Efficient Large-Scale Media Selection Optimization for Online Display Advertising: Web Appendix

Courtney Paulson, Lan Luo, and Gareth M. James

March 23, 2018

Appendix A Comparing Proposed and Danaher Methods

Danaher et al. (2010) provides one state-of-the-art method for optimal budget allocation of Internet display ads. Danaher considers a traditional advertising setup, in which costs are fixed and a small number of websites are considered for advertising purposes. A basic premise of this method is that the number of web pages viewed by individuals at websites (denoted as a n by p matrix Z in our context) can be characterized by a multivariate negative binomial distribution (referred to as MNBD hereafter). In particular Danaher's method models the full exposure distribution as follows:

$$\begin{aligned} & P(X_1 = x_1, \ldots, X_p = x_p) \qquad \text{(A1)} \\ = & \left(\prod_{j=1}^{p} P(X_j = x_j | s_j, r_j, \alpha_j, t_j)\right) \left[1 + \sum_{j<k} \omega_{j,k}\phi_j(x_j)\phi_k(x_k) + \sum_{j<k<l} \omega_{j,k,l}\phi_j(x_j)\phi_k(x_k)\phi_l(x_l)\right] \end{aligned}$$

where $P(X_j = x_j | s_j, r_j, \alpha_j, t_j) = \binom{x_j + r_j - 1}{x_j} \left(\frac{\alpha_j}{\alpha_j + t_j s_j}\right)^{r_j} \left(\frac{t_j}{\alpha_j + t_j s_j}\right)^{x_j}$, i.e., X_j is modeled using a Negative Binomial distribution, the $\phi_j(x_j)$ terms have a given functional form, $0 \le s_j \le 1$ corresponds to the share of impressions purchased, t_j represents a time interval, and $r_j, \alpha_j, \omega_{j,k}, \omega_{j,k,l}$ are all parameters that are estimated from $\mathbf{Z}$.

By comparison, our approach models the full exposure distribution as follows:

$$\begin{aligned} P(X_1 = x_1, \dots, X_p = x_p|\mathbf{c}) &= \int_{\mathbf{z}} P(X_1 = x_1, \dots, X_p = x_p|\mathbf{c}, \mathbf{Z} = \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} \\ &= \int_{\mathbf{z}} \prod_{j=1}^{n} P(X_j = x_j|c_j, Z_j = z_j) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} \\ &= \int_{\mathbf{z}} \prod_{j=1}^{n} \left(\frac{e^{-\gamma_j}\gamma_j^{x_j}}{x_j!} \right) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} \\ &\approx \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{e^{-\gamma_{ij}}\gamma_{ij}^{x_{ij}}}{x_{ij}!} \right) \qquad \text{(A2)} \end{aligned}$$

where $\gamma_{ij} = s_j(c_j)z_{ij}$. An important difference between our approach and that of Danaher et al.'s (2010) method is that we do not attempt to model a parametric distribution for $f_{\mathbf{Z}}(\mathbf{z})$ but instead use a sample of individual's page views among our p websites, which provides an empirical approximation for $f_{\mathbf{Z}}(\mathbf{z})$.

When one wishes to model reach, both methods will then set $P(X_1 = 0, \dots, X_p = 0)$. However, a key distinction between Equation A1 and Equation A2 is that in the latter case the expression simplifies to $\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j \gamma_{ij}}$, while in the former case no such simplification occurs. This difference between the two methods is the main explanation for why our method is still computationally feasible for up to thousands of websites, while Danaher's method is more suitable for much smaller numbers of websites.

In this appendix, we compare performance of the proposed method with that of Danaher's method in small-scale campaigns assuming that the CPM to advertise at each website is known and fixed. It is also worth noting that, under the fixed costs assumption, the approach we outline below (as a slight variation of the proposed method discussed in the main paper) is directly applicable to large-scale nonprogrammatic and programmatic direct display ad campaigns.

Let c_j represent the cost to purchase 1000 impressions. Using the proposed method, we now solve for optimal budget spent at website j, w_j, for $j = 1, \dots, p$. The total number of impressions purchased will then be given by $1000w_j/c_j$, where c_j is fixed. It can be shown that our objective function under this setup becomes:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} e^{-\gamma_i} \quad \text{subject to} \quad \sum_j w_j \leq B \quad \text{and} \quad w_j \geq 0, \quad j = 1, \dots, p, \qquad \text{(A3)}$$

where $\gamma_{ij} = \frac{z_{ij}}{\tau_j \frac{c_j}{1000}} w_j$. Then the algorithm can be run in a similar fashion as in the Methodology Section, but now with θ_{ij} and η_{ij} in Equation 8 as $\theta_{ij} = \frac{z_{ij}}{\tau_j c_j/1000}$ and $\eta_{ij} = \theta_{ij}^2$.

Note that while Danaher et al. (2010) optimizes over *share of impressions* at a given website (what we call s_j) and we optimize over total budget spent w_j here, there is a one-to-one correspondence between s_j and w_j as follows: $s_j = \frac{1000 w_j}{\tau_j c_j}$. Therefore, we use this correspondence to compare the performances of these two methods below. Further, because Danaher et al.'s method assumes a MNBD in the web page matrix, we develop two settings to test the methods: (1) a simulated data example, where the data is simulated from the assumed MNBD distribution, and (2) a real data setting, where the data is taken from the comScore data used in Empirical Investigation Section.

A.1 Comparison using Data Simulated from MNBD

To examine how our method performs under the basic premise of Danaher's approach, we first generate the Internet usage matrix Z with 5000 rows (users) and 7 columns (websites), based on a MNBD with α_j and r_j, $j = 1, ..., 7$, the usual parameters associated with a MNBD, and $\omega_{j,j'}$, a set of correlation parameters denoting the correlation coefficient in viewership between websites j and j'. To make our simulation as realistic as possible, we establish α_j, r_j, and $\omega_{j,j'}$ as the values from the seven most-visited websites from the December 2011 comScore data. We also use the CPMs provided by comScore's 2010 Media Metrix (Lipsman, 2010) in this stimulation. Given that Danaher's method models page views as a MNBD while we directly use empirically observed page views in the Z matrix without assuming any underlying distribution, each method has its own definition of the reach function. Thus, to ensure an unbiased comparison of the two methods, we report the results using a neutral reach definition. Both methods share a common metric, the probability of being served the ad at the jth website is s_j. Thus we can naturally define a binomial reach function, $1 - \frac{1}{n} \sum\limits_{i=1}^{n} \prod_j (1 - s_j)^{z_{ij}}$, for comparison purposes. In what follows, we use this definition of reach in model comparisons.

Figure A1 shows the reach curves for the average reach estimate for our approach (Proposed Poisson), the Danaher estimate[1], and Danaher's method with all interaction terms

[1] Since Danaher's objective function is highly nonconvex, it can find local optima during optimization. Consequently, we run the optimization with several initialization points and choose the results with the highest reach for comparisons.

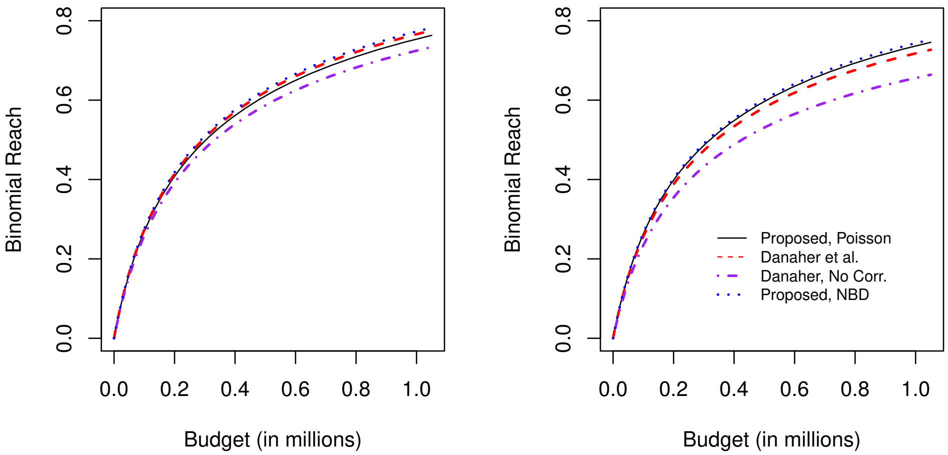

Figure A1: Performance comparison between the proposed method and Danaher's using simulated data (left) and real data (right).

set to zero, at each budget across 100 simulation runs using the binomial definition of reach (both for this simulated data study on the left and the following real data study on the right). Finally, we also demonstrate a variant of our Poisson model. Specifically we replace the Poisson distribution in Equation 3 with a NBD, with the same expected value $\gamma_j = s_j z_{ij}$, which corresponds to

$$P(X_j = 0|Z_j = z_{ij}, \mathbf{c}) = (1 + s_j)^{-z_{ij}}, \quad P(Y = 0|\mathbf{c}) \approx \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}(1 + s_j)^{-z_{ij}} \tag{A4}$$

Inserting (A4) into the left hand side of Equation 5 provides a NBD version of our standard objective function which can be optimized in a similar fashion. We plot the associated reach estimate using the blue dotted line (Proposed NBD).

As expected, Danaher's method performs slightly better than our Poisson approach on the simulated data, because the data is generated from the MNBD as assumed by his model. However, when using the NBD objective function on the simulated data, we see the proposed NBD method slightly outperforms even Danaher. All three methods here outperform the No Correlation version of Danaher's method, since it is the only approach to assume independence across websites.

The right plot in Figure A1 shows the average reach at each budget across the 100 sample runs using the December 2011 comScore Media Metrix data. Specifically, we use Internet usage data from the top seven most-visited websites that support Internet display advertisements. The data contained 51,093 users who visited one of the seven websites at least once in December 2011. We fit all methods to 100 randomly chosen user subsets of size 5,109 (approximately 10% of the population), then calculated reach using the budget allocations on the remaining 90% holdout data. Again, we use the CPMs as given in comScore Inc.'s Media Metrix data from May 2010 (Lipsman, 2010). In this scenario, the proposed Poisson reach estimate slightly outperforms the reach obtained under Danaher's method. Presumably this occurs because the real data does not precisely follow an MNBD. Overall, as seen in the left panel of Figure A1, the two methods yield very similar reach results. Again, the Danaher No Corr. results are considerably worse than for the other methods, due to the inherent correlations among the websites in the real data.

One interesting feature here is that the Proposed NBD method slightly outperforms the Proposed Poisson in both plots of Figure A3. However, this is to be expected, since the reach measure being used here is the binomial reach (and thus slightly favors the optimization based on a binomial distribution rather than an exponential). There is one major potential disadvantage to using the NBD over the Poisson objective function though. Unlike when using the Poisson distribution for the objective function, the resulting objective function for the NBD approach is no longer convex so can be harder to globally optimize. As noted with Danaher's approach, these criteria can get stuck at local optima and require multiple initializations to achieve reliable results. As the optimizations under consideration get more and more complex, this optima issue could potentially become problematic.

Appendix B Algorithm Details, Convergence, and Efficiency

B.1 Intuition Behind the Sparsity of Websites Chosen Using Our Method

It is not obvious why Equation 5 should produce a sparse solution set, just as it is not obvious that the L1 penalty in the Lasso will have the same effect. Figure A2, taken from James et al. (2013), provides additional intuition. The blue diamond in the left hand plot represents the constraint region for the standard Lasso, while the ellipses represent the objective function that is being minimized (in this case sum of squares). All points on an ellipse represent equal sum of squares, with larger ellipses corresponding to higher sum of squares. The goal then is to find a point within the shaded region which has smallest possible

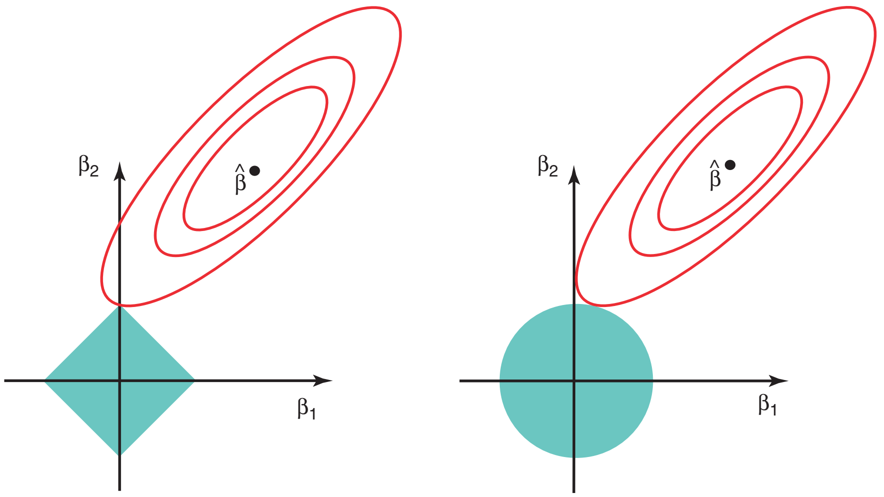

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$*, while the red ellipses are the contours of the RSS.*

Figure A2: Geometric explanation of the Lasso's ability to produce sparse solutions.

sum of squares. The fact that the constraint region has sharp points on the axes means that the point which minimizes the criterion, while still lying within the constraint, often falls on one of these axes. For example, in the above plot the ellipse which first touches the shaded region corresponds to $\beta_1 = 0$. By comparison the right hand plot corresponds to ridge regression where the constraint set has no points and hence the solution does not fall on the axes. For our method, while the criterion to be optimized is not sum of squares, the form of the constraint set is similar to that for the Lasso (it corresponds to the positive part of the shaded region), so our approach also produces sparse solutions.

B.2 Algorithm

Our objective function of Equation 7 in the Methodology Section can be written in statistical form using an ℓ_1 penalty:

$$f(\mathbf{w}) = g(\mathbf{w}) + \|\mathbf{w}\|_1, \tag{A5}$$

where $g(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} e^{-\gamma_i}$ is a differentiable convex function of $\mathbf{w}$, and $\|\mathbf{w}\|_1 = \sum_{j=1}^{p} |w_j|$ is a separable convex but not differentiable function. It has been shown in Luo and Tseng (1992) that a coordinate descent algorithm, which iteratively minimizes the objective as a function of one coordinate at a time, will achieve a global minimum for functions of the form in Equation A5. Thus convergence using coordinate descent is guaranteed for our objective function, since it is in the form specified by Luo and Tseng.

Because no closed-form solution exists for Equation 7, we employ a Taylor approximation to Equation A5, resulting in Equation 8. To minimize Equation 8 over w_j, with all w_k $(k \neq j)$ fixed, we first compute the partial derivative with respect to w_j which is given by

$$\sum_{i=1}^{n} [\eta_{ij}(w_j - \tilde{w_j}) - \theta_{ij}] + \lambda \tag{A6}$$

for $w_j > 0$. Setting Equation A6 equal to zero gives Equation 9.

We can also use Equation A6 to find a starting point for our algorithm, i.e., the λ value corresponding to $B = 0$. To do this, we employ the same procedure for calculating H_j as used in Equation 9. H_j measures whether our algorithm has set a coefficient to drop below zero. We can use this same procedure to initialize the first λ at which $B = 0$. In particular, we first set $\tilde{w_j} = 0$ for all $j = 1, \dots, p$, which corresponds to zero budget. Then, we calculate H_j for each website and set our initial λ value to $\max H_j$, with $j = 1, \dots, p$. To calculate increasing budgets, we use this value as λ_{max} and incrementally decrease λ by steps. The step size and number of steps are both parameters of the algorithm and are thus specified by the researcher depending on desired granularity and maximum budget. For example, for the McRib case study we used 500 steps at a step size of 0.01.

B.3 Convexity

To demonstrate that $\frac{1}{n}\sum_i e^{-\gamma_i}$ is convex, we compute the Hessian of partial second derivatives, $\mathcal{H}$, and show that the p by p matrix is positive semi-definite. First, standard calculations show that $\mathcal{H} = \frac{1}{n}\sum_i \mathcal{H}_i$ where the (j,k)th component of $\mathcal{H}_i$ is given by

$$\mathcal{H}_{ijk} = e^{-\gamma_i} \times \begin{cases} (z_{ij}s'_j)^2 - z_{ij}s''_j & j = k \\ z_{ij}z_{ik}s'_j s'_k & j \neq k \end{cases}.$$

Second, to demonstrate that $\mathcal{H}$ is positive semi-definite we need to show that $\mathbf{x}^T\mathcal{H}\mathbf{x} \geq 0$ for all $\mathbf{x}$. But

$$\mathbf{x}^T\mathcal{H}\mathbf{x} = \mathbf{x}^T\frac{1}{n}\sum_i \mathcal{H}_i\mathbf{x} = \frac{1}{n}\sum_i \mathbf{x}^T\mathcal{H}_i\mathbf{x} = \frac{e^{-\gamma_i}}{n}\sum_i \left[\left(\sum_j x_j z_{ij} s'_j\right)^2 - \sum_j x_j^2 z_{ij} s''_j\right]$$

This term is guaranteed to be non-negative provided s_j'' is negative for all j. Hence, a sufficient condition for our objective function to be convex is that the s_j functions are concave.

Appendix C NCL Campaign: Unobserved Consumer Heterogeneity

In this appendix we provide a demonstration of how our method may be used to incorporate unobserved consumer heterogeneity. In addition to observed heterogeneity such as demographics and/or past browsing behavior, the extent to which a consumer pays attention to a particular Internet display ad might also depend on additional heterogeneous factors that are unobservable to campaign managers. For example, a consumer who had a recent bad travel experience is more likely to ignore an ad by NCL compared to someone who is enthusiastically planning for his/her next family vacation. Such unobserved heterogeneity can be incorporated into our framework by modeling γ_i as coming from a random distribution with $E(\gamma_i) = \sum_j s_j z_{ij}$. Specifically, if γ_i is a random variable then

$$P(Y=0|\mathbf{c}) = E_\gamma\left(P(Y=0|\mathbf{c},\gamma)\right) = E_\gamma(e^{-\gamma}|\mathbf{c}) \approx \frac{1}{n}\sum_{i=1}^{n} E_{\gamma_i}(e^{-\gamma_i}|\mathbf{c}). \tag{A7}$$

This last term in Equation A7 is in fact the moment generating function (mgf) for γ_i evaluated at -1. Hence, the expression can be easily computed for a variety of possible distributions on γ_i. For example, we could model $\gamma_i \sim N\left(\mu_v \sum_j s_j z_{ij}, \sigma_v^2 \left(\sum_j s_j z_{ij}\right)^2\right)$, where $\mu_v = E(v_i)$, $\sigma_v^2 = Var(v_i)$. Here v_i is a random variable representing the amount of attention consumer i pays to the ad when it is served on a given website. Then, using the Gaussian mgf, $P(Y=0|\mathbf{c})$ can be approximated by

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\left(\mu_v - \sigma_v^2 \sum_j s_j z_{ij}/2\right)\sum_j s_j z_{ij}}. \tag{A8}$$

Although a bit more complicated than Equation 5, Equation A8 can be optimized by updating the values for θ_{ij} and η_{ij} in Equation 8. Thus, optimizing reach with the additional unobserved heterogeneity in γ_i would proceed in an almost identical fashion to our main model.

Nevertheless, this extension requires campaign managers to estimate values for μ_v and σ_v^2. Note that, for $\mu_v = 1$ and $\sigma_v^2 = 0$, Equation A8 is identical to Equation 5 in our main model. In such cases, all consumers are assumed to pay full attention to the ad once it is served. In practice, as discussed above, it is possible that consumers are heterogeneous with regard to the amount of attention they devote to a particular ad. To capture such unobserved consumer heterogeneity, campaign managers can collect additional data from auxiliary studies such as the increasingly popular eye-tracking study (e.g., Wedel and Pieters 2000; Aribarg et al. 2010; Wedel and Pieters 2012). For example, campaign managers can post the ad on test websites and use eye-tracking devices to measure the duration at which each consumer fixates on the ad. With observed values of fixation duration over a random sample of n^* consumers, campaign managers can easily estimate values of μ_v and σ_v^2. For example, if we consider that a consumer pays full attention to an ad after 100 milliseconds of eye fixation (Wedel and Pieters, 2012), we can record the number of milliseconds each consumer fixates on the ad (d_i) and compute $v_i = \min\left(1, \frac{d_i}{d_{max}}\right)$, with $d_{max} = 100$. Using natural sample statistics, we can then estimate μ_v and σ_v^2 as:

$$\hat{\mu}_v = \frac{1}{n^*}\sum_{i=1}^{n^*} v_i, \quad \hat{\sigma}_v^2 = \frac{1}{n^*-1}\sum_{i=1}^{n^*}\left(v_i - \hat{\mu}_v\right)^2.$$

We provide a demonstration of this extension below.

For this example, we again consider the subset of users who visited at least one aggregate travel website in January 2011 (6,431 users). We rerun our method following the details in the Methodology Section, where we assume values of μ_v and σ_v are known to the NCL campaign manager, presumably through an eye-tracking study. For the purposes of this example, we use $\mu_v = 0.5$ and $\sigma_v = 0.15$. We modify both the proposed method and the benchmark greedy method to account for these additional variables.

Figure A3 shows the results of this scenario, with the modified proposed and benchmark greedy methods, keeping the equal and Danaher allocations unchanged. We consider two situations. In the left-hand plot of Figure A3, σ_v and μ_v are known to advertisers, and we incorporate those values into the optimization directly using the methodology from the Methodology Section. As this left-hand plot shows, there is a clear advantage for the proposed method. In contrast, the plot on the right demonstrates the case where μ_v and σ_v are still 0.5 and 0.15, respectively, but this information is not available to advertisers. In this case, advertisers could run the method without adjusting for μ and σ, i.e. running the optimization with $\mu_v = 1$ and $\sigma_v = 0$. The estimates used in the right-hand figure are thus calculated from the original optimization, but reach is calculated using the underlying

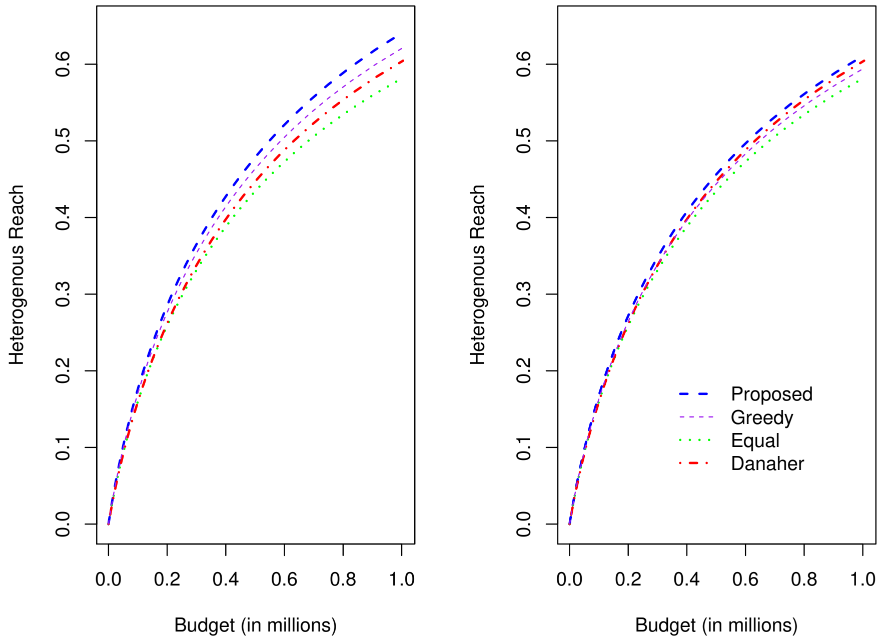

Figure A3: Overall reach when true unobserved heterogeneity values are incorporated into the optimization (left) and when optimization does not incorporate unobserved heterogeneity (right).

"truth" of $\mu_v = 0.5$ and $\sigma_v = 0.15$ to show what happens when the optimizations are run without accounting for the unobserved heterogeneity.

As the right-hand plot of Figure A3 shows, the proposed method still performs better than the other three methods, though it sacrifices much of the advantage it gained by correctly incorporating the heterogeneity information in the left-hand plot. In particular, the estimates given by Danaher's method on the eight travel websites actually perform better than the greedy method and only slightly behind the proposed method. Having accurate heterogeneity estimates can greatly assist advertisers in adjusting their allocations to accommodate for any unobserved consumer heterogeneity. However, our example shows that, even if NCL fails to account for heterogeneity, the proposed method still outperforms the other benchmark methods.

Appendix D Properties of Method as n Approaches Infinity

D.1 p fixed

First we consider the setting where p is fixed but $n \to \infty$. Let $\mathbf{c}_m$ represent the value of $\mathbf{c}$ that maximizes reach over the entire population subject to a specific budget B i.e.

$$\mathbf{c}_m = \arg\min_{\mathbf{c}} E[e^{-\sum_j s_j(c_j)Z_j}] \quad \text{such that} \quad \sum_{j=1}^{p} c_j s_j \tau_j \le B, \quad \text{and} \quad c_j \ge 0, \quad j = 1, \dots, p,$$

while $\mathbf{c}_n^*$ is the corresponding value that maximizes reach over our sample, $\mathbf{Z}$. Then our goal is to prove that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} \to E[e^{-\sum_j s_j(c_{mj})Z_j}] \quad a.s. \tag{A9}$$

First we select a dense, but finite, grid over $\mathbf{c}$ such that all points in the grid satisfy the constraints in Equation 1, and for any value of $\mathbf{c}$, there exists a point $\mathbf{c}'$ on the grid such that

$$|s_j(c_j) - s_j(c_j')| < \epsilon/p \tag{A10}$$

for all $j = 1, \dots, p$, where $\epsilon > 0$ is a suitably small value. Note that this is guaranteed to be possible as long as s_j is continuous with $|s_j'| < K < \infty$ for all j. This is a reasonable assumption since s_j is bounded between 0 and 1. We also include $\mathbf{c}_m$ as one of the points on the grid. Then by the strong law of large numbers, $\exists N_1$ such that $\forall n > N_1$,

$$\left| \frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_j')Z_{ij}} - Ee^{-\sum_j s_j(c_j')Z_j} \right| < \epsilon, \quad a.s. \tag{A11}$$

for every $\mathbf{c}'$ on our grid. Let $M = \max_j E[Z_j]$. Then, also by the strong law of large numbers, $\exists N_2$ such that $\forall n > N_2$,

$$\frac{1}{n}\sum_{i=1}^{n} Z_{ij} < M + \epsilon \tag{A12}$$

for all j a.s.

Consider

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - Ee^{-\sum_j s_j(c_{mj})Z_j} = A + B + C$$

where $A = \frac{1}{n}\sum_i e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - \frac{1}{n}\sum_i e^{-\sum_j s_j(c_j')Z_{ij}}$, $B = \frac{1}{n}\sum_i e^{-\sum_j s_j(c_j')Z_{ij}} - Ee^{-\sum_j s_j(c_j')Z_j}$, $C = Ee^{-\sum_j s_j(c_j')Z_j} - Ee^{-\sum_j s_j(c_{mj})Z_j}$, and $\mathbf{c}'$ is a point on our grid. Then clearly, by the definition of $\mathbf{c}_m$, for any $\mathbf{c}'$, $C \geq 0$. Similarly by Equation A11 for any $\mathbf{c}'$, $B > -\epsilon$ a.s. for $n > N_1$. Finally note that, for $n > N_2$,

$$\begin{aligned}
|A| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - e^{-\sum_j s_j(c_j')Z_{ij}}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{j=1}^{p}\left(s_j(c_{nj}^*)Z_{ij} - s_j(c_j')Z_{ij}\right)\right| \quad \text{(since } de^{-x}/dx < 1 \text{ for } x > 0\text{)} \\
&\leq \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left|s_j(c_{nj}^*)Z_{ij} - s_j(c_j')Z_{ij}\right| \\
&= \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left|s_j(c_{nj}^*) - s_j(c_j')\right| Z_{ij} \\
&\leq \frac{\epsilon}{p}\sum_{j=1}^{p}\frac{1}{n}\sum_{i=1}^{n} Z_{ij} \quad \text{(for appropriately chosen } \mathbf{c}' \text{ by Equation A10)} \\
&\leq \frac{\epsilon}{p}\sum_{j=1}^{p}(M+\epsilon) = \epsilon(M+\epsilon) \quad a.s. \quad \text{(by Equation A12)}
\end{aligned}$$

Hence, $A + B + C \geq -\epsilon(M+\epsilon) - \epsilon + 0 = -\epsilon_2$. Thus, for $n > \min(N_1, N_2)$,

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} > Ee^{-\sum_j s_j(c_{mj})Z_j} - \epsilon_2 \quad a.s. \tag{A13}$$

Now consider

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - Ee^{-\sum_j s_j(c_{mj})Z_j} = D + E$$

where $D = \frac{1}{n}\sum_i e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - \frac{1}{n}\sum_i e^{-\sum_j s_j(c_{mj})Z_{ij}}$ and $E = \frac{1}{n}\sum_i e^{-\sum_j s_j(c_{mj})Z_{ij}} - Ee^{-\sum_j s_j(c_{mj})Z_j}$. Then, by the definition of $\mathbf{c}_n^*$, $D < 0$. Also, by Equation A11, $E < \epsilon$ for $n > N_1$ a.s. Thus

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} < Ee^{-\sum_j s_j(c_{mj})Z_j} + \epsilon \quad a.s. \tag{A14}$$

Thus, by Equation A13 and Equation A14, for $n > \min(N_1, N_2)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c_{nj}^*)Z_{ij}} - Ee^{-\sum_j s_j(c_{mj})Z_j}\right| < \epsilon_3 \quad a.s.$$

where $\epsilon_3 = \max(\epsilon, \epsilon_2)$. Hence, for suitably small ϵ these two quantities are arbitrarily close a.s. so Equation A9 is proved.

D.2 p increasing with n

Now we consider the setting where p and n both approach infinity. We make the following assumptions, as $n \to \infty$:

$$p \to \infty \tag{A15}$$

$$\frac{p}{\log n} \to 0 \tag{A16}$$

$$E[Z_j] < M < \infty, \quad j = 1, 2, \ldots \tag{A17}$$

$$|s_j'| < K < \infty, \quad j = 1, 2, \ldots \tag{A18}$$

and as $p \to \infty$

$$E[e^{-\sum_j s_j(c_{mj})Z_j}] \to r \quad \text{where } r \text{ is a constant.} \tag{A19}$$

Note that Equation A16 assumes that, while p grows to infinity, it grows slower than n. This assumption matches the data we observe in practice where the number of websites (p) under consideration may be very large, in the thousands, but the number of potential customers (n) is even larger, in the hundreds of thousands or millions.

Our goal is to prove that, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} \to r \quad a.s. \tag{A20}$$

First we note that, by Equation A19, for any $\epsilon > 0$, $\left|E[e^{-\sum_j s_j(c_{mj})Z_j}] - r\right| < \epsilon$ for large enough p. Thus, for large p,

$$\left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} - r\right| < \left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} - E[e^{-\sum_j s_j(c_{mj})Z_j}]\right| + \epsilon.$$

We can use essentially the same arguments as in the previous section to show that, for any fixed p, $\left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} - E[e^{-\sum_j s_j(c_{mj})Z_j}]\right| \to 0 \quad a.s.$ The key point to observe is that quantities such as $\mathbf{c}_m, \mathbf{c}^*_n, \mathbf{c}', \mathbf{Z}, N_1, N_2$ as well as the grid over $\mathbf{c}$ are now functions of p. However, Equation A16 guarantees that for any fixed p, we can grow n large enough that Equations A10, A11, and A12 all hold.

First, note that Equation A18 guarantees Equation A10 will hold. Second, for Equation A12 to hold, we only need that n grows faster than p (since we need to bound p different values of $\bar{Z}_j$). This is certainly guaranteed by Equation A16. Finally, the most challenging part is to show that Equation A11 holds, since the bound must hold for all points in the grid, and the grid could grow at an exponential rate in p. However, the $\log n$ term in Equation A16 accounts for this possibility so n can still grow large enough to ensure Equation A11 holds.

Once we observe that, for any fixed p, we can grow n large enough so that Equations A10, A11, and A12 all hold, the argument in the previous section can be applied to show that, for $n > \min(N_1(p), N_2(p))$, $\left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} - E[e^{-\sum_j s_j(c_{mj})Z_j}]\right| < \delta$ for some small $\delta > 0$. Hence, for large enough p, we can always find a corresponding $n > \min(N_1(p), N_2(p))$ such that

$$\left|\frac{1}{n}\sum_{i=1}^{n} e^{-\sum_j s_j(c^*_{nj})Z_{ij}} - r\right| < \epsilon + \delta$$

Thus our reach estimate is guaranteed to converge to the population reach as n and p approach infinity.

Appendix E Illustration of Correlation in Website Viewership

In this appendix, we provide both analytical and empirical illustrations of how the proposed method incorporates correlations into the objective function. Our approach models reach as $1 - E_Z\left(\prod_j G_j\right)$, where $G_j = e^{-s_j Z_j}$ and Z_j is the number of page views a random person has at website j. In the case $p = 2$ this expression reduces to

$$\text{Reach} = 1 - E_Z\left(G_1 G_2\right) = 1 - E_Z(G_1)E_Z(G_2) - Cov(G_1, G_2)$$

Note that $1 - E_Z(G_1)E_Z(G_2)$ is the reach if Z_1 and Z_2 are independent, so our objective function for reach can be seen as modeling reach as a term assuming independence plus an adjustment for the covariance between G_1 and G_2. One important observation is that the adjustment is in terms of the covariance between G_1 and G_2 rather than between Z_1 and Z_2. If the covariance in viewership between sites is positive, then reach is lower relative to the independent case and vice versa if covariance is negative. This result matches our intuition, since we are likely to reach fewer unique customers when two sites are positively correlated because the same people tend to visit both sites. Conceptually this idea extends to the case $p > 2$. For example, for $p = 3$ the expression becomes

$$\text{Reach} = 1 - E_Z\left(G_1 G_2 G_3\right) = 1 - E_Z(G_1)E_Z(G_2)E_Z(G_3) - Cov(G_1, G_2 G_3) - EG_1 Cov(G_2, G_3),$$

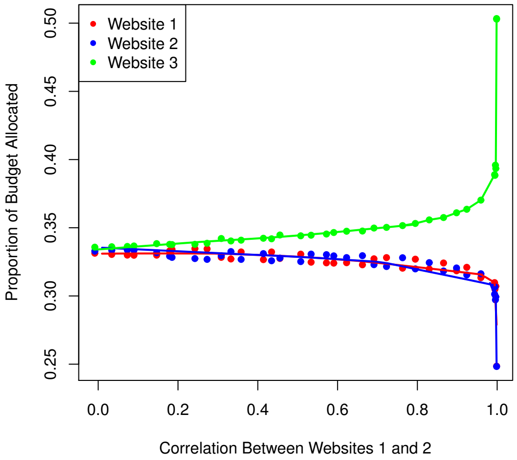

Figure A4: Illustration of budget allocation with varying correlations in website viewership.

which also corresponds to the reach under independence less a reduction if the covariances of G_1, G_2 and G_3 are positive. However, the reach becomes harder to express for larger p since it involves terms of order p so cannot be expressed using a single covariance term.

We can also empirically illustrate the effects of correlation on budget allocation by considering a case with $p = 3$ websites, all generated from the same distribution with the same cost. However, the viewership for websites 1 and 2 has a measurable correlation ranging from 0.0 (fully independent) to 1.0 (perfect positive correlation), and website 3's viewership is generated entirely independently of the other two websites (correlation of 0).

Figure A4 shows the change in budget allocation across the three websites as the correlation between websites 1 and 2 changes. When the correlation between websites 1 and 2 is zero, all three websites are independent. In this case, the algorithm allocates one-third of the budget to each of the three websites, since no website has a clear advantage over the other two. As the correlation between websites 1 and 2 increases, the algorithm gradually allocates more budget to website 3 and splits the remaining budget among websites 1 and 2. When these two websites become perfectly correlated, the algorithm divides the budget in half, allocating one half to website 3 and the other half across websites 1 and 2.

Appendix F Change in Reach as a function of B

Let $Q(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} e^{-\gamma_i}$ where $\gamma_i = \sum_j s_j(w_j)z_{ij}$ and reach equals $1 - Q$. Then $\frac{\partial Q}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n} e^{-\gamma_i} s_j'(w_j)z_{ij}$. At any solution point $w_1, \ldots, w_p$ will be chosen such that $\sum_j w_j = B$. Hence, a Δ increase in B will be associated with a Δ increase in $\sum_j w_j$. In particular we will select the w_j associated with the largest value of $-\frac{\partial Q}{\partial w_j}$ as this will give the largest instantaneous increase in reach. Hence, a Δ increase in B will be associated with a

$$\Delta \max_j \frac{1}{n}\sum_{i=1}^{n} e^{-\gamma_i} s_j'(w_j)z_{ij}$$

increase in reach. In practice for larger values of B there may be several websites whose derivatives are all equal to this maximum value. In that setting the Δ increase in B would be shared among all these websites. However, this would not impact the overall change in reach since we would simply apportion this derivative among several websites. Since all the derivatives would be equal, the conclusion is unaffected.

Appendix G Complete Enumeration

One approach to maximize reach involves computing a complete enumeration of all possible budget allocations. In theory, by enumerating these possible allocations and calculating their subsequent reach measures, we are guaranteed to find the global optimal solution. However, in practice this is computationally prohibitive, especially for a large number of websites and high budgets.

In what follows, we attempt to verify results of the proposed method using complete enumeration over the eight aggregate travel websites identified in the Norwegian Cruise Lines (NCL) Section. This gives us a reasonably small set of websites over which we can run the enumeration. In addition, for feasibility we choose a moderate budget of \$50,000. Lastly, we run the complete enumeration over the 6,431 users who visited at least one of the eight aggregate travel websites in January 2011, again as used in the NCL Section. Ideally we would compute a complete enumeration of all possible combinations of budget points across all eight websites. However, testing all combinations of a budget allocation of \$50,000 on eight websites in increments of \$1 would involve on the order of 10^{37} total calculations, which even at one million calculations a second would require 10^{23} years to compute. Therefore, we employ a modified iterative enumeration approach as follows.

We first initialize the method with 10 evenly-spaced budget points (for \$50,000, this represents an incremental increase of \$5000 per budget point), examining all possible budget allocations which total \$50,000. Then, after finding the optimal solution at this coarsest level, we create a finer grid of 10 budget points around the solution, with five points above the solution and five below. The grid size of these budget increments are first done as 10% of the budget allocated to a given website, then 5%, then 2.5%, etc. For example, if website j was allocated \$1,000, the first iteration would have budget increments of \$100 (e.g. ten budget points ranging from $w_j = 500$ to $w_j = 1500$). For the same budget of \$1,000, the second iteration would have grid size of \$50, and the third iteration would have grid size of \$25. This ultimately results in a more precise solution around the values initially chosen by the enumeration. We repeat this process until the overall reach achieved does not change by more than 0.1%, or 0.001.

We then compare the reach performance of the proposed method and complete enumeration using binomial reach from Appendix A.1, i.e. $1 - \frac{1}{n} \sum_{i=1}^{n} \prod_j (1 - s_j)^{z_{ij}}$ where s_j is the probability that the ad is served on the jth website. We use binomial reach in this comparison because the reach definition under our approach would favor the proposed method. Under this comparison, the enumeration method obtains a reach higher by only 0.3%. The mean absolute deviation (MAD) between the enumeration optimal schedule and the proposed optimal schedule is \$935.

We further carry out an alternative enumeration method using an initialization around the solution given by the proposed method on the eight aggregate sites. Here, instead of initializing the enumeration with an evenly-spaced grid of budget points, we create an initial grid from the solution provided by the proposed method. Again, we run this procedure iteratively. As described above, we start with a coarse grid around the proposed method's solution (i.e., 10% of the budget allocated to the website) and then narrowing the grid around the best allocation (i.e., then 5%, 2.5%, etc.) until the reach achieved does not change by more than 0.1%. Under this comparison, the enumeration method achieves a reach higher by 0.2% and the MAD of the budget allocations between the enumeration and the proposed method is \$33.5.

Lastly, we enumerate all possible budget allocations which total \$50,000 for three aggregate travel websites. While impractical for eight websites, a complete enumeration over a fine grid is much more reasonable for a problem of this size. Here, we employ a grid size of \$5 per budget point, meaning we test 10,000 budget points per website. Within this problem setting, the MAD of budget allocations reduce to \$11. And the complete enumeration obtains a reach higher by 0.2%.

Appendix H Comparing Proposed and Benchmark Greedy Methods

Here we explore performance comparisons between the proposed method and the benchmark Greedy algorithm when websites vary in 1) their relative attractivenesses with respect to cost and/or total visitation; and 2) the degree of correlation in their viewership. Ceteris paribus, a website is more attractive to an advertiser when it entails low cost and/or high visitation. Recall that a key difference between the proposed and the benchmark greedy methods is that the latter will not adjust funds allocated to previously chosen websites once a subsequent website has entered the allocation. When the websites are clearly distinguishable, in not only the order in which they should be chosen but also how much budget should be allocated, incremental gains from freely adjusting budgets across sites should be relatively small. Therefore, we expect similar performance of the two methods under such scenarios. In contrast, when the relative attractivenesses of websites are similar, given diminishing returns to additional funds spent on each website, the proposed method can benefit a great deal from iteratively adjusting budgets across sites to achieve the highest possible marginal return. In this setting we expect the proposed method to outperform the benchmark method.

Additionally, the comparitive performance between the two methods can also be impacted by viewership correlation across websites. When websites exhibit overlap in viewership, the proposed method may find it advantageous to allocate more or less funds to previously chosen websites in attempts to maximize reach under a given budget. In contrast, because the benchmark greedy algorithm lacks the flexibility of adjusting budget once a website has already been chosen, we expect that a complex correlation structure can ultimately hurt the benchmark greedy algorithm's performance.

As an illustration we consider an example of three websites with visitation generated in the same manner as the simulated websites in the Simulation Studies Section. For simplicity,

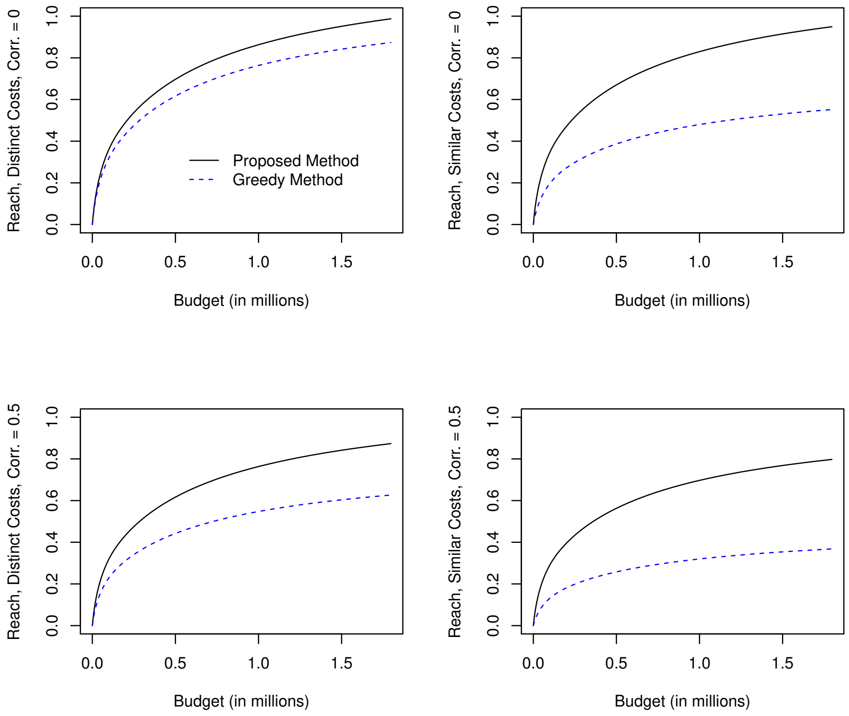

Figure A5: Comparison in varying Cost and Correlation settings: proposed vs. benchmark greedy

visitation at all three websites is drawn from the same distribution. Therefore, the relative attractiveness of the websites is gauged by their different cost curves, which are generated in the same manner as described in our main paper. We tested a total of four scenarios: 1) distinct attractiveness, no viewership correlation; 2) similar attractiveness, no viewership correlation; 3) distinct attractiveness, correlated viewership; and 4) similar attractiveness, correlated viewership.

In the distinct attractiveness condition, the average CPM's of the first, second and third websites are respectively \$2.50, \$5.00, and \$7.50. In the similar attractiveness condition, the three websites have average CPMs of \$4.75, \$5.00, and \$5.25, respectively. Within each condition, we further generated two different data matrices, one with no viewership correlation and the other with viewership correlations of $\rho = 0.5$ across the three websites.

Figure A5 shows the relative performance of the two methods under these four conditions. All results are consistent with our expectations. The top left plot, corresponding to distinct attractiveness and no viewership correlation, shows the closest performance comparison between the two methods, while the bottom right plot, corresponding to similar attractiveness and high viewership correlation, shows the greatest performance difference. The remaining two figures illustrate more moderate reductions in performance for the benchmark Greedy method.

To summarize, while the two methods are both computationally efficient, from a marketing perspective, the proposed method generally outperforms the benchmark greedy algorithm to a practically significant degree, with the magnitude of gains varying based on the relative attractivenesses of websites and the degree of viewership correlation. Nevertheless, we do note that since the benchmark greedy algorithm does not have to reallocate across websites, it necessarily provides a faster solution. This gain in computation time may prove preferable in situations where speed is prioritized over performance maximization or where comparatively small gains are expected due to websites being distinct and uncorrelated, as in the upper left scenario of Figure A5. However, this scenario is unlikely to occur in practice.

Appendix I Supplementary Information on Empirical Results

Figure 6 in the main text shows clear differences among the methods in terms of performance. These reach values are the averages across all 100 runs with different 10% subsets of the data. It is worth noting these results are remarkably stable, even with only 10% of the total data used for any given run. Table A1 compares the reach achieved by the proposed method and the benchmark greedy method at the runs corresponding to the given percentiles (that is, the 50th percentile here is the median reach achieved by the method) at three example budgets: \$50,000, \$100,000, and \$250,000. The proposed method consistently outperforms the benchmark greedy method, even when comparing the 5th percentile of the proposed method to the 95th percentile of the benchmark greedy method. Further, these curves do not overlap for any of the 100 runs. That is, the largest reach achieved by the benchmark greedy method at any budget is always lower than the smallest reach achieved by the proposed method at any of the 100 runs. We further verify this using 99.9% intervals for the mean reach achieved by both the proposed method and the benchmark greedy algorithm; there is no overlap between the intervals, even at a confidence level of 99.9%.

Table A2 provides an overview of correlation in viewership among the 16 website groups in the McRib example, both within groups and among groups. Within group correlations in the table (diagonal elements) are calculated by taking the mean of all absolute correlations

	$50,000		**$100,000**		**$250,000**	
Percentile	Proposed	Benchmark	Proposed	Benchmark	Proposed	Benchmark
5th	0.0679	0.0361	0.1149	0.0698	0.2023	0.1444
25th	0.0684	0.0363	0.1158	0.0702	0.2045	0.1458
50th	0.0685	0.0366	0.1162	0.0705	0.2053	0.1463
75th	0.0688	0.0368	0.1165	0.0709	0.2060	0.1470
95th	0.0691	0.0369	0.1171	0.0712	0.2072	0.1475

Table A1: Reach achieved by percentile across 100 runs at budgets of $50,000, $100,000, and $250,000 for both the proposed and benchmark greedy methods.

between websites in a particular group. For example, the Newspaper category shows moderately high average correlation in viewership among websites with a value of 0.48. In contrast, there is not much correlation in viewership among websites in the E-mail category, only 0.01 on average. The off-diagonal elements of Table A2 show the maximum absolute correlation between each pair of groups. This is calculated by taking the maximum correlation between two websites from the respective groups. For example, there is a high correlation of 0.96 between Newspaper and Portal sites. In contrast, there is a low correlation between Filesharing and E-mail sites, only 0.03.

References

Aribarg, Anocha, Michel Wedel and Rik Pieters (2010), ‘Raising the BAR: Bias Adjustment of Recognition Tests in Advertising.’, *Journal of Marketing Research* **47**(3), 387–400.

Danaher, Peter J., Janghyuk Lee and Laouchine Kerbache (2010), ‘Optimal Internet Media Selection’, *Marketing Science* **29**(2), 336–347.

Lipsman, Andrew (2010), The New York Times Ranks as Top Online Newspaper According to May 2010 U.S. comScore Media Metrix data, Technical report, comScore, Inc.

Wedel, Michel and Rik Pieters (2000), ‘Eye Fixations on Advertisements and Memory for Brands: A Model and Findings’, *Marketing Science* **19**(4), 297–312.

Wedel, Michel and Rik Pieters (2012), ‘Ad Gist: Ad Communication in a Single Eye Fixation.’, *Marketing Science* **31**(1), 59–73.

Category	Com	Email	Ent	File	Game	Gen	Info	News	Onl	Photo	Port	Ret	Serv	Soc	Sport	Travel
Community	0.02	0.14	0.82	0.14	0.77	0.14	0.47	0.16	0.55	0.88	0.21	0.39	0.21	0.26	0.12	0.15
Email	.	0.01	0.07	0.03	0.28	0.04	0.07	0.05	0.09	0.04	0.87	0.10	0.10	0.06	0.12	0.04
Entertainment	.	.	0.02	0.78	0.32	0.90	0.76	0.92	0.28	0.83	0.90	0.30	0.69	0.24	0.79	0.10
Fileshare	.	.	.	0.05	0.27	0.05	0.15	0.56	0.67	0.13	0.17	0.10	0.13	0.14	0.10	0.07
Gaming	.	.	.	.	0.01	0.12	0.82	0.32	0.85	0.12	0.25	0.14	0.95	0.09	0.51	0.09
General News	.	.	.	.	.	0.28	0.76	0.94	0.08	0.04	0.96	0.08	0.10	0.34	0.85	0.11
Information	.	.	.	.	.	.	0.02	0.77	0.51	0.18	0.76	0.30	0.11	0.24	0.65	0.27
Newspaper	.	.	.	.	.	.	.	0.48	0.10	0.05	0.96	0.36	0.12	0.26	0.86	0.15
Online Shop	.	.	.	.	.	.	.	.	0.03	0.49	0.16	0.26	0.75	0.42	0.19	0.10
Photos	.	.	.	.	.	.	.	.	.	0.02	0.11	0.09	0.09	0.41	0.04	0.05
Portal	.	.	.	.	.	.	.	.	.	.	0.06	0.19	0.19	0.12	0.87	0.09
Retail	.	.	.	.	.	.	.	.	.	.	.	0.04	0.19	0.18	0.25	0.12
Service	.	.	.	.	.	.	.	.	.	.	.	.	0.01	0.15	0.19	0.05
Social Network	.	.	.	.	.	.	.	.	.	.	.	.	.	0.02	0.10	0.26
Sports	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.07	0.08
Travel	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0.18

Table A2: Overview of viewership correlation within and across the sixteen website categories in the McRib example.