014	0.012	0.148	0.006	0.001	0.090	0.072	0.011	0.001	0.458	0.013
F65+	0.003	0.008	0.081	0.015	0.018	0.007	0.173	0.018	0.008	0.071	0.005	0.008	0.007	0.568

Table 3: Misclassification matrix for severe correction simulation

	Unadjusted Cookies	Unconstrained Regression	Normalized Regression	Panel
M17-	0.0388	0.0036	0.0247	0.0131
M18-24	0.0286	0.0048	0.0259	0.0158
M25-34	0.0269	0.0035	0.0286	0.0201
M35-44	0.0218	0.0040	0.0208	0.0148
M45-54	0.0314	0.0058	0.0284	0.0150
M55-64	0.0409	0.0025	0.0264	0.0137
M65+	0.0354	0.0026	0.0198	0.0101
F17-	0.0306	0.0034	0.0252	0.0133
F18-24	0.0496	0.0044	0.0270	0.0166
25-34	0.0754	0.0054	0.0286	0.0196
F35-44	0.0415	0.0069	0.0191	0.0158
F45-54	0.0464	0.0038	0.0271	0.0158
F55-64	0.0323	0.0056	0.0276	0.0132
F65+	0.0186	0.0025	0.0224	0.0097

Table 4: Root Mean Squared Error for severe correction simulation

Figures

Figure 1: Correction performance for medium correction – Males 18-24 – using (a) unadjusted cookies, (b) unconstrained regression, (c) normalized regression, and (d) panels.

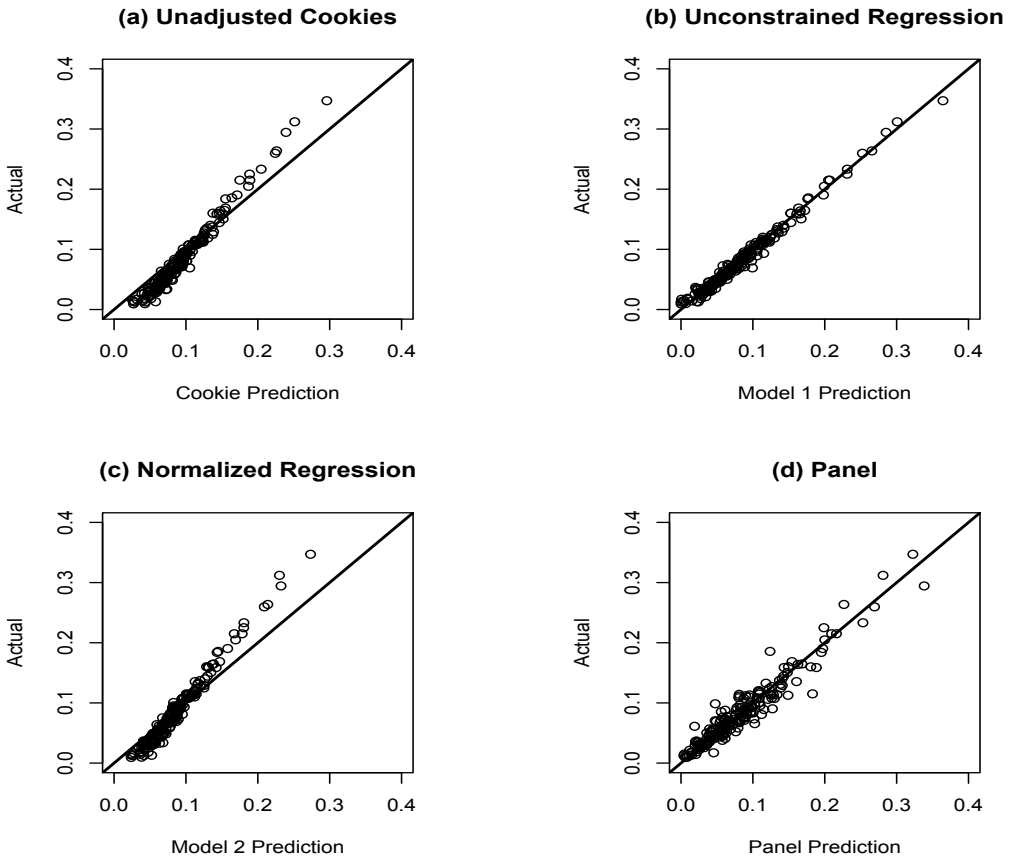

Figure 2: Correction performance for medium correction – Females 45-54 – using (a) unadjusted cookies, (b) unconstrained regression, (c) normalized regression, and (d) panels.

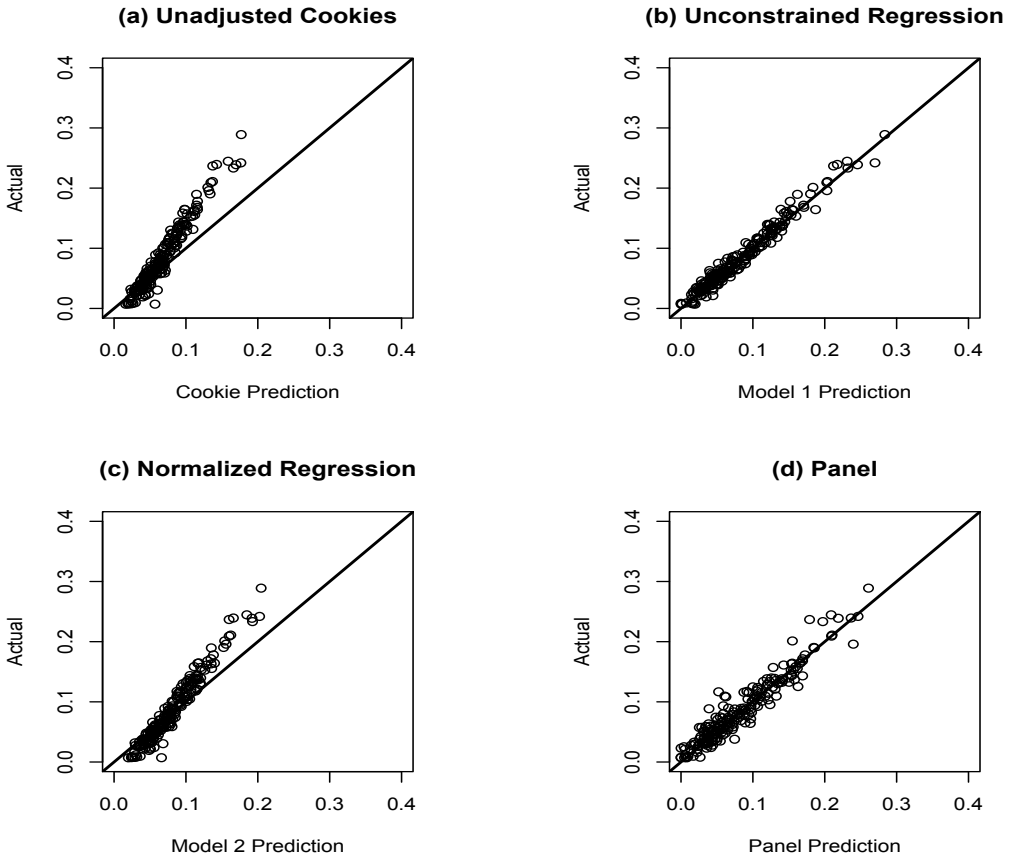

Figure 3: Correction performance for severe correction – Males 18-24 – using (a) unadjusted cookies, (b) unconstrained regression, (c) normalized regression, and (d) panels.

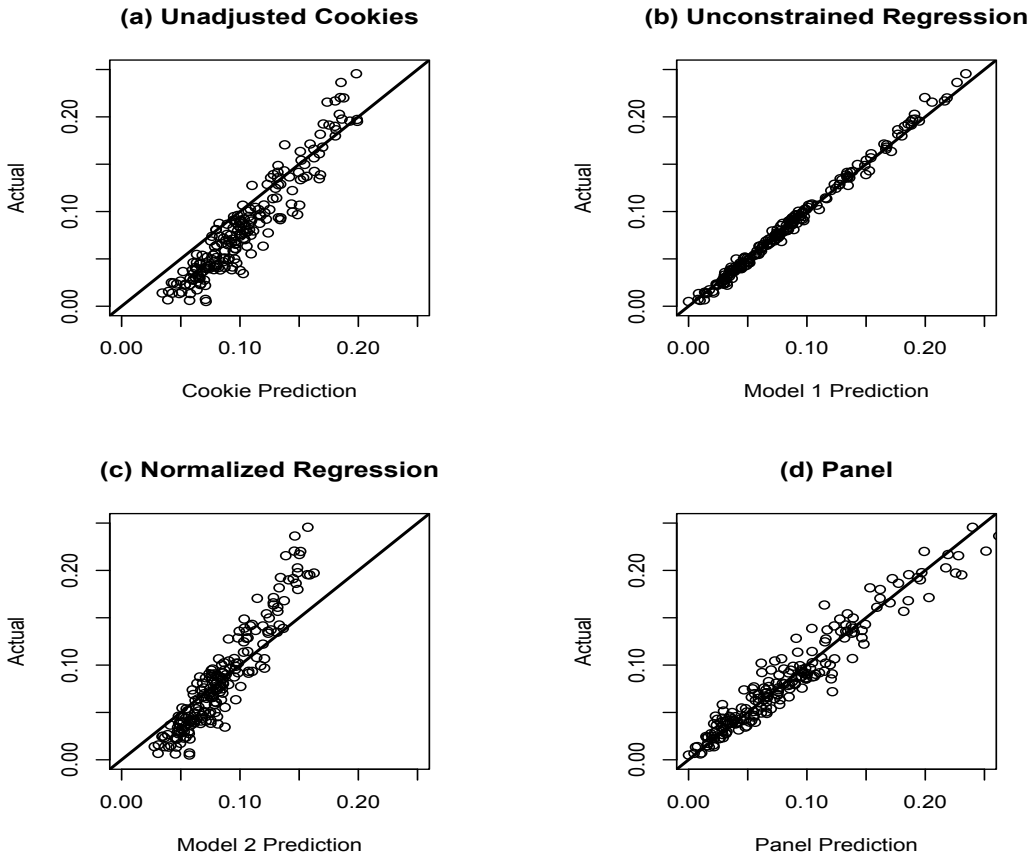

Figure 4: Correction performance for severe correction – Females 45-54 – using (a) unadjusted cookies, (b) unconstrained regression, (c) normalized regression, and (d) panels.

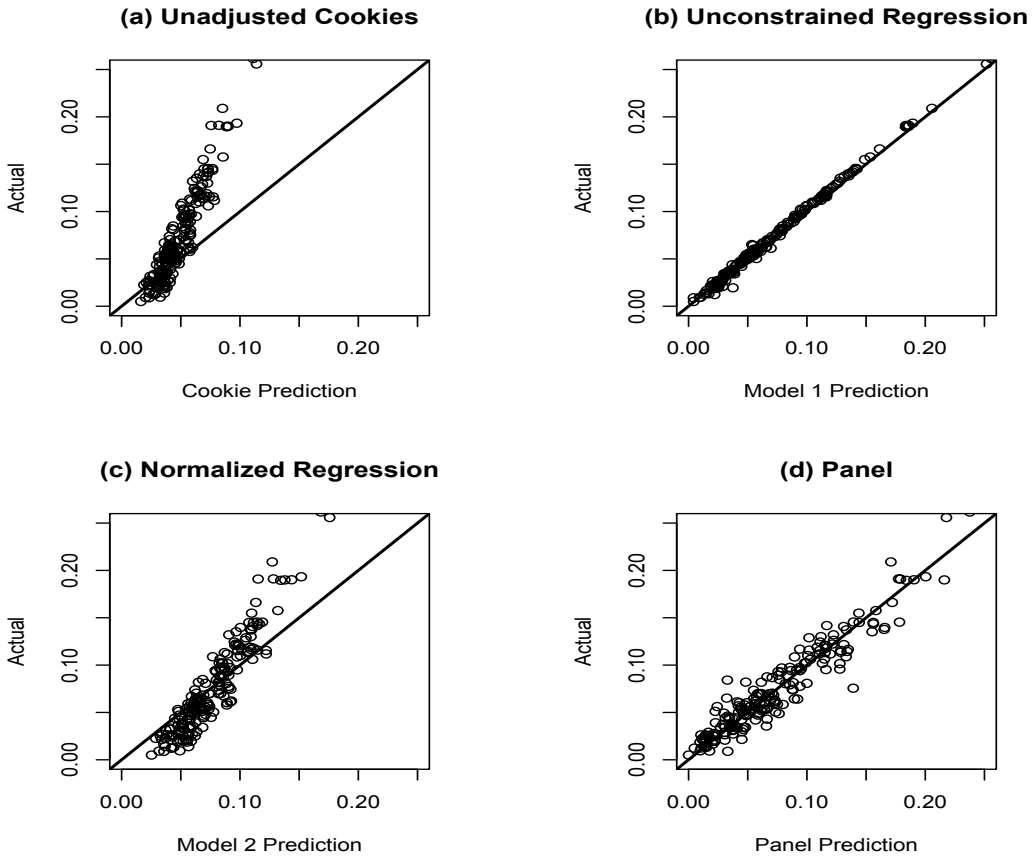

Figure 5: Relationship between people and cookies reached for a set of domains

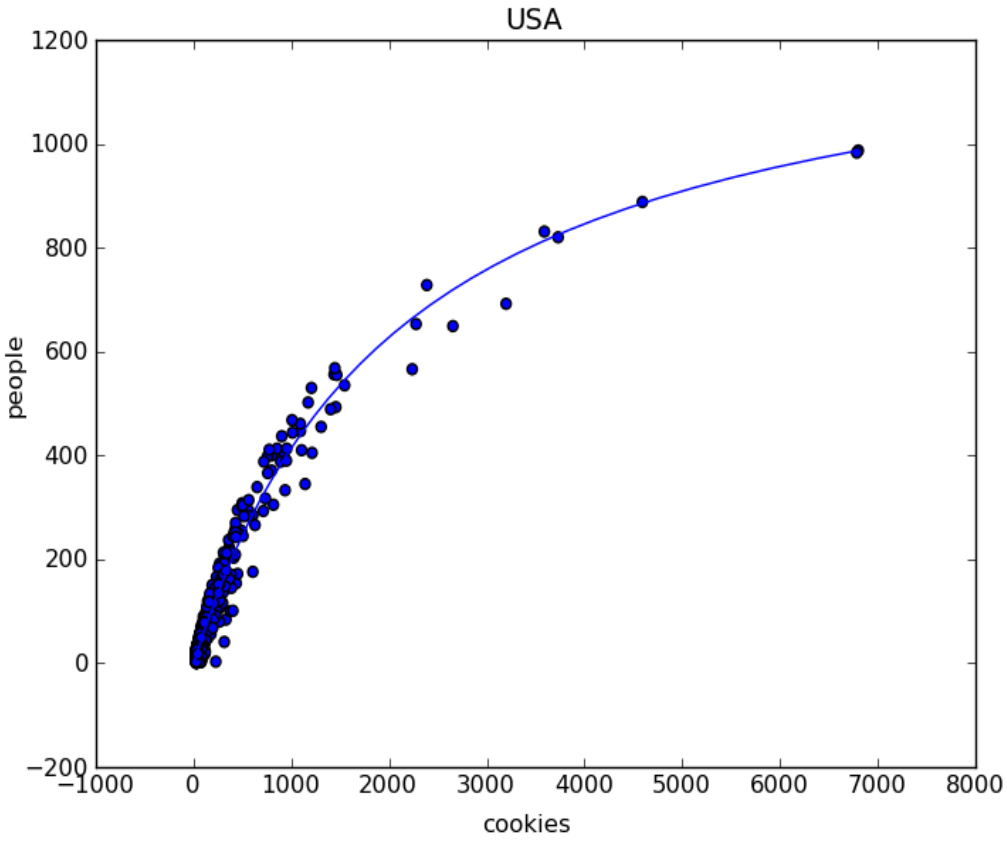

Figure 6: Relationship between people and cookies reached for a set of synthetic media plans

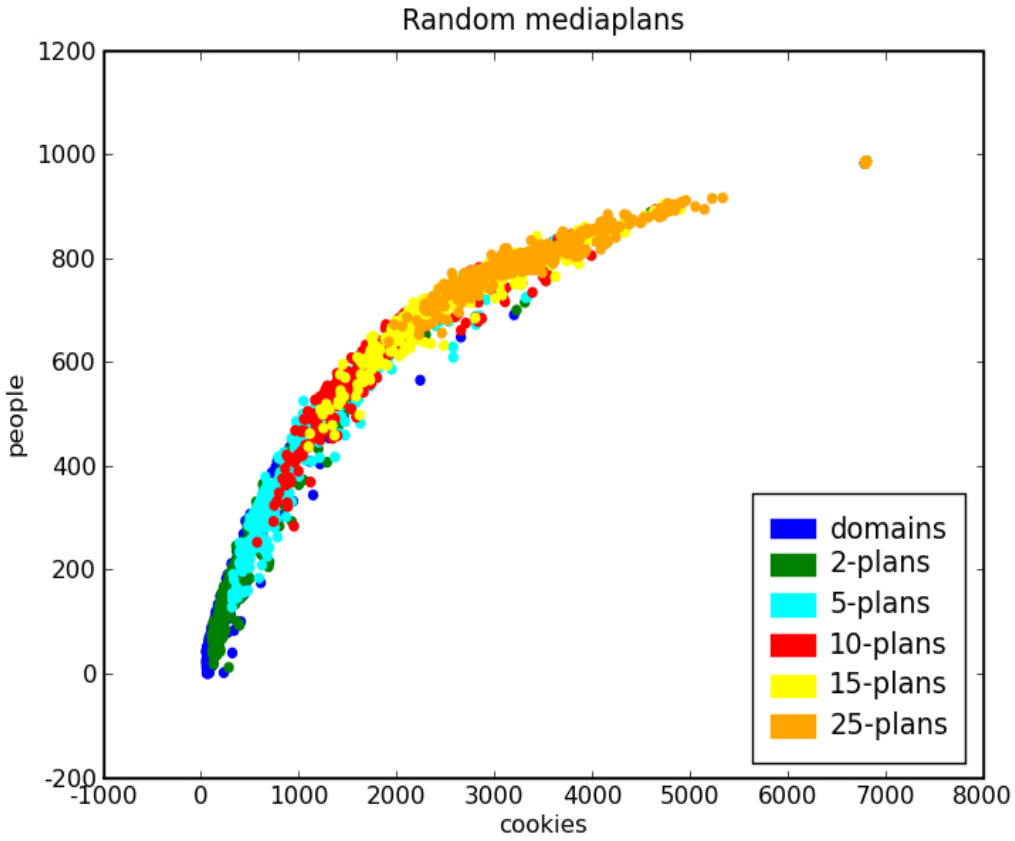

Figure 7: Overall audience composition estimates for Google Nexus campaign