# Data Enrichment and Cross Panel Imputation

Yunting Sun, Jim Koehler, Nicolas Remy, Wiesner Vos

Google Inc.

## 1 Introduction

Many empirical microeconomics studies rely on consumer panels. For example, TV and web metering panels track TV and online usage of individuals to estimate reach and frequency of a campaign: reach is the fraction of the population that has been exposed to an ads and frequency measures how often they have seen it on average. As reach and frequency are used in media planning, i.e., optimal mix between online ads and TV ads (Jin et al. (2012)), it is critical to obtain accurate reach and frequency from panel data. However, panels often suffer from underreporting, i.e., they record only a fraction of all events. Missingness can stem from various sources such as non-compliance, work usage or the use of unregistered devices (see Sudman (1964a), Sudman (1964b) for details).

To tackle missingness problem, Fader and Hardie (2000) build a Poisson model for underreported counts; Schmittlein et al. (1985) apply beta-binomial negative binomial (BBNB) model to panel data when not every purchase occasion is recorded; Goerg et al. (2015b) extend the BBNB model with a hurdle component (BBNBH) to account for excess zeros in the data-generating process of actual counts; Goerg et al. (2015a) add categorical covariates to the BBNBH model and propose a categorical missingness estimation via a penalized maximum likelihood estimator (MLE) in order to capture heterogeneity across categories, e.g., demographic groups.

Sometimes more than one panel representing the same subject population are available. These panels may use different metering technologies and are subject to varying degrees of missingness. The problem we consider here is how to do data enrichment (see Chen et al. (2013) for details) and imputation based on two panels which have similar but not identical statistical characteristics. Each panel measures a set of variables and some common variables are observed in both panels, however, there are no direct observations of the joint distribution of the two sets of variables.

Variables appear in one panel but not in the other are called target variables. For example, Figure 1 shows that panel A measures TV and publisher provided desktop impressions and panel B measures publisher provided desktop and mobile impressions. The common variables observed in both panels are desktop impressions and demographic profiles, while the target variables are TV and mobile impressions. We want to estimate a count of ad impressions across all three-screens out of the two two-screen panels. Both panels should be probabilistically recruited and calibrated to the same population. They do not have any overlap in the panelists recruited.

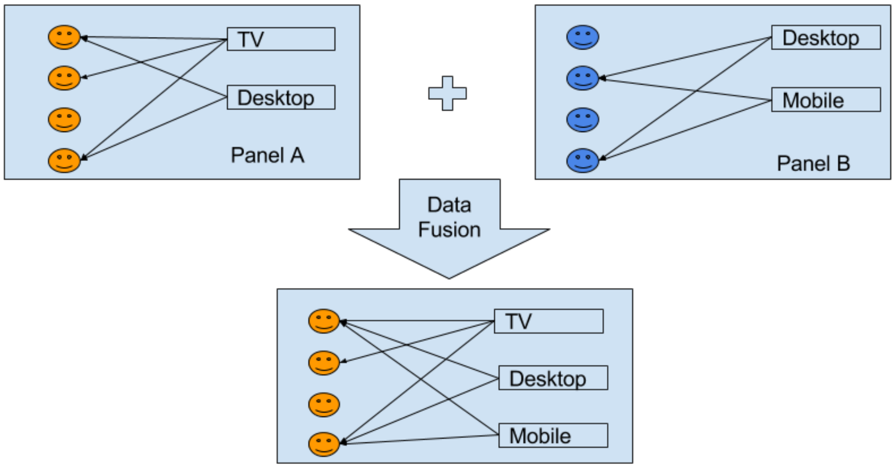

Figure 1: Data fusion of panel A and panel B. Panel A measures TV and desktop impressions from the publisher and panel B measures desktop and mobile impressions from the publisher. We want to fuse panel A and panel B to create a single panel measuring TV, desktop and mobile impressions.

The term "data fusion" was coined for this problem. Kadane (2001) view the data fusion problem as a matching problem by forming groups of observations that are "similar" as measured by their common variable values. The groups of observations can be used to impute the values of the target variables that are not observed or missing in a particular data set. Gilula et al. (2006) develop an approach that directly estimates the joint distribution of binary variables of interest that are consistent with the marginal distribution observed in each panel. As the joint distribution is not identifiable given the marginal distributions, conditional independence of target variables is usually assumed. To account for possible departure from conditional independence, they propose a multinomial model with an additional parameter allowing conditional dependence among the target variables.

In this work, we extend previous work on the BBNBH model (Section 2) and develop a joint imputation model by pooling desktop impression counts observed in both panels (Section 3). The idea is that we fix the distribution of true desktop impressions to be the same across panels but allow the missing schemes to be different. After imputing the panels, we model the distribution of mobile visits using the panel B (Section 4) and then ”fuse” the mobile impressions to panel A by assuming conditional independence (Section 5). We apply the methodology to two US panels to facilitate measurements of TV, Youtube viewership across devices in Section 6. All computations and figures were done in R.

## 2 Review of BBNBH Model

Let $N_i$ be the actual (but unobserved) number of visits by panelists i from a specific device type (e.g., desktop, mobile). The population is a mixture of people who do not visit the publisher's site at all and those who visit at least once. We assume that

$$N_i \sim \text{NBH}(N; q_0, r, q_1),$$

where $q_0 = \mathcal{P}(N_i = 0)$ indicates the fraction of people who do not visit the publisher's site at all and $N_i - 1 | N_i > 0$ follows a negative binomial distribution with the number of failures $r$ and success probability $q_1$. Let $p_i$ be the probability a visit is recorded in the panel. Assuming independence across visits, the recorded visits by panelist i, $K_i$, follows a binomial distribution $K_i | N_i \sim \text{Bin}(N_i; p_i)$. To account for heterogeneity across the population we assume $p_i \sim \text{Beta}(\mu, \phi)$, which has mean $\mu$ and precision $\phi$. Here $\mu$ represents the expected non-missing rate. Integrating out $p_i$ gives a Beta-Binomial (BB) distribution,

$$K_i | N_i \sim \text{BB}(N_i; \mu, \phi).$$

Let the negative binomial distribution parameter $\theta_{NBH} = (q_0, r, q_1)$ and beta-binomial distribution parameter $\theta_{BB} = (\mu, \phi)$ and $\theta = (\theta_{NBH}, \theta_{BB}) = (q_0, r, q_1, \mu, \phi)$. The marginal distribution for recorded visits can be derived explicitly as a function of $\theta$. The maximum likelihood estimator (MLE) for $\theta$ can be obtained by numerical optimization. More details can be found at Goerg et al.

(2015a).

# 3 Joint Imputation of Two Panels

Let $\mathbf{k^s} = \{k_1^s, \cdots, k_{P^s}^s\}$ be the number of recorded events for all $P^s$ panelists in panel $s$, $s \in \{a, b\}$. Each panelist in panel $s$ is assigned a weight $w_i^s$ based on socio-economic characteristics, indicating the number of people he represents in the entire population. As both panels are representative of the same population, $W = \sum_i w_i^a = \sum_j w_j^b$ equals the total population count (obtained from, e.g., census data).

Assume the recorded events in both panel A and B follow the BBNBH model and the corresponding parameters are $\theta^a, \theta^b$. As panel A and B are both probabilistically recruited and representing the same population, we can treat them as independent samples and assume that the distribution of true impressions is the same across panels, i.e., $\theta^a_{\text{NBH}} = \theta^b_{\text{NBH}}$. The log-likelihood of $\theta^{a,b} = (\theta^a, \theta^b)$,

$$l(\theta^{a,b}; \mathbf{x^a}, \mathbf{x^b}) = l(\theta^a; \mathbf{x^a}) + l(\theta^b; \mathbf{x^b}),$$

is the sum of log-likelihood of events in each panel and it depends on the sufficient statistics $\mathbf{x}^s = \{x_k^s | k = 0, 1, \cdots, \max(\mathbf{k}^s)\}, s \in \{a, b\}$, where $x_k^s = \sum_{i|k_i^s = k} w_i^s$ is the total weight of panelists with $k$ visits in the panel $s$. The maximum likelihood estimator (MLE)

$$\hat{\theta}^{a,b} = \arg\max \Big( l(\theta^a; \mathbf{x}^a) + l(\theta^b; \mathbf{x}^b) \Big), \text{s.t.} \quad \theta^a_{\text{NBH}} = \theta^b_{\text{NBH}}$$

can be obtained by numerical optimization.

## 3.1 Fix expected non-missing rate $\mu$

The above optimization takes place over a 7-dimensional parameter space. If we have access to internal publisher logs, we can reduce it to 5-dimensional space by fixing the expected non-missing rate in each panel a-priori by comparing panel data with publisher logs. Let $k_W^s = \sum_{i=1}^{P^s} w_i^s k_i^s, s \in \{a, b\}$ be the recorded visits in panel $s$ projected to the entire population. In the internal publisher logs, we get $n_W$ visits during the same sample period. The ratio $\mu^s_{\text{Logs}} = \frac{k_W^s}{n_W}, s \in \{a, b\}$ indicates the fraction of actual events captured by panel $s$ on average. By fixing the expected non-missing rates to be $\mu^s_{\text{Logs}}, s \in \{a, b\}$, the parameter $\theta^{a,b}$ can be obtained by a constrained MLE

$$\begin{aligned} & \hat{\theta}^{a,b} = \arg\max \left( l(\theta^a; \mathbf{x}^a) + l(\theta^b; \mathbf{x}^b) \right) \\ \text{s.t.} \quad & \theta^a_{\text{NBH}} = \theta^b_{\text{NBH}} \\ & \mu^a = \mu^a_{\text{Logs}} \\ & \mu^b = \mu^b_{\text{Logs}}. \end{aligned}$$

In applications we found that the additional linear constraint on $\mu^a, \mu^b$ gives more stable estimates.

### 3.2 Demographic-dependent estimation

As advertisers use panels to measure viewing behavior of specific target audience, e.g., young males, demographic-specific inference is important. However the missing scheme may be different across demographic groups. Relying on the same $\hat{\theta}^{a,b}$ for all demographic groups does not provide good demographic-specific inference. We thus extend the model with categorical parameters $\Theta^{a,b(1:G)} = \left(\Theta^{a,b(1)}, \cdots, \Theta^{a,b(G)}\right)$, one for each of G exhaustive demographic subgroups $D_1, \cdots, D_G$ and $\Theta^{a,b(g)} = \left(\theta^{a(g)}, \theta^{b(g)}\right), g = 1, \cdots, G$, where $\theta^{a(g)}, \theta^{b(g)}$ are the BBNBH parameters for demographic group $D_g$ in panel A and B, respectively. We can estimate the overall expected non-missing rate in each panel using the internal publisher logs $\mu^s_{\text{Logs}} = \frac{k^s_W}{n_W}, s \in \{a, b\}$ but we can not estimate the expected non-missing rate for each demographic group of each panel as most internal publisher logs do not contain reliable demographic labels for viewers. If we have mandatory login with trusted demographic labels, we might set the non-missing rate of $D_g$ of panel $s$ $\mu^{s(g)} = \mu^{s(g)}_{\text{Logs}} = \frac{k^{s(g)}_W}{n^{(g)}_W}$ directly, where $k^{s(g)}_W$ and $n^{(g)}_W$ are recorded visits in the panel $s$ and in the logs restricted to demographic group $D_g$. By fixing the overall expected non-missing rates to be $\mu^s_{\text{Logs}}, s \in \{a, b\}$, $\Theta^{a,b(1:G)}$ can be obtained by a constrained MLE

$$\begin{aligned}\hat{\Theta}^{a,b(1:G)} &= \arg\max \sum_{g=1}^{G} \Big( l(\theta^{a(g)}; \mathbf{x}^{a(g)}) + l(\theta^{b(g)}; \mathbf{x}^{b(g)}) \Big) \\ \text{s.t.} \quad & \theta_{\text{NBH}}^{a(g)} = \theta_{\text{NBH}}^{b(g)} \quad g = 1, \cdots, G \\ & \sum_{g=1}^{G} \frac{k_W^{a(g)}}{\mu^{a(g)}} = \frac{k_W^a}{\mu_{\text{Logs}}^a} \\ & \sum_{g=1}^{G} \frac{k_W^{b(g)}}{\mu^{b(g)}} = \frac{k_W^b}{\mu_{\text{Logs}}^b},\end{aligned}$$

where $\mathbf{x}^{s(g)} = \left\{ x_k^{s(g)} | k = 0, 1, \cdots, \max(\mathbf{k}^{s(g)}) \right\}$, $s \in \{a, b\}$ and $x_k^{s(g)}$ is the total weight of panelists with $k$ visits in demographic group $D_g$ of panel $s$. It is an optimization with $7G - 2$ degrees of freedom, which is numerically challenging when $G$ is large. Hence, we develop an alternative method which can be parallelized easily and reduces the degrees of freedom to $6G$.

We first combine the panelists from panel A and B as if they were coming from the same enlarged panel. The sufficient statistics for demographic group $D_g$ of the combined panel is

$$\mathbf{x}^{(g)} = \left\{ x_k^{(g)} = x_k^{a(g)} + x_k^{b(g)} | k = 0, 1, \cdots, \max(\mathbf{k}^{a(g)}, \mathbf{k}^{b(g)}) \right\}.$$

Let the BBNBH parameter for demographic group $D_g$ of the combined panel be $\Theta^{(g)}$ and $\Theta^{(1:G)} = \left(\Theta^{(1)}, \cdots, \Theta^{(G)}\right)$. By fixing the expected overall non-missing rate of the combined panel to be $\mu_{\text{Logs}} = \frac{k_W^a + k_W^b}{2 \times n_W}$, $\Theta^{(1:G)}$ can be obtained by a constrained MLE

$$\begin{aligned}\hat{\Theta}^{(1:G)} &= \arg\max \sum_{g=1}^{G} l(\Theta^{(g)}; \mathbf{x}^{(g)}) \\ \text{s.t.} \sum_{g=1}^{G} \frac{k_W^{a(g)} + k_W^{b(g)}}{\mu^{(g)}} &= \frac{k_W^a + k_W^b}{\mu_{\text{Logs}}}.\end{aligned}$$

We take the approach of iterative exact-constraint estimator from Section 3.3.2 of Goerg et al. (2015a) for the optimization. The algorithm alternates between the estimation of $\mu^{(1:G)}$ and the rest of the parameters $\Theta_{-\mu}^{(1:G)}$. By fixing the expected non-missing rate in demographic group $D_g$ to be $\hat{\mu}^{(g)}$, we can estimate the panel specific BBNBH parameters $\Theta^{a,b(g)} = \left(\theta^{a(g)}, \theta^{b(g)}\right)$ by another constrained MLE

$$\begin{aligned}
&\hat{\Theta}^{a,b(g)} = \arg\max \left( l(\theta^{a(g)}; \mathbf{x}^{a(g)}) + l(\theta^{b(g)}; \mathbf{x}^{b(g)}) \right) \\
\text{s.t.} \quad &\theta^{a(g)}_{\text{NBH}} = \theta^{b(g)}_{\text{NBH}} \\
&\frac{k^{a(g)}_W}{\mu^{a(g)}} + \frac{k^{b(g)}_W}{\mu^{b(g)}} = \frac{k^{a(g)}_W + k^{b(g)}_W}{\hat{\mu}^{(g)}}.
\end{aligned}$$

This step can be parallelized across G demographic groups.

# 4 Modeling mobile impressions

The mobile impression counts in panel B can be imputed the same way as desktop in Goerg et al. (2015a) and only for panelists with mobile devices. Panelists without mobile devices are assigned zero mobile impression. Panelists with mobile devices may or may not have a different weight from that of panelists with desktops. It would not be an issue as imputation are done separately for desktop and mobile visits. Let $N^a_d, N^b_d$ be the actual desktop visits in panel A and B and $K^a_d, K^b_d$ be the recorded visits. Let $N^b_m$ be the actual mobile visits in panel B and $K^b_m$ be the recorded visits. Assuming that desktop and mobile missing schemes are independent given recorded cross-device visits, we can simulate M replicates of paired $N^b_d$ and $N^b_m$ for each panelist from imputation model inferred distributions $N^b_d|K^b_d$ and $N^b_m|K^b_m$ independently. Each simulated replicate becomes a new panelist with weight shrunken proportionally. Let $r(i), i = 1, \cdots, P^b M$ be the index of the original panelist from whom the $i$th replicate comes from and then $\tilde{w}^b_i = \frac{w^b_{r(i)}}{M}$ is the demographic weight of the $i$th replicate.

Let $Z$ be the number of desktop visits and demographic features of a panelist. For example, $Z = (N_d, \text{age, gender})$. The demographic features are not limited to age and gender and could be generalized to include education and income as well if such data is available. In order to model the conditional distribution of $N_m|Z$, we assume that it follows a negative binomial hurdle model:

$$N_m|Z \sim \text{NBH}(N; p_0(Z), \tau(Z), \eta),$$

- $p_0(Z) = \exp(Z^T\gamma)/(1 + exp(Z^T\gamma))$ is the probability of not being reached by the publisher on mobile devices
- $\tau(Z) = \exp(Z^T\beta)$ is the mean of the negative binomial distribution
- $\eta$ is the dispersion of the negative binomial distribution.

Parameters $(\gamma, \beta, \eta)$ can be obtained by maximizing the log-likelihood function

$$l((\gamma,\beta,\eta)\,;\mathbf{z^b},\mathbf{n_m^b}) = \sum_{i=1}^{P^bM} \tilde{w}_i^b \log P(N_m = n_{m(i)}^b | Z = z_i^b; (\gamma,\beta,\eta)),$$

where $\mathbf{z^b} = \left\{z_1^b, z_2^b, \cdots, z_{P^bM}^b\right\}$ are the number of desktop visits and demographic features and $\mathbf{n_m^b} = \left\{n_{m(1)}^b, \cdots, n_{m(P^bM)}^b\right\}$ are the number of mobile visits of $P^bM$ replicated panelists from the panel B.

### 4.1 Right-truncation of impressions

We have found empirically that panel observations could have heavy tails, that is, some panelists have an extremely large number of recorded visits. Figure 4 illustrates the heavy tail property of the empirical distribution of online visits on desktop and mobile devices. To make the estimation more robust to these extremes, we right-truncate the desktop impressions at some $n_d = n_q^d$:

$$N_d = \min(N_d, n_q^d).$$

Besides, we also right-truncate the summation in the log-likelihood at some $n_m = n_q^m$, and add the cumulative probability for the event $N_m > n_m$. The approximation of the log-likelihood function can be written as:

$$\begin{aligned} l((\gamma,\beta,\eta)\,;\mathbf{z^b},\mathbf{n_m^b})_{\text{trunc}} = & \sum_{i|n_{m(i)}^b <= n_q^m} \tilde{w}_i^b \log P(N_m = n_{m(i)}^b | Z = z_i^b; (\gamma,\beta,\eta)) \\ & + \sum_{i|n_{m(i)}^b > n_q^m} \tilde{w}_i^b \log P(N_m > n_q^m | Z = z_i^b; (\gamma,\beta,\eta)). \end{aligned}$$

As $N_m, N_d$ grows with length of the time period, it is not possible to propose a universal value for $n_q^d, n_q^m$. Thus we choose $n_q^d, n_q^m$ based on the empirical quantile. We found that using sample quantile $q = 0.995$ works well in practice.

### 4.2 Demographic-dependent estimation

The right-truncated negative binomial hurdle (RNBH) regression model can be applied independently for each demographic group to account for heterogeneity and interactions between predictors. The demographic groups may or may not be the same as the ones used in section 3.2.

### 4.3 Model inferred marginal distribution of mobile impressions

With the estimated distribution $N_m|N_d$ from the RNBH model and $N_d|K_d$ from the extended BBNBH imputation, we can calculate the model implied distribution of actual mobile visits given recorded desktop visits $N_m|K_d$ for panel $s \in \{a, b\}$,

$$\begin{aligned} P^s(N_m = r|K_d = h) &= \sum_{l \geq h} P^s(N_m = r, N_d = l|K_d = h) \\ &= \sum_{l \geq h} P(N_m = r|N_d = l) P^s(N_d = l|K_d = h). \end{aligned}$$

Panel A and B are sharing $N_m|N_d$ but have separate $N_d|K_d$ distributions. The model implied marginal distribution of mobile impressions for panel $s$ can thus be expressed as

$$P^s(N_m = r) = \frac{\sum_{i=1}^{P^s} w_i^s P^s(N_m = r|K_d = k_{d(i)}^s)}{\sum_{i=1}^{P^s} w_i^s},$$

where $k_{d(i)}^s$ is the recorded desktop visits for panelist $i$ in panel $s$.

## 5 Data Fusion of Two Panels

Let $N_m$ be the mobile visits, $N_d$ be the desktop visits and $N_t$ be the TV visits. As panel A captures TV and desktop visits, we can obtain $p(N_t, N_d)$ after imputing the recorded desktop visits. Let $n_{t(i)}^a$ be the TV visits and $k_{d(i)}^a$ be the recorded desktop visits for panelist i in panel A. The joint distribution of TV and desktop visits can be calculated as

$$P(N_t = r, N_d = l|\mathbf{n_t^a}, \mathbf{k_d^a}) = \frac{\sum_{i|n^a_{t(i)}=r} w_i^a P(N_d = l|K_d = k^a_{d(i)})}{\sum_{i=1}^{P^a} w_i^a}.$$

As panel B captures desktop and mobile visits, we can obtain $p(N_m|N_d)$ from section 4. However, the joint distribution $p(N_d, N_m, N_t)$ is not identifiable given the two marginal distributions. Hence we have to assume conditional independence,

$$p(N_m, N_t|N_d) = p(N_m|N_d)p(N_t|N_d).$$

And it implies

$$p(N_m, N_t, N_d) = p(N_t, N_d)p(N_m|N_t, N_d) = p(N_t, N_d)p(N_m|N_d).$$

The joint distribution of TV, desktop and mobile visits for panel A can be written as

$$P(N_t = r, N_d = l, N_m = h|\mathbf{n_t^a}, \mathbf{k_d^a}) = P(N_t = r, N_d = l|\mathbf{n_t^a}, \mathbf{k_d^a})P(N_m = h|N_d = l).$$

The particular quantity of interest is the incremental reach of the publisher across devices versus TV $P(N_d + N_m > 0, N_t = 0)$. In practice, for each panelist in the panel A we simulate actual desktop impression counts $n^a_{d(i)}$ from $N_d|K_d = k^a_{d(i)}$ and actual mobile impression counts $n^a_{m(i)}$ from $N_m|N_d = n^a_{d(i)}$. As long as the simulated panel size is large enough, the incremental reach of the publisher over TV can be estimated by

$$P(N_d + N_m > 0, N_t = 0) = \frac{\sum_{i|n^a_{m(i)}+n^a_{d(i)}>0, n^a_{t(i)}=0} w_i^a}{\sum_i w_i^a}.$$

# 6 Case Study

We now illustrate the cross panel imputation methodology on data from two panels in the United States. Panel A monitors TV and Youtube desktop watchpage usage and panel B monitors Youtube desktop and mobile watchpage usage for the period from 2014-12-01 to 2015-01-01 (31 days). We know the mobile device ownership in the panel B but not in the panel A. After data cleaning, we remain with 17352 panelists in panel A and 7728 panelists in panel B both representing the online population of the United States. We know the age and gender of each panelist and divide the panelists into 6 demographic groups, 0-35Female, 0-35Male, 36-50Female, 36-50Male, 50+Female and 50+Male.

## 6.1 Joint imputation of desktop impressions

Figure 2 shows the empirical frequency of recorded desktop impressions in the two panels. The proportion of zero visits is quite high, around 80% for panel A and 60% for panel B. As both panels are probabilistically recruited, the panel B does a better job of capturing desktop impressions than panel A in general and thus a higher non-missing rate across demographic groups.

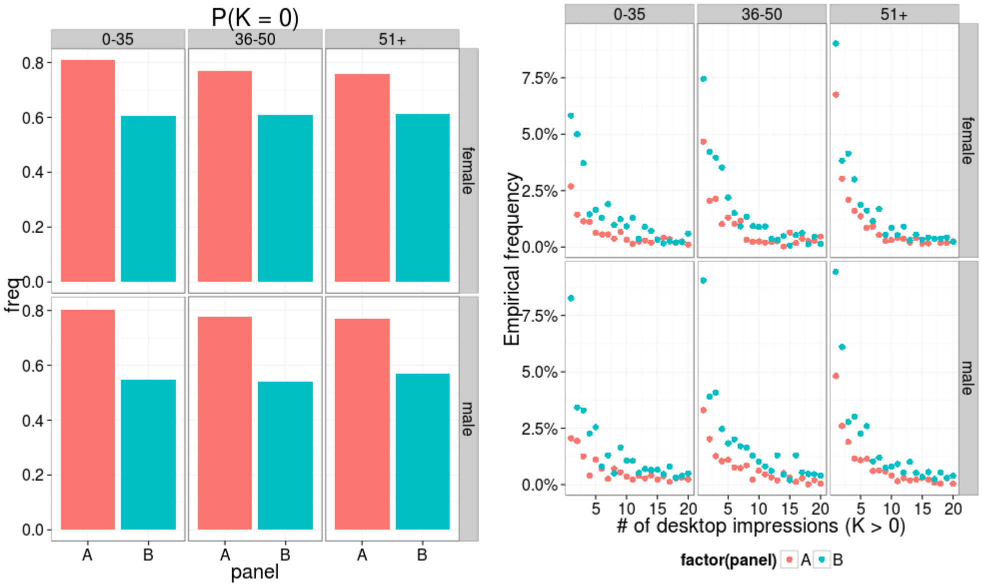

Figure 2: Empirical frequency of desktop Youtube impressions in panel A and panel B. Left figure shows the proportion of no visits and the right figure shows the fraction of panelists having $k|k = 1, 2, \cdots$ visits in each panel by demographic groups.

Our internal Youtube logs show that the combined panel has a non-missing rate of $\hat{\mu}_{\text{Logs}} = 0.26$ for Youtube desktop visits. We carry out the demographic specific joint imputation for the desktop impressions of the two panels. Figure 3 shows the estimated model parameters for each demographic group. As we impose the constraint that the actual impression distributions are the same for panel A and B, the estimated parameters $q_0, q_1, r$ are the same across panels. The estimated non-missing rates in panel B is higher than that in panel A in all demographic groups. Consider females aged between 36 and 50, the estimated hurdle probability of $\hat{q}_0 = 0.46$ suggests that the excess zeros in the panels are the result of missingness and a high probability of not visiting a Youtube watchpage at all.

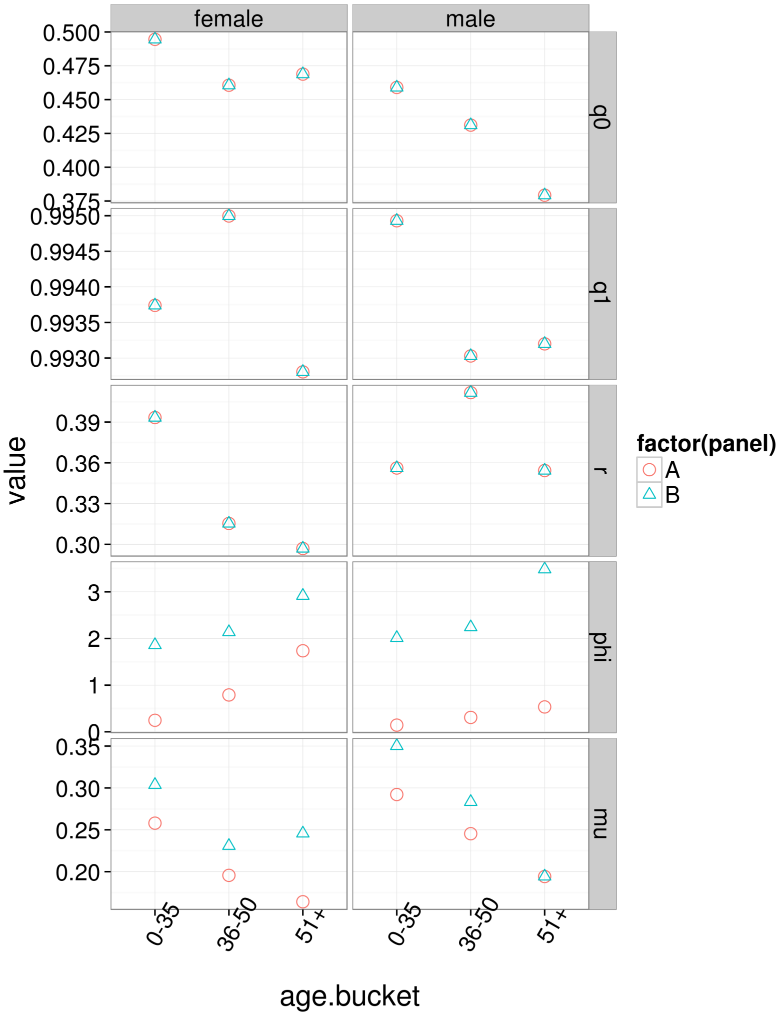

Figure 3: Estimated imputation parameters for desktop Youtube impressions across panels.

## 6.2 Mobile Imputation

Figure 4 shows the empirical cumulative distribution functions of mobile and desktop visits in the panel B. Some panelists have more than 1000 desktop visits during the 31-day sample period. The 0.995 sample quantile of desktop visits and mobile visits are 235 and 135, respectively. We fit a right-truncated negative binomial hurdle (RNBH) regression of mobile visits over desktop visits and age for each demographic group.

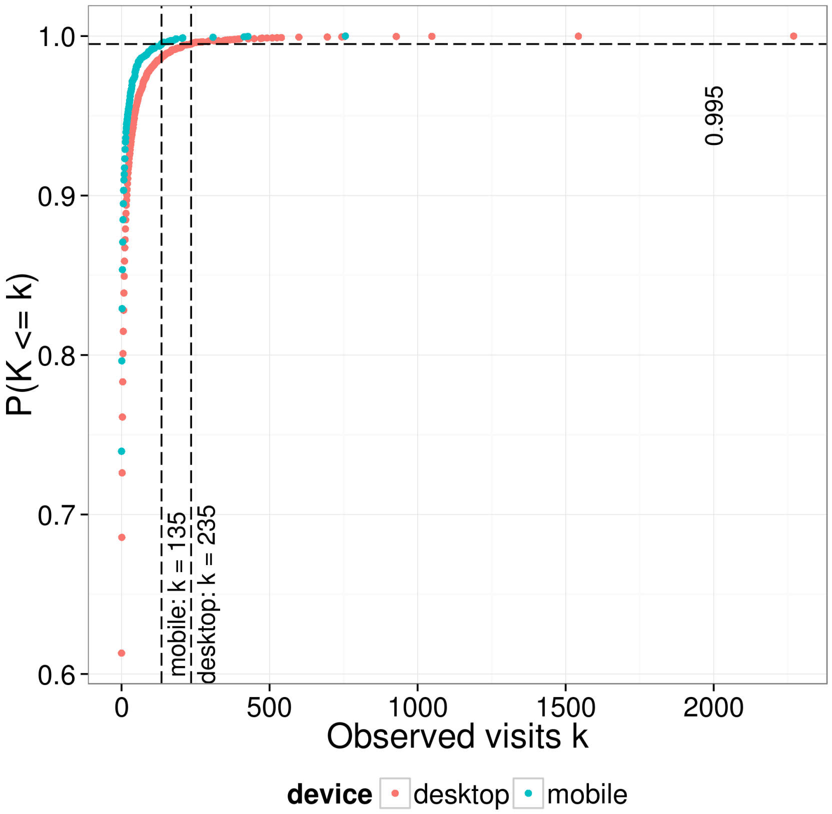

Figure 4: Empirical CDF of desktop and mobile recorded visits for panel B.

In order to evaluate the performance of the RNBH regression model, we randomly split panel B into two equal-sized sub-panels $B1$ and $B2$. The plan is to build the model using $B1$ and evaluate its performance using $B2$ where we know the ground truth.

We first perform joint imputation (section 3) on desktop impressions and mobile impressions over sub-panels $B1$ and $B2$ so that the distribution of actual desktop and mobile impressions per panelist are the same across the sub-panels. The joint imputation for mobile impressions is only performed over panelists who use mobile devices. Panelists without mobile devices get zero imputed mobile visits and panelists with mobile devices get imputed mobile visits following $(N_m|K_m, \text{age, gender})$, where $K_m$ is the recorded mobile visits.

We then fit a RNBH regression for mobile visits using all the panelists from the sub-panel $B1$ and then predict the mobile visits for each panelist in $B2$, which follows the distribution $(N_m|K_d, \text{age, gender})$, where $K_d$ is the recorded desktop visits. This model is agnostic to whether a panelist uses mobile devices at all. On the other hand, each mobile device owner in the sub-panel

$B2$ has imputed mobile visits $(N_m|K_m, \text{age, gender})$, which serve as the ground truth for evaluating the model. Figure 5 shows the cumulative distribution functions for the RNBH model (section 4.3) and the extended BBNBH model inferred marginal mobile visits in the sub-panel $B2$. The two distributions are close in general.

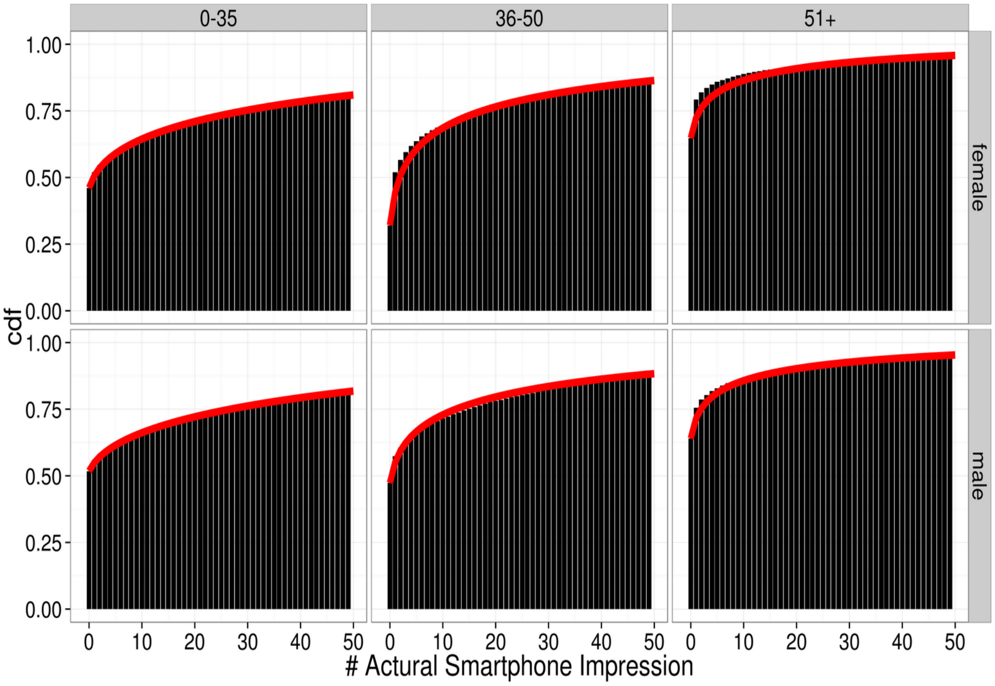

Figure 5: RNBH regression model inferred and imputed marginal mobile visits cumulative distribution functions. The red curve is the cdf based on RNBH regression and the black vertical bar is the cdf from the extended BBNBH imputation of recorded mobile visits.

We also study the model performance at the individual panelist level. For each panelist, we compute his model inferred mobile visits expectation $\xi = \mathcal{E}(N_m|K_d, \text{age, gender})$ as well as the imputed mobile visits expectation $\zeta = \mathcal{E}(N_m|K_m, \text{age, gender})$. The relative error rate is

$$\text{Relative Error} = \frac{\sum_i w_i (\xi_i - \zeta_i)^2}{\sum_i w_i \zeta_i^2},$$

where the sum is over all the panelists in the sub-panel $B2$. Figure 6 shows the relative errors across demographic buckets. The female36-50 bucket has the lowest relative error and the female50+ bucket has the highest relative error. The overall relative error rate is 76% and the weighted pearson correlation between imputed and RNBH model fitted expectation is 0.23 over all the panelists in

$B2$.

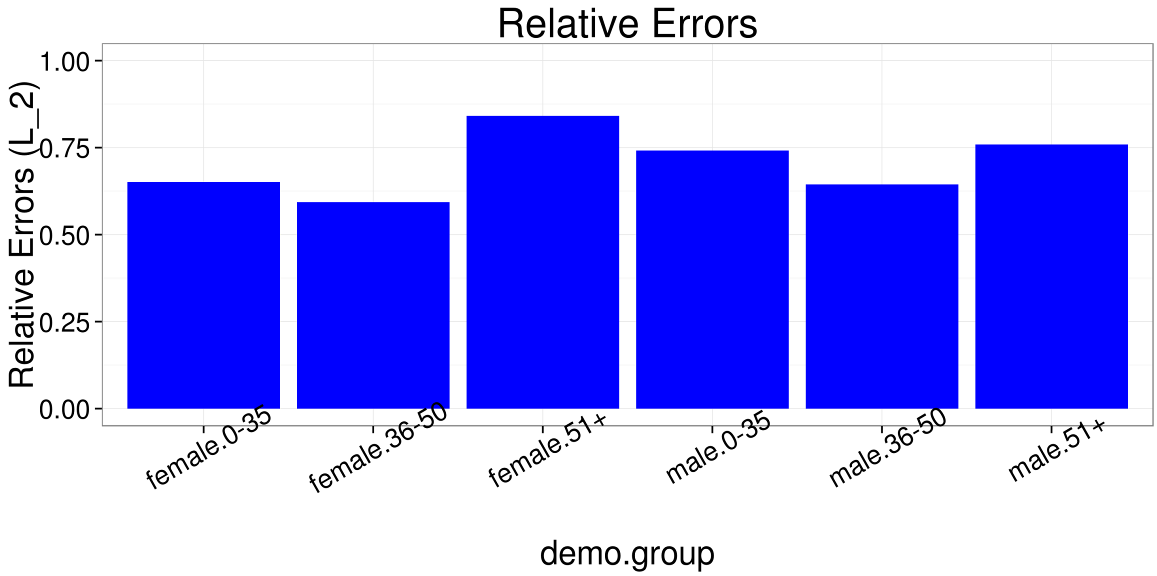

Figure 6: Relative error rates across demographic buckets

The above RNBH regression model assumes that mobile device ownership is unknown in the sub-panel $B2$. However, in practice, we may collect panelist device ownership through panel surveys. In that case, we would assign zero mobile visits to panelist without mobile devices in the sub-panel $B2$ directly. We then fit the RNBH regression model over panelists with mobile devices in $B1$ and apply the model only to mobile device users in $B2$. With the knowledge of mobile device ownership, the overall relative error rate decreases to 66% and the weighted pearson correlation between imputed and RNBH model fitted expectation increases to 0.42 over all the panelists in $B2$.

### 6.3 Data Fusion

We fit the RNBH model using the panel B and then ”fuse” mobile impressions to the panel A. Figure 7 compares the ”fused” Youtube mobile 1+ reach in the panel A with the imputed Youtube mobile 1+ reach in the panel B. Figure 8 compares the ”fused” Youtube cross-device 1+ reach in the panel A with the imputed Youtube cross-device 1+ reach in the panel B. The reach curves are close to each other in general. Although the relative error of mobile visits for individual panelist is high, the model does a good job in capturing the marginal distribution of mobile visits and cross-device visits and thus successfully recovers the reach curves in the panel A.

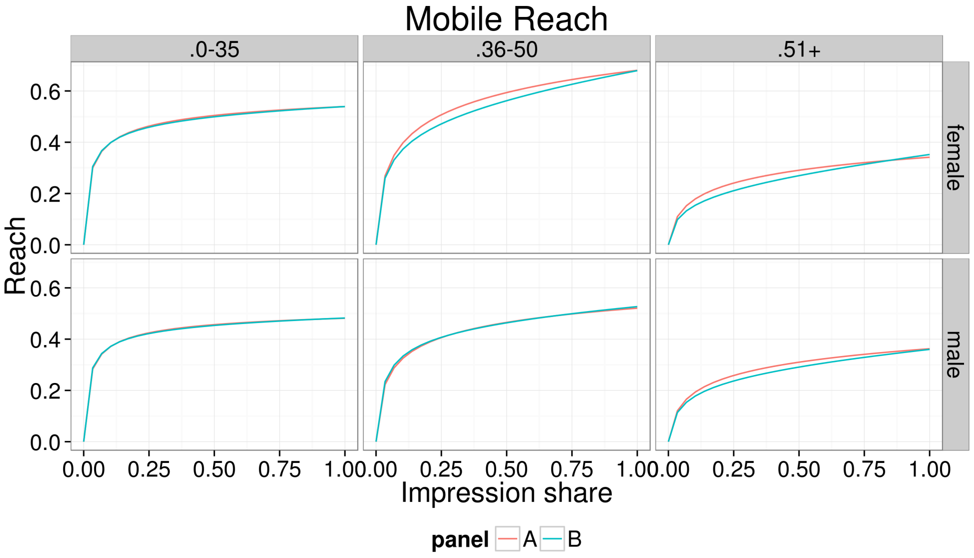

Figure 7: Youtube mobile 1+ reach curves across panels.

# 7 Discussion

Motivated by the applied problem of data enrichment using two panels representing the same population, we extend the BBNBH model and propose a constrained likelihood approach to obtain imputation estimates. The methodology is used to estimate incremental cross-device reach of Youtube versus TV in the United States.

In future work, we aim to extend the methodology to combine more than two panels with similar but not identical characteristics and introduce additional parameters to account for departure from conditional independence. Another direction is to combine biased panels with calibrated panels for cross-device measurement.

# Acknowledgement

We want to thank Georg Goerg, Yuxue Jin, Xiaojing Wang, Tristan Launay, Diane Lambert, Penny Chu, Tony Fagan, Raimundo Mirisola, Andras Orban and Mikko Sysikaski for constructive discussions and feedbacks.

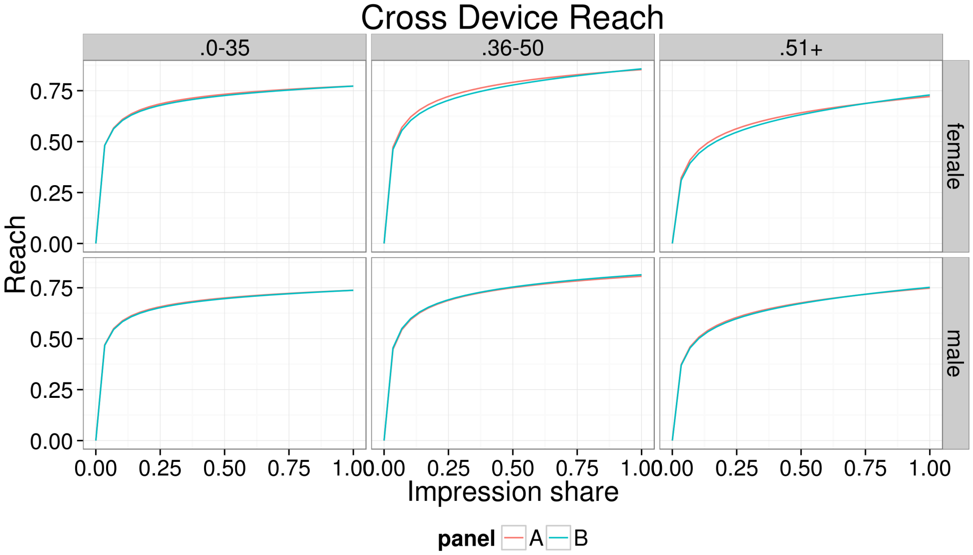

Figure 8: Youtube cross-device 1+ reach curves across panels.

## References

Chen, A., Koehler, J., Owen, A., Remy, N., and Shi, M. (2013). Data enrichment for incremental reach estimation.

Fader, P. and Hardie, B. (2000). A note on modelling underreported poisson counts. *Journal of Applied Statistics*, pages 27(8):953–964.

Gilula, Z., McCulloch, R., and Rossi, P. (2006). A direct approach to data fusion. *Journal of Marketing Research*, 43:73–83.

Goerg, G. M., Jin, Y., Remy, N., and Koehler, J. (2015a). How many millenials visit youtube? estimating unobserved events from incomplete panel data conditioned on demographic covariates. Technical report, Google Inc.

Goerg, G. M., Jin, Y., Remy, N., and Koehler, J. (2015b). How many people visit youtube? imputing missing events in panels with excess zeros. *Proceedings of 30th International Workshop on Statistical Modelling, Linz, Austria.*

Jin, Y., Shobowale, S., Koehler, J., and Case, H. (2012). The incremental reach and cost efficiency of online video ads over tv ads. Technical report, Google Inc. `http://research.google.com/pubs/pub40426.html`.

Kadane, J. (2001). Some statistical problems in matching data files. *Journal of Official Statistics*, 17:423–433.

Schmittlein, D. C., Bemmaor, A. C., and Morrison, D. G. (1985). Why does the nbd model work? robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Science*, pages 4(3):255–266.

Sudman, S. (1964a). On the accuracy of recording of consumer panels: I. *Journal of Marketing Research*, 1:14–20.

Sudman, S. (1964b). On the accuracy of recording of consumer panels: II. *Journal of Marketing Research*, 1:69–83.