Online Appendix for "Optimal Internet Media Selection"

(Version dated 26 February 2009)

A1: The MNBD in detail.

The exposure distribution model which underpins our optimization method is the multivariate NBD (MNBD). Briefly, the MNBD is constructed via a class of distributions named "Sarmanov distributions" (Lee 1996). Sarmanov distributions have the ability to combine univariate distributions into multivariate distributions, with the resulting model being akin to a Taylor series expansion for a probability distribution. An alternative way to view the construction of the bivariate NBD, for example, is that it is the mixture of a bivariate gamma distribution with two independent Poisson distributions. The Sarmanov model provides a way to derive the bivariate gamma distribution. Danaher and Hardie (2005) illustrate an analogous situation for a bivariate beta-binomial distribution. We now give details on the construction of the univariate and multivariate NBD.

A1.1 One Website

Danaher (2007) and Huang and Lin (2006) show that a negative binomial distribution (NBD) model for a single website fits the observed exposure distribution better than several other plausible models. Indeed, in a holdout sample at a future time point, Danaher (2007) shows that the error in predicting reach is only 10%, whereas competing models are significantly higher at 13 to 14%. Therefore, for a single website, we also use the NBD with mass function

$$\Pr(X_i = x_i \mid r_i, \alpha_i, t_i) = \binom{x_i + r_i - 1}{x_i} \left(\frac{\alpha_i}{\alpha_i + t_i}\right)^{r_i} \left(\frac{t_i}{\alpha_i + t_i}\right)^{x_i}, x_i = 0,1,2,\ldots, \quad \text{(A1)}$$

where r_i and α_i are the usual parameters for the gamma distribution. An additional parameter, t_i, not incorporated by Danaher (2007), permits the NBD to be rescaled depending on the time interval used to estimate the parameters (see Lilien, Kotler and Moorthy 1992, p.34). For instance, if one week of data are used to estimate the model, but an advertiser wants to predict the exposure

distribution for a 4 week period, then all that changes is t_i goes from 1 to 4. This results from the additive property of the Poisson distribution which is the basis for the NBD, and makes it extremely versatile for Internet exposure distribution applications.

Equation (A1) allows us to estimate the reach for website *i* from the NBD as

$$\text{Reach}_i = 1 - \Pr(X_i = 0) = 1 - \left(\frac{\alpha_i}{\alpha_i + t_i} \right)^{r_i}. \quad \text{(A2)}$$

Furthermore, the mean number of exposures per person is

$$E[X_i] = r_i t_i / \alpha_i \quad \text{(A3)}$$

A1.2 Multiple Websites

Equation (A1) is an excellent model for the exposure distribution of just one website. However, advertisers are likely to spread their budget over multiple websites, as this is a proven way to increase the reach of an ad campaign (Rust 1986). When estimating the exposure distribution for several websites, the duplication in audience must be taken into account (Danaher 1992; Rust 1986). For example, Table A1 gives the pairwise correlations for the 10 websites we later use for optimal media scheduling. Although none of the correlations exceed .5, some of them are large enough (e.g., .2678 between sites 4 and 7) to indicate considerable overlap in visitors to different websites. Even such moderate overlap is an impediment to increasing reach and so needs to be incorporated into the objective function (see §A2.1 below). In addition, there is intra-vehicle correlation due to repeat visits by the same person to the same website, which is also commonplace in traditional media (Barwise and Ehrenberg 1988; Morrison 1979). The higher the incidence of repeat visits, the lower the overall reach. The NBD model already captures this intra-vehicle correlation for single websites, but we still require a method for handling correlation among exposure to different websites.

Table A1: Correlation Matrix for the 10 Korean Websites

Website no.	1	2	3	4	5	6	7	8	9	10
1	1	0.0415	0.0529	-0.0018	0.0111	0.0027	0.0059	0.1039	0.1086	0.0772
2		1	0.0245	0.0262	0.0234	0.0338	0.0217	0.0359	0.0353	0.0380
3			1	0.0125	0.0083	0.0010	0.0114	0.0528	0.0283	0.0491
4				1	0.2343	0.1662	0.2678	0.0126	0.0356	0.0442
5					1	0.1614	0.2661	0.0196	0.0198	0.0584
6						1	0.1875	0.0180	0.0181	0.0611
7							1	0.0275	0.0259	0.0565
8								1	0.0523	0.0582
9									1	0.0256
10										1

Danaher (2007) considered several possible models for multiple websites, including a multivariate generalization of the NBD. This MNBD performed the best empirically as a model for predicting audiences to website ad campaigns, so we also employ it in this study. The MNBD model for the full exposure distribution, with truncation after third-order terms is given in equations (2) and (3) of the main paper and enables us to write down closed form expressions for reach, average frequency and effective reach, as given, respectively in equations (4) through (6) of the main paper.

An additional consideration is that the omega parameters in equation (4) are subject to the following constraints. General parameter restrictions for ω are given in Lee (1996), but in our case, the restrictions for ω_{12} (and similarly for all the other associations) in the MNBD are

Lower bound is $-1/\max\{L_1(1)L_2(1),(1-L_1(1))(1-L_2(1))\}$

Upper bound is $1/\max\{L_1(1)(1-L_2(1)),L_2(1)(1-L_1(1))\}$,

where $L_i(1)=\left(\frac{\alpha_i}{\alpha_i+t_i(1-e^{-1})}\right)^{r_i}$ i=1,2. We generally find these bounds are satisfied in practice and so place no practical limitations on the implementation of our model in the website exposure application.

A1.3: FORTRAN Code for the Optimization Algorithm

We now detail the FORTRAN code and calls to the IMSL(1997) library subroutine NCONF, which is used to obtain the maximum reach, staying within the budget constraint. The call to this subroutine is as follows.

CALL NCONF(FCN,M,ME,N,XGUESS,IBTYPE,XLB,XUB,XSCALE,IPRINT, MAXITN,X,FVALUE)

The values and description for each of these parameters is:

N is the number of websites (10 in our case), being the dimension of the X vector, which contains the share of impression values. It is this X vector which is varied so as to maximize reach (the components of X are labeled s_i in equations (10) and (11)). Initial values of X must be provided in the XGUESS vector. We set all the values of XGUESS to be 0.0. We also tried other values to ensure the resulting optimum is consistent across different start points. XLB and XUB are lower and upper bounds for X, which are set to 0.0 and 1.0, since share of impressions must be within this range.

M is the number of constraint equations, being 1 in our case, as per the constraint in equation (11) of the main paper. Notice that the constraints on X being between 0 and 1 have already been handled by the XLB and XUB vectors. ME is the number of constraints with exact equality, as opposed to an inequality, as occurs for equation 11. In our case ME=0, and additionally, IBTYPE=0, IPRINT=2 and MAXITN=100.

FCN is a FORTRAN function which calculates and outputs the reach value for a given value of X. It must be a separate function to NCONF, but it uses the parameters M, ME, N, X as input. The exact values for the FCN function are calculated by using equation (11). The output of the FCN function is parsed to FVALUE in the NCONF subroutine. The NCONF FORTRAN routine finds the optimal s_i values in less than a quarter second.

A2: Additional Issues for the Fitted MNBD

Given the relatively small correlations in Table A1, it is not immediately apparent that it is worth the trouble of allowing for correlation among websites at the optimization stage. Hence we now discuss two issues for the MNBD when applied to optimization. The first is whether or not it is worth including terms in the Sarmanov model for correlations, and the second is how many terms to include in the Sarmanov model.

A2.1 Allowing for Correlations in the Sarmanov Model

We employ Danaher's (2007) MNBD model as it is robust and more accurate than competing models (see his Table 4). If it happens that pairwise and higher-order associations are not important then his MNBD simplifies to being the product of univariate NBDs. Our paper is about the selection of Internet media vehicles, so we are agnostic about whether associations among websites are significant. If they are, then the MNBD easily accommodates them. If they aren't then the MNBD reduces to the product of NBDs, which presents no issues for optimization. The key is to have a model which can handle either situation, which the MNBD can do. Notwithstanding these comments, Table A2 shows the effect of ignoring correlations when determining the optimal schedule.

Table A2: Optimal Reach Values for MNBD and Independence Models

Budget	Original Data		Double Correlation	
	MNBD	Indep Model	MNBD	Indep Model
$100k	34.98	34.67	28.52	26.42
$50k	21.81	21.67	18.85	17.78

The 2nd and 3rd columns of Table A2 contain the optimal reach values obtained using the original data. As observed in Table A1, the correlations among websites for our data are not particularly large, so we artificially inflated thesm by dividing the observed non-exposure between each pair of websites by 1.2. This approximately doubled the correlations. Those results are in the 4th and

5th columns of Table A2. As expected, increasing the bivariate correlations among websites lowers the reach, since there is a higher overlap in the audience. The results for the original data using the MNBD model for budget levels $100k and $50k (34.98 and 21.81, respectively) are as for the Full model with no frequency capping in Tables 2a and 2b of the main paper. For the independence model we re-ran the optimization algorithm, but used a model that assumes independence among websites (i.e., the MNBD with no $\omega_{j_1 j_2}$ and $\omega_{j_1 j_2 j_3}$ parameters). This gives a different set of s_i values for each website. Using these alternative s_i values we re-estimated the reach using the full MNBD (with $\omega_{j_1 j_2}$ and $\omega_{j_1 j_2 j_3}$ parameters). It can be seen that the optimal reach under the independence model is lower than when the full MNBD is used. However, the differences are not large, but this is because the correlations between the websites are not large. When the correlations are approximately doubled, Table A2 reveals that there is a 2.1 percentage point difference for the $100k budget and a 1.1 percentage point difference for the $50k budget. Given the flatness of the reach surface, these differences are actually quite large. Obviously, the differences will widen as the correlation increases further. In our example for popular Korean websites, the correlations are not large, but in other instances, particularly for niche websites like gardening, or hobbies, the correlations are likely to be much higher. The key point is that the MNBD is robust enough to handle all these eventualities.

A 2.2 How many terms to include in the Sarmanov model

The following are results for a comparison of model accuracy when second-order and third-order terms are used in the Sarmanov model. The empirical reach for all 10 websites is 87.1%, while the predicted reach using the Sarmanov with 2nd order and 3rd order terms are, respectively, 90.2% and 89.2%. Hence, the predicted reach is closer to the true value when 3rd order terms are used. Repeating this test for just the first 5 websites: the empirical reach for all 5 websites is

76.3%, while the predicted reach using the Sarmanov with 2nd order and 3rd order terms are, respectively, 76.7% and 76.6%. Again, the model with 3rd order terms is closer to the true reach. This is not as rigorous a test as conducted by Danaher (2007), but it certainly indicates some improvement in reach prediction when third-order terms are included in the Sarmanov model. Moreover, the improvement in reach prediction accuracy increases as the number of websites increases.

A3: Omission of Daum.net from Optimal Schedule

The omission of the most popular site, daum.net, may seem counterintuitive at first, but closer scrutiny reveals the reason. Table A3 shows that in a one-week period daum.net attracts over 64 million page views, amounting to a reach of 70.4%. However, the CPM for this website is the highest of the 10 sites, at $15. At this CPM, the total cost of achieving the 70.4% reach is $969,821, which is nearly 20 times the budget of $50,000. For the $50,000 advertising budget, daum.net is not cost effective. The next obvious question is "What reach can I get for $50,000". The NBD can be used to answer this question, with the results being in the two rightmost columns of Table A3.

Table A3: Impressions and Reach for each Website

Website	Total impressions	Affordable impressions	Affordable Reach,%
daum.net	64,654,720	3,333,333	15.5
dreamwiz.com	7,259,028	6,250,000	11.3
msn.co.kr	3,269,306	5,000,000	12.2
chosun.com	3,477,751	5,000,000	9.2
joins.com	3,137,656	6,250,000	9.5
hani.co.kr	2,589,114	6,250,000	6.2
donga.com	2,267,304	5,000,000	7.3
naver.com	31,405,812	5,000,000	17.6
kr.yahoo.com	37,571,416	4,166,667	17.1
empas.com	12,009,395	5,000,000	16.1

It can be seen now that 3.33 million impressions can be purchased from daum.net for $50,000, and the achieved reach is 15.5%, which is now only the fourth highest of the 10 websites. When

other issues are taken into account, such the correlation between websites and repeat visits, other less-obvious websites, like msn.co.kr, suddenly become more attractive as advertising vehicles.

As a second face validity check, we repeated the optimization of reach but set the CPM of each website to be the same, at $10. The optimal reach for a budget of $50,000 is 22.6%, close to the previous value of 21.8%. The optimal schedule is given in Table A4, where it can be seen that all websites are now included in the schedule. Indeed, the highest spend allocation goes to daum.net this time.

Table A4: Optimal Schedule for Equal Website CPMs

Website	SOI	Impressions	Cost, $
daum.net	.0288	1,862,922	18,629
dreamwiz.com	.0265	192,623	1,926
msn.co.kr	.0587	191,839	1,918
chosun.com	.0466	162,224	1,622
joins.com	.0405	127,143	1,271
hani.co.kr	.0225	58,332	583
donga.com	.0288	65,228	652
naver.com	.0215	673,948	6,739
kr.yahoo.com	.0302	1,135,165	11,352
empas.com	.0442	530,575	5,306

A4. Estimating the exposure distribution empirically.

The MNBD is a model-based estimate of the exposure distribution. Historically, the accuracy of media models has been gauged by comparing such model-based estimates with observed exposure distributions obtained from individual-level data (Danaher 2007; Rust 1986). Indeed, Danaher (2007) compared nearly 3000 empirical Internet exposure distributions with those estimated by the MNBD and found the model to be very accurate. The principle of testing a model against actual data (the “true” exposure distribution) is very appealing. However, to do so in our context of Internet media selection, we need to employ a simulation method, as typically

only a share, rather than all, of the available impressions are purchased, as explained in §3.1. We now show how the simulation is operationalized.

Suppose person *n* in the raw data has I_{ni} total impressions to website *i* in a fixed time period. Suppose also that the share of impressions to site *i* in a campaign is s_i, then generate a binomial $bin(I_{ni}, s_i)$ random variable, which gives an integer between 0 and I_{ni}, denoted Y_{ni}, as the simulated total impressions[1]. Repeat this for all *m* websites, giving the total number of ad impressions for the n^{th} person as $\sum_{i=1}^{m} Y_{ni}$, which can then be used to calculate the simulated "observed" exposure distribution for the entire campaign. These simulation steps are repeated 100 times and the results averaged.

To incorporate frequency capping, the same simulation procedure can be employed, but s_i is replaced with $s_i / \Pr(X_i \le cap)$. In addition, the simulated $bin(I_{ni}, s_i / \Pr(X_i \le cap))$ random variable is now truncated at the cap, so that $0 \le Y_{ni} \le cap$.

We give an example of simulated exposure distributions in Table A5 for the three websites daum.net, dreamwiz.com and msn.co.kr. As a benchmark, the first row is the observed exposure distribution when the SOI is 100% for all three websites. The next two rows are for the cases where each website has a SOI of 10% and 20%, respectively. As would be expected, the reach and average frequency are lower when only a share of impressions is purchased. The final row of Table A5 illustrates the situation where the SOI values differ across websites (10, 20 and 30%, respectively, for sites 1 through 3). Here the reach is between that achieved when all the SOIs are 10% or 20%. The key point is that the simulation method is capable of generating model-free empirical exposure distributions for any combination of s_i values. Therefore, it is a

[1] This simulation mimics precisely the way in which ad impressions are delivered to visitors at a website.

good benchmark for evaluating both the accuracy of the MNBD model and for assessing whether the computationally-fast model-based schedule is close to that obtained from a laborious complete enumeration using just the raw data.

Table A5: Simulated Exposure Distributions for Three Websites.

SOI	Exposures											Reach	Av.
	0	1	2	3	4	5	6	7	8	9	10+		Freq
100% -all	26.0	12.2	8.3	7.1	7.8	6.5	6.1	4.8	4.0	3.1	14.0	74.0%	5.19
10% -all	71.2	19.6	6.6	2.0	0.5	0.1	0.1	0	0	0	0	28.8%	1.45
20% -all	55.3	22.7	12.2	5.6	2.4	1.1	0.5	0.2	0	0	0	44.7%	1.87
10,20,30%	68.1	20.6	7.5	2.6	0.8	0.2	0.2	0	0	0	0	31.9%	1.53

A5: Validation

Media planners typically use current or historical data to predict reach or the exposure distribution at a time period in the future (Rossiter and Danaher 1998). Therefore, validation of model accuracy in a future period is very important. Online audiences are not always stable over time, as the last two columns in Table 1 illustrate for the observed reach and average frequency values in June for the same 10 websites. The reach for each website is lower than for March, except for naver.com, which increases substantially.

Although the MNBD has already been successfully validated as an audience exposure model by Danaher (2007), in this study we still need to validate the downstream optimal advertising schedules based upon the MNBD. We do this in several ways. Table A6 reports the optimal reach values for budgets at $50,000 and $100,000 for one week campaigns with no frequency capping. The actual schedules that produce these optimal reach values, for the $50,000 budget, are given in Table 3 of the main paper. Our first validation check is to see if the model-based optimal reach is an accurate estimate of the actual reach obtained empirically (as detailed in Section A4 of the Online Appendix) in the March estimation period. Table A6 confirms that in

all cases the first two columns of numbers are very close, with the MNBD model tending to slightly overestimate the “true” empirically-derived reach.

Our second validation check compares the model-based estimate of reach in the estimation period with that derived empirically for the optimal schedule in the validation period three months later. For example, the model-based optimal reach for the fully flexible schedule is 21.8%. Table 3 shows that the optimal schedule places ads in all but the first website. If we now use that same schedule in June, using the March-derived SOI values from Table 3 in each website, the empirically-derived reach is almost identical, at 21.7 percent. The first and third numerical columns in Table A6 are always very close, with the optimal reach being a little lower in the validation period due to our earlier observation that most websites have lower reach in June compared with March. This is good evidence that our method produces optimal schedules that deliver a reach level at a later time consistent with that estimated in the initial estimation (or planning) time period.

Our final, and most challenging, validity check is to recalculate the optimal schedule derived by complete enumeration in the June period and see if the resulting reach is different to that obtained using our model-based method on the March data. That is, how does our model-based March optimal reach compare with the maximum reach achievable in June? The last column in Table A6 is the optimal reach based on the empirical complete enumeration method using just the June data. For the $50,000 budget the selected websites in the June complete enumeration are the same as obtained by our method in March. Moreover, the impressions for the fully flexible schedules are very similar. Therefore, it can be seen that the maximum reach values at this budget level are the same, being, respectively 21.9 and 19.9 percent for the fully flexible and fixed cost methods.

For the $100,000 budget, as noted previously, there are many more feasible solutions, so this time the revised optimal schedule in June is no longer the same as that based on our model in March. Table A6 shows the June optimal reach for the fixed cost method is 33.5 percent, being higher than the 33.2 percent achieved by the March optimal schedule. This time the two schedules are not the same, although they are not very different, with seven of the ten websites having the same SOI. The major point of difference between the model-based March schedule and the optimal empirical schedule in June is that daum.net has double the SOI, while msn.co.kr has half the SOI in the June schedule. Despite these differences for the larger budget, the model-based optimal schedules obtained in March are reasonably robust, since updated schedules in June have either the same reach or are just a little higher.

Table A6: Reach for the Optimal Schedules – estimation and validation periods

Model-based optimization method	Time Period			
	Estimation - March		Validation - June	
	Model – based	Empirical	Empirical -based on March schedule	Empirical -based on revised schedule for June
Budget=$50,000				
Fully flexible	21.8	21.7	21.9	21.9
Fixed cost per week	20.4	20.1	19.9	19.9
Budget=$100,000				
Fully flexible	35.0	34.7	34.7	-*
Fixed cost per week	33.6	33.4	33.2	33.5

*As the projected computation time is 174 days we did not obtain this empirical reach value.