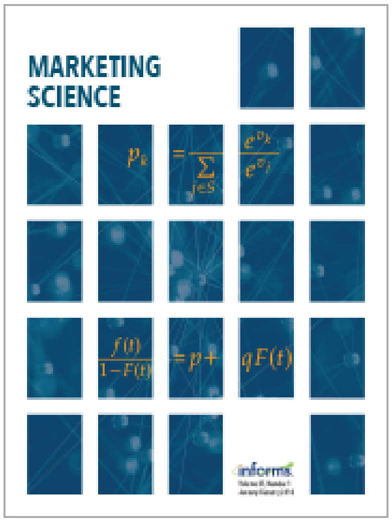


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Modeling Page Views Across Multiple Websites with an Application to Internet Reach and Frequency Prediction

Peter J. Danaher,

MARKETING SCIENCE
Vol. 26, No. 3, May–June 2007, pp. 422–437
ISSN 0732-2399 | EISSN 1526-548X | 07 | 2603 | 0422

DOI 10.1287/mksc.1060.0226

Modeling Page Views Across Multiple Websites with an Application to Internet Reach and Frequency Prediction

Peter J. Danaher
Melbourne Business School, The University of Melbourne, Carlton, Victoria 3053, Australia, p.danaher@mbs.edu

In this study, we develop a multivariate generalization of the negative binomial distribution (NBD). This new model has potential application to situations where separate NBDs are correlated, such as for page views across multiple websites. In turn, our page view model is used to predict the audience for Internet advertising campaigns. For very large Internet advertising schedules, a simple approximation to the multivariate model is also derived. In a test of nearly 3,000 Internet advertising schedules, the two new models are compared with some proprietary and nonproprietary models previously used for Internet advertising and are shown to be significantly more accurate.

Key words: advertising; Internet marketing; media; probability models
History: This paper was received July 18, 2005, and was with the author 3 months for 2 revisions; processed by Bruce Hardie.

1. Introduction

A key measure of website activity is page views, also known as page impressions or page requests, which are the number of distinct pages served to a Web user over the duration of his or her visit to a domain (Bhat et al. 2002). The underlying reason for the increasing importance of page views is that Web pages often carry banner ads or sponsored search links to other websites. Now that annual Internet advertising spending has reached $12.5 billion and is growing consistently and strongly (Internet Advertising Bureau 2006), accurate models for estimating campaign audiences are required (Meskauskas 2003). Models that produce well-known advertising audience measures such as reach and frequency are in particular demand to help give online advertising credibility alongside traditional advertising media such as television (Smith 2003).

A necessary starting point for a model of Internet advertising exposure is a probability model for page views. However, an alternative method to a probability model is an empirical distribution based on historical data. For instance, in recent years, Nielsen/Netratings and comScore Media Metrix have enlarged their Web user panels to the point where a probability model is seemingly redundant. In particular, the comScore Media Metrix data set in the United States totals 100,000 panelists, which allows an advertiser to accurately compile an empirical exposure distribution to an Internet ad campaign comprising the entire panel as well as demographic subgroups. Indeed, empirical distributions are the basis of Nielsen's and Telmar's online media planning software. In §5.1, we show that the *real* challenge for Internet media models is not to estimate the audience for a historical data set since the availability of large panels enables audience measures to be estimated empirically with high accuracy. Instead, the pertinent problem is to *predict* audiences for a time period in the future based on historical data. For example, none of the probability models considered by Leckenby and Hong (1998) are adaptable to the predictive environment. Therefore, the purpose of this research is to develop a model that can accurately predict audience exposure measures for Internet advertising campaigns by using a model for page views. In doing so, we address the limitations of previous models by tailoring our model to the Internet environment, but we retain the ability to report media exposure measures that are familiar to traditional media buyers. Internet ad campaigns often run across several websites simultaneously, with sites being similar in purpose such as travel websites, so there is a resulting correlation in page view counts among the sites (Li et al. 2002). Such correlations will affect the audience measures for Internet campaigns (Leckenby and Hong 1998) and thereby present a modeling challenge. Hence, our model of page views is multivariate without assuming independence between pairs of sites, and additionally allows for the possibility of different time periods for ad delivery on each website.

Our results show our model to be significantly more accurate than all previous models including those based in empirical distributions from large panels.

2. Modeling Page Views and Internet Advertising

There is a long history of models being used in traditional media such as television and print (Chandon 1986, Danaher 1992, Leckenby and Kishi 1982, Rust 1986). Models have been used in these media primarily to estimate *reach*, the proportion of the target audience exposed to at least one ad; *frequency*, the average number of exposures among those reached; gross rating points (GRPs), the average number of exposures (with GRPs = reach × frequency); and the *exposure distribution* (ED), the proportion of the target audience exposed to none, just one, just two, etc., ads.

2.1. Modeling Issues Specific to Online Advertising

For online media to be accepted by advertisers and advertising agencies, online publishers must also be able to apply traditional media language, particularly GRPs, to their medium (Meskauskas 2003). However, when it comes to online advertising there is a fundamental difference between the way advertising space is bought compared with offline media.[1] The primary difference is that online campaigns are often purchased on the basis of "ad impressions." An online ad impression is some form of advertisement (e.g., banner, interstitial, pop up, etc.) that is served to a website's user during the course of a visit. A typical online ad campaign might comprise 50,000 to 200,000 ad impressions (http://computer.howstuffworks.com/banner-ad.htm). While a user is surfing a particular website, he downloads different pages depending on the links he clicks. As each page is assembled, advertisements are added to the page by the site's server. Each page served could have different ads embedded within it. However, the more pages a user requests the more likely it is that he will receive several exposures to the same ad, especially if he visits the site multiple times over several weeks. Hence, the key issue in the online environment is that users determine the rate of page view delivery depending on where and how often they click on items within a session. This contrasts with traditional media, where the broadcaster/publisher controls the delivery of advertisements to its audience.

Estimating the audience for an Internet advertising campaign is further complicated by issues such as possibly having multiple ads per page, ads on just the homepage, and frequency capping whereby websites limit the number of ads served to a computer by using "cookies." Handling these issues requires data not just from the page server, but also from the ad server. User centric Web browsing data (such as the comScore Media Metrix used in this study) has only page views and no record of ads served. Hence, we model page views/impressions which are conceptually similar to ad impressions, since all ads are served on Web pages. If an advertiser is fortunate enough to additionally have data on the advertising regime, then the "ad view data" simply replaces the page view data.

A further difference between online and offline media models is the data available for model fitting. Models in traditional media are generally limited to using published figures on single-vehicle reach and pairwise duplications (Rust 1986). However, online media models usually have large-scale individual-level panel data available, as detailed in §4.2.

2.2. Previous Page View and Internet Advertising Models

To the best of our knowledge, only one page view model exists, being a multivariate discretized version of the Tobit model developed by Li et al. (2002). Their use of the Tobit model is justified because a large proportion of Web users do not visit particular sites, creating a "spike" at zero page views for each website. In addition, page views are nonnegative integers, so the Tobit must be "discretized." Last, Li et al. (2002) recognize the need to allow for correlations in page views across different website categories so they generalize the univariate Tobit model to one that has a multivariate normal distribution. They apply their model to page views of comScore Media Metrix data, as we do, but their primary purpose is to uncover patterns in browsing behavior across categories of websites like auction and portal sites and test the effects of user demographics on such browsing. Still, their model can be adapted to predicting Internet audiences, so we compare our model with their multivariate discretized Tobit in §5.4.

To date, only three nonproprietary models have been developed specifically for online advertising. Of these, the most comprehensive are Leckenby and Hong's (1998) and Huang and Lin's (2006) studies, with Wood's (1998) model essentially a curve-fitting method rather than a formal model. Leckenby and Hong compare some well-known models from offline media such as the betabinomial distribution (BBD)

[1] We distinguish between banner ads and a rapidly growing form of online advertising revenue, namely paid search advertising (IAB 2006). Paid search advertising is the primary source of revenue for Google.com, for example. Our model is best suited to the situation where banner ads are placed on a website but can accommodate sponsored links, as they are conveyed by pages served up to Web users who may view the links, then click on them.

(Metheringham 1964) and the Dirichlet-multinomial (Leckenby and Kishi 1984). To use these models, Leckenby and Hong (1998) had to artificially aggregate the panel-based website exposure data in a way that forced it into the same format as that used in offline media. Rather than restricting the number of exposures to coincide with a prespecified time period, as done by Leckenby and Hong (1998), our model allows each person's exposure level to range from zero to infinity. This is more appropriate for the Internet, where there is varying exposure opportunity per website visitor. Huang and Lin (2006) avoid the problems of Leckenby and Hong's (1998) model by allowing exposures to range upward from zero, but their model requires the duration of advertising on each website to be the same and ignores duplication of exposure between websites. Our model does not have these limitations.

Proprietary models for reach and frequency prediction include Nielsen/Netratings' "WebRF" and Telmar's "WebPlanner" models. Both use individual-level panel data to build an empirical exposure distribution. Another proprietary model is one developed by Atlas DMT (www.atlasdmt.com), which combines site-centric ad server information with comScore Media Metrix panel data (Smith 2003). No technical details about this model are available except that it is based on a simulation method, although Chandler-Pepelnjak (2004) reports that the average prediction error for reach for this model is 20%. Later, we demonstrate that this is much higher than the 5% average prediction error for our model.

3. Model Development

3.1. Notation

A formal statement of the exposure distribution (ED) setup is as follows. Let X_i be the number of exposures[2] a person has to media vehicle i, $X_i = 0, 1, 2, \ldots, i = 1, \ldots, m$, where m is the number of different vehicles. The exposure random variable to be modeled is $X = \sum_{i=1}^{m} X_i$, the total number of exposures to an advertising schedule. Although X is a simple sum of random variables, two nonignorable correlations make modeling it difficult (Danaher 1989). One is the intravehicle correlation due to repeat viewing/visits to the same vehicle (Danaher 1989, Morrison 1979) and the other is intervehicle correlation, where there might be an overlap in exposure to two vehicles.

In the case of the print media, for example, observed empirical EDs are known to be particularly "lumpy" due to strong intravehicle correlation. As a consequence, Danaher (1988, 1989, 1991) shows that it is necessary to first model the joint multivariate distribution of $(X_1, X_2, \ldots, X_m)$, from which the distribution of total exposures $X = \sum_{i=1}^{m} X_i$ can be derived. This is less of a problem with television EDs (Rust 1986) where loyalty from episode to episode is generally moderate, with intra-exposure duplication factors of the order 0.28 (Ehrenberg and Wakshlag 1987). In addition, for the television environment there are more vehicle choices than for the print medium (Krugman and Rust 1993) and this helps to reduce both intra- and intervehicle correlation. As a result, models for just X rather than the full multivariate $(X_1, X_2, \ldots, X_m)$ are often adequate for television EDs, which tend to be smooth (Rust and Klompmaker 1981). The Internet has many more possible vehicles than even television and so it might be the case that Internet EDs are also relatively smooth. Indeed, Leckenby and Hong (1998) have already noted that Internet EDs are reasonably well-behaved, that is, they tend to be smooth rather than lumpy. To assess the need for a model of the full joint distribution $(X_1, X_2, \ldots, X_m)$, in §3.5 we also develop a direct model for just X, which is an approximation to the model based on $(X_1, X_2, \ldots, X_m)$.

3.2. One-Vehicle Model

In the case of Internet advertising, X_i is the number of ad/page impressions a person has to website i, $i = 1, \ldots, m$. Over the course of a fixed time period, X_i can range from 0 to infinity as there is no limit to the number of times a site can be visited and how many pages can be viewed.

Modeling the number of exposures a person has to website i in a fixed time period (say, a month) is analogous to the well-known problem in marketing of modeling the number of purchases a person has in a product category (say, in a year or month for a frequently purchased product). Goodhardt et al. (1984) and Ehrenberg (1988) show that the number of purchases, given a purchase rate λ_i, is modeled accurately by a Poisson distribution with mean λ_i. That is, the number of purchases follows a Poisson process,[3] conditional on a person's purchase rate λ_i. To allow for heterogeneity in λ_i across individuals, we

[2] For Internet advertising, by "exposure" we mean an ad impression. As mentioned earlier, since our data do not have explicit information on ad impressions, only page views/impressions, throughout this paper we consider an exposure and a page impression to be synonymous.

[3] Website users can accrue page impressions over a month-long period by having multiple visits to the site and/or spending more time on a site during their visits. In the category purchase situation, this is analogous to multiple store trips and greater quantity purchased at each trip, respectively. While it might be of interest to model the *way* in which page impressions are accumulated, it is not necessary in this application. A sufficient statistic for reach and frequency estimation is simply the *total* number of page impressions.

assume λ_i comes from a gamma distribution, as was also proposed by Ehrenberg (1988) and Morrison and Schmittlein (1988) to allow for heterogeneity in purchase rates in packaged goods. Compounding the $X_i \mid \lambda_i \sim Poisson(\lambda_i)$ with a gamma distribution results in the well-known negative binomial distribution (NBD) with mass function

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

where r_i and α_i are the usual parameters for the gamma distribution. Hence, a reasonable model for ad exposures to a single website is the NBD (see also Huang and Lin 2006). We later show that it fits observed EDs very well.

3.3. Two-Vehicle Model

While the NBD is a natural model for one website and derives easily by making an analogy between website page impressions and category purchases, an extension to two websites is less obvious. Although Chintagunta and Haldar (1998) have studied purchase timing across two categories, there are no models for the total number of purchases across two product categories. A naïve approach is to assume that page impressions across two websites are statistically independent. The bivariate mass function is then simply the product of the two marginal mass functions. Given the vast number of websites, many of which have relatively few visitors, this may seem like a reasonable assumption. However, we obtained the pairwise correlations between the top 45 websites in our data and found many instances of reasonably high correlations. For example, the correlation between the two Web hosting sites angelfire.com and tripod.com is 0.56, while that between msn.com and msnbc.com is 0.22. Correlations of this order make the independence assumption questionable.

Park and Fader (2004) faced a similar issue to ours when examining the visit times between two websites. They use an exponential-gamma model for a single website and initially assume independence between the sites. Like us, they find the independence assumption to be unsupported by their data. Instead, they develop a model in which the marginal distributions are still exponential gamma, but allowance is made for a correlation between the univariate random variables. The bivariate model employed by Park and Fader (2004) belongs to a general class of bivariate distributions developed by Sarmanov (1966) and first applied by Lee (1996) in the statistics literature.[4] The general form of the Sarmanov bivariate distribution for (X_1, X_2) is

$$f(X_1, X_2) = f_1(X_1) f_2(X_2)[1 + \omega\phi_1(x_1)\phi_2(x_2)], \qquad (2)$$

where $f_i(X_i)$ is the marginal distribution for random variable X_i and $\phi_i(x_i)$ are called "mixing functions," with the requirement that $\int \phi_i(t) f_i(t)\, dt = 0$. Notice that the general form of this bivariate distribution is the product of the marginal distributions, with a "correction factor" to allow for correlation. This general bivariate distribution clearly has strong appeal in our application, where we would ideally like to have a bivariate model for page impressions that has NBD marginal distributions to retain the modeling accuracy and simplicity of the univariate model. Notice, however, that our bivariate model is different from that of Park and Fader (2004), since theirs has exponential-gamma marginals while ours has NBD marginals. Furthermore, they report only the bivariate case whereas we extend the bivariate to a multivariate version of the Sarmanov distribution to Web ad campaigns with up to 15 websites.

There is often a choice for the mixing functions, but Lee (1996) recommends the appropriate mixing function for the NBD to be

$$\phi_i(x_i) = e^{-x_i} - \left(\frac{\alpha_i}{1 + \alpha_i - e^{-1}}\right)^{r_i}. \qquad (3)$$

Now, substituting Equation (3) into Equation (2) we obtain a model for the bivariate distribution of (X_1, X_2) as

$$f(X_1, X_2) = f_1(X_1) f_2(X_2)\left[1 + \omega\left(e^{-x_1} - \left(\frac{\alpha_1}{1+\alpha_1-e^{-1}}\right)^{r_1}\right) \cdot \left(e^{-x_2} - \left(\frac{\alpha_2}{1+\alpha_2-e^{-1}}\right)^{r_2}\right)\right], \qquad (4)$$

where $f_i(X_i)$, $i = 1, 2$ are NBD distributions with parameters r_i and α_i, as given in Equation (1). Due to the property of the mixing functions, namely $\sum_t \phi_i(t) f_i(t) = 0$, it is easy to verify that the marginal distributions of (X_1, X_2) in Equation (4) are NBD distributions. This bivariate advertising exposure model has the same functional form as one previously developed in the print media by Danaher (1991), namely the canonical expansion model. Danaher's

[4] When modeling page views across multiple websites, we can not simply adapt Park and Fader's (2004) model since theirs is developed for just two websites, whereas Internet ad campaigns can have many more than two sites. Our model can be applied to any number of websites, with the empirical results showing good predictive accuracy for as many as 15 websites in a campaign. In addition, they model intervisit times and visit rates, but we require a model for the count of the number of pages viewed at each of several websites. Hence, there is a fundamental difference in the underlying random variables between our application and that of Park and Fader (2004).

(1991) model is a product of univariate marginal distributions with a correction factor to account for correlation in exposure between media vehicles. Another way to construct Equation (4) is to use Equation (2) to build a bivariate gamma distribution and then mix it with two conditionally independent Poisson distributions.

Lee (1996) gives a general expression for the correlation between random variables in the Sarmanov distribution. In the case of the bivariate NBD distribution in Equation (4), the correlation between X_1 and X_2 is

$$\begin{aligned}\operatorname{corr}(X_1, X_2) &\\ &= \omega(1-e^{-1})^2 \frac{\sqrt{r_1 r_2 (1+\alpha_1)(1+\alpha_2)}}{\alpha_1 \alpha_2} \\ &\quad \cdot \left(\frac{\alpha_1}{1+\alpha_1-e^{-1}}\right)^{r_1+1} \left(\frac{\alpha_2}{1+\alpha_2-e^{-1}}\right)^{r_2+1}. \quad (5)\end{aligned}$$

Therefore, X_1 and X_2 are independent if and only if $\omega = 0$, so the parameter ω largely determines the correlation between exposure to two websites.

3.4. Model for Three or More Vehicles

To extend the bivariate model in Equation (4) to m vehicles, we again draw on the Sarmanov model to allow for correlation in advertising exposures among websites. Lee (1996) provides a generalization of the bivariate Sarmanov model to m variates, being

$$\begin{aligned} f(X_1, X_2, \ldots, X_m) &\\ &= \left\{\prod_{i=1}^{m} f_i(X_i)\right\} \left[1 + \sum_{j_1<j_2} \omega_{j_1,j_2} \phi_{j_1}(x_{j_1}) \phi_{j_2}(x_{j_2}) \right. \\ &\qquad \left. + \cdots + \omega_{1,2,\ldots,m} \prod \phi_i(x_i)\right]. \quad (6)\end{aligned}$$

Equation (6) is a series expansion of bivariate, trivariate, up to m-variate terms. Estimating parameters for such a model would require observed multivariate duplications among the m websites. To reduce the number of terms in the multivariate Sarmanov model, we truncate Equation (6) after just the trivariate terms. This gives an approximation to the full Sarmanov expansion, with accuracy up to third-order terms. This means that bivariate and trivariate exposure interactions among the websites are modeled, but fourth- and higher-order interactions are not explicitly modeled. Again, this is very similar to the approximation made by Danaher's (1991) canonical expansion model, but he truncated the full canonical expansion after just the second-order terms. We show later that even after truncating the multivariate Sarmanov model, the resulting website exposure model is still very accurate.

The multivariate model with truncation after third-order terms is

$$\begin{aligned} f(X_1, X_2, \ldots, X_m) &\\ &= \left\{\prod_{i=1}^{m} f_i(x_i \mid r_i, \alpha_i)\right\} \\ &\quad \cdot \left[1 + \sum_{j_1<j_2} \omega_{j_1,j_2} \left(\exp(-x_{j_1}) - \left(\frac{\alpha_{j_1}}{1+\alpha_{j_1}-e^{-1}}\right)^{r_{j_1}}\right)\right. \\ &\quad \cdot \left(\exp(-x_{j_2}) - \left(\frac{\alpha_{j_2}}{1+\alpha_{j_2}-e^{-1}}\right)^{r_{j_2}}\right) \\ &\quad + \sum_{j_1<j_2<j_3} \omega_{j_1,j_2,j_3} \left(\exp(-x_{j_1}) - \left(\frac{\alpha_{j_1}}{1+\alpha_{j_1}-e^{-1}}\right)^{r_{j_1}}\right) \\ &\quad \cdot \left(\exp(-x_{j_2}) - \left(\frac{\alpha_{j_2}}{1+\alpha_{j_2}-e^{-1}}\right)^{r_{j_2}}\right) \\ &\quad \left.\cdot \left(\exp(-x_{j_3}) - \left(\frac{\alpha_{j_3}}{1+\alpha_{j_3}-e^{-1}}\right)^{r_{j_3}}\right)\right]. \quad (7)\end{aligned}$$

Since this model is a multivariate generalization of the NBD and has NBD marginals, we name it the *multivariate negative binomial distribution* (MNBD).

Our model for X, the total number of exposures, can now be obtained from the multivariate distribution for $(X_1, X_2, \ldots, X_m)$ by summing over the relevant probabilities as follows:

$$f_X(x) = \sum_{\{(x_1,\ldots,x_m):\, x_1+\cdots+x_m=x\}} f(X_1, X_2, \ldots, X_m), \quad x = 0, 1, 2, \ldots. \quad (8)$$

3.5. Model for Very Large Schedules

For advertising schedules with a very large number of sites ($m \geq 10$), the multivariate NBD model for X given by Equation (8) takes a lot of computation time, making it impractical. A solution to this difficulty arises from a convenient additive property of the Poisson distribution, as we now explain.

Recall that we assume the conditional univariate distribution $X_i \mid \lambda_i$ is a Poisson distribution with mean λ_i. In the multivariate case, the corresponding assumption is that $X_1, X_2, \ldots, X_m$ are conditionally independent Poisson distributions, with respective parameters $\lambda_1, \lambda_2, \ldots, \lambda_m$. It is easy to show that conditional on $(\lambda_1, \lambda_2, \ldots, \lambda_m)$, $X = \sum_{i=1}^{m} X_i \sim Poisson(\sum_{i=1}^{m} \lambda_i)$. Remembering that X and not $(X_1, X_2, \ldots, X_m)$ is our ultimate modeling objective, the additive property of the Poisson distribution shows it is possible to get a direct model for the ED without first estimating a model for the full multivariate distribution. The trade-off is a reduction in prediction accuracy, as will be seen later.

We can think of $\sum_{i=1}^{m} \lambda_i$ as simply another parameter, denoted by λ. Then, as for the univariate NBD

Table 1 Data and Parameter Estimates for an Online Advertising Schedule

	Univariate parameter estimates							Pairwise parameters			
	Estimation period (September)			Validation Period (November)				netflix.com		travelzoo.com	
Website	Page impressions	$\hat{r}_i$	$\hat{\alpha}_i$	Page impressions	δ	$\hat{\alpha}_i$	Non-reach $\hat{f}_i(0)$	f_{00}	$\hat{\omega}$	f_{00}	$\hat{\omega}$
aol.com	97,701	0.0922	0.0094	104,970	1.07	0.0088	0.6501	0.586	0.847	0.614	1.736
netflix.com	2,449	0.1091	0.4453	5,768	2.36	0.1891	0.8795	—	—	0.812	1.886
travelzoo.com	1,293	0.1362	1.053	2,129	1.65	0.6397	0.9131	—	—	—	—

model, to allow for individual-level heterogeneity we assume that λ has a gamma distribution with parameters r^* and α^* resulting in $X \sim NBD(r^*, \alpha^*)$.[5] Hence, the mass function for the large-schedule model is

$$f_X^*(x) = f^*(x \mid r^*, \alpha^*), \tag{9}$$

where $f^*(\cdot)$ denotes the NBD mass function as given by Equation (1). Since this model is intended to approximate the full multivariate model in Equation (8), we subsequently refer to it as the approximate NBD (ANBD).

3.6. Parameter Estimation—Full Model

It can be seen from Equation (7) that only univariate marginals and bi- and trivariate associations are needed to model the joint probability of $(X_1, X_2, \ldots, X_m)$. Indeed, the model in Equation (7) has NBD marginals, each with parameters r_i and α_i, $i = 1, \ldots, m$. Hence, a straightforward method for estimating r_i and α_i for each website is to fit a univariate NBD separately to each website's observed univariate ED. As Goodhardt et al. (1984) show, an efficient method for estimating the parameters of the NBD is the method of means and zeros (Anscombe 1950).[6] Here, the observed sample mean number of exposures to website i is fit to the parametric mean of each NBD, namely r_i/α_i, and the observed nonreach (0th frequency of each univariate ED) is fit to $f_i(0 \mid r_i, \alpha_i) = (\alpha_i/(1+\alpha_i))^{r_i}$, giving estimates $\hat{r}_i$ and $\hat{\alpha}_i$. A straightforward iterative Newton-Raphson method is required for this estimation.

The remaining parameters to be estimated are the second- and third-order measures of association among the websites, ω_{j_1, j_2}, $1 \le j_1 < j_2 \le m$ and ω_{j_1, j_2, j_3}, $1 \le j_1 < j_2 < j_3 \le m$. We start by discussing several possible methods for estimating ω_{j_1, j_2}, all of which we tried. The first is to use the observed bivariate exposure distribution and estimate each ω_{j_1, j_2} by maximum likelihood. The second is to use a method of moments-type estimator and equate the empirical correlation to the parametric correlation given in Equation (5). As each r_i and α_i have already been estimated, an estimate of ω_{j_1, j_2} is obtained by solving for ω_{j_1, j_2} in Equation (5). A third possible estimator has its roots in the print media industry (Chandon 1986) and is based on the unduplicated reach, meaning the proportion of people exposed to either one or both sites (any number of times) within a given time interval. For websites, this is now routinely being reported by ACNielsen's Nielsen/Netratings software, for example. One minus the unduplicated reach is the nonreach, which can be used to estimate the pairwise associations, as we now show.

Consider for the moment just the bivariate model, as given by Equation (4). The nonreach for two vehicles is

$$f(0,0) = f_1(0) f_2(0)[1 + \omega \phi_1(0) \phi_2(0)], \tag{10}$$

where

$$f_i(0) = f_i(0 \mid r_i, \alpha_i) \quad \text{and}$$

$$\phi_i(0) = 1 - \left(\frac{\alpha_i}{1 + \alpha_i - e^{-1}}\right)^{r_i}, \quad i = 1, 2.$$

Now obtain the observed nonreach of two website vehicles and denote this $\hat{f}(0,0)$. Using the $\hat{r}_i$ and $\hat{\alpha}_i$ from the univariate ED estimation stage, from Equation (10) we obtain a consistent estimate of ω as

$$\hat{\omega} = \left[\frac{\hat{f}(0,0)}{\hat{f}_1(0)\hat{f}_2(0)} - 1\right] \frac{1}{\hat{\phi}_1(0)\hat{\phi}_2(0)}. \tag{11}$$

This method of parameter estimation is used for all $m(m-1)/2$ pairwise combinations of websites to obtain $\hat{\omega}_{j_1, j_2}$. We illustrate the estimation of $\hat{\omega}_{j_1, j_2}$ with an example comprised of the websites aol.com, netflix.com, and travelzoo.com as given in Table 1 (full details on the data are given later in §4.1). The univariate NBD parameters are estimated by the method of means and zeros, while the bivariate $\hat{\omega}_{j_1, j_2}$ parameters are estimated with Equation (11). For example, the estimate of ω for the aol.com and netflix.com pairing is

$$\hat{\omega} = \left[\frac{0.586}{(0.6501)(0.8795)} - 1\right] \frac{1}{(0.3222)(0.0919)} = 0.847.$$

[5] It is *not* the case that we are assuming that each λ_i has a gamma distribution and that their sum is another gamma distribution. We initially set $\lambda = \sum \lambda_i$ only as a parameter and then later allow for heterogeneity in λ by permitting it to come from a flexible distribution, being the gamma distribution in this case.

[6] We also fit the NBD by maximum likelihood and found very little difference in parameter estimates and model fit.

To estimate the third-order association parameters, ω_{j_1,j_2,j_3}, $1 \le j_1 < j_2 < j_3 \le m$, we use an analogous argument as for the second-order terms. The starting point is the trivariate Sarmanov model at the zero exposure levels for all three websites, which is

$$f(0,0,0) = f_1(0)f_2(0)f_3(0) \cdot [1 + \omega_{12}\phi_1(0)\phi_2(0) + \omega_{13}\phi_1(0)\phi_3(0) + \omega_{23}\phi_2(0)\phi_3(0) + \omega_{123}\phi_1(0)\phi_2(0)\phi_3(0)]. \tag{12}$$

Substituting the respective second-order estimates from Equation (11) into Equation (12) gives a consistent estimate of the third-order association:

$$\hat{\omega}_{123} = \left[\frac{\hat{f}(0,0,0)}{\hat{f}_1(0)\hat{f}_2(0)\hat{f}_3(0)} - 1 - \hat{\omega}_{12}\hat{\phi}_1(0)\hat{\phi}_2(0) - \hat{\omega}_{13}\hat{\phi}_1(0)\hat{\phi}_3(0) - \hat{\omega}_{23}\hat{\phi}_2(0)\hat{\phi}_3(0)\right] \cdot \frac{1}{\hat{\phi}_1(0)\hat{\phi}_2(0)\hat{\phi}_3(0)}. \tag{13}$$

A very appealing feature of the estimators in Equations (11) and (13), which is not shared by the first two estimators discussed above, is that they result in very accurate estimates of reach, which is the most important measure of advertising audience size. This is because reach is one minus the nonreach, and the estimators in Equations (11) and (13) ensure an exact match between the model estimate and observed values of bivariate and trivariate nonreach. For four or more websites, the match is no longer exact but it is closer than we obtained for the first two estimators described above. Moreover, the estimators in Equations (11) and (13) are superior empirically when compared against maximum likelihood and method-of-moments estimators.

3.7. Parameter Estimation—Approximate Model

To estimate the two parameters of the ANBD model, we follow the common practice in marketing of using the method of means and zeros for the NBD (Goodhardt et al. 1984). First, we need a good estimate of the probability of zero exposures, i.e., nonreach. This can be obtained from the estimated nonreach given by the MNBD in Equation (7). Hence, the MNBD is a crucial component of the ANBD. Second, the observed mean of X is simply the sum of the observed mean exposures for each website. With these estimates of mean exposures and nonexposure, estimates of r^* and α^* can be obtained.[7]

4. Data

4.1. Estimation Data

We use comScore Media Metrix data to fit and test our model. comScore employs a user-centric method of measurement, whereby a panel of Internet-enabled homes have all their Web browser activity monitored (see Coffey 2001 for particular details relating to Media Metrix methodology). Panelists install proprietary software on their computers that unobtrusively captures the URL, page impressions, and visit duration for each URL as they proceed through their Internet sessions. The Web activity is tied to a particular machine in the home rather than a particular person, allowing for the possibility of multiple machines within the same home. The data are aggregated to the domain level, enabling reporting of the total number of page impressions within that domain. In addition, if several Internet browsing sessions occur on the same day, data are aggregated across that day. Moe and Fader (2004a, b) and Park and Fader (2004) also use comScore data aggregated to the daily level.[8] It is important to note that we do not use click stream data and do not have information on the navigation history within a website. Such detailed information is not required in our application, even if it were available.

Our particular comScore data come from the Wharton Research Data Service (www.wrds.upenn.edu), which stores a six-month sample of comScore panelist records from the United States for the months July through December 2002 and is available by subscription for research purposes. The panel size comprises 100,000 machines located within panelists' homes.[9] At recruitment, each household reports a number of demographic variables, including the education level of the most educated person, household size, region, age of oldest member, income, presence of children, race, and the speed of the household's Internet connection. There is a long history of media planners targeting advertising campaigns at particular demographic groups (Cannon 2001, Gal-Or and Gal-Or 2005, Iyer et al. 2005), which enhances the appeal of user-centric data, like comScore's panel.

As the data quantity is enormous, comprising over 120 MB each day, we take a representative subset of the full data set, as we now detail. A panel of 100,000 machines is clearly very large and a smaller panel size for a shorter time period could easily suffice without much loss of accuracy. Hence, we take a systematic random sample of every tenth panelist just for the

[7] This parameter estimation method ensures that the reach predictions for the multivariate NBD in Equation (7) will agree identically with those of the approximation given by Equation (9).

[8] Park and Fader (2004), in particular, report that daily aggregations are not problematic as multiple visits to the same site on the same day are not common for the vast majority of websites.

[9] Even though a house may have more than one machine, we use the terms machine and household interchangeably.

Table 2 Top 45 Websites and Their Reach and Frequency for Various Demographics

		All households, $n = 10{,}000$		College educated, $n = 2{,}865$		Household with children, $n = 4{,}420$	
Website name	Site type	Reach, %	Average frequency	Reach, %	Average frequency	Reach, %	Average frequency
about.com	portal	8.8	4.5	10.1	4.2	10.1	4.6
aim.com	messaging	4.6	3.5	5.5	3.7	6.1	3.6
altavista.com	search	4.6	8.1	5.2	10.2	5.0	7.2
amazon.com	retail	12.8	5.2	14.3	5.7	13.8	5.7
angelfire.com	hosting	8.6	4.6	9.1	5.5	10.8	4.8
aol.com	portal	35.0	27.9	36.2	28.4	38.1	29.1
ask.com	search	5.0	5.6	5.0	5.7	5.8	4.9
bonzi.com	software	2.6	8.9	2.2	7.8	2.6	7.9
cnet.com	service	3.6	2.4	4.5	2.7	3.5	2.4
cnn.com	news	10.3	7.5	13.1	9.0	10.6	7.7
ebay.com	auction	21.7	33.8	21.7	33.0	23.1	33.2
excite.com	portal	4.6	21.5	4.6	21.1	4.9	23.2
expedia.com	travel	6.9	6.1	9.1	6.4	6.2	6.0
ezboard.com	messaging	3.5	14.4	3.9	8.4	4.4	13.9
flowgo.com	greetings	5.6	4.3	4.2	2.4	4.8	3.6
gamespot.com	games	1.9	8.2	2.7	9.7	2.6	7.4
gator.com	software	25.1	24.6	24.4	26.9	26.9	23.1
go.com	portal	14.5	14.8	17.3	15.9	16.4	14.4
gohip.com	portal	1.8	27.0	1.2	23.5	2.2	24.0
google.com	search	22.7	13.5	26.9	16.2	24.5	13.4
hotbar.com	software	3.7	75.3	2.6	78.6	3.6	77.6
hotmail.com	e-mail	7.2	41.2	7.1	29.4	7.6	37.0
iwon.com	lottery	6.7	48.0	6.7	42.6	6.6	41.6
kazaa.com	music	11.1	10.3	11.7	10.1	13.3	10.8
looksmart.com	search	2.6	4.4	2.8	3.5	2.4	4.3
lycos.com	portal	19.0	12.9	19.5	10.7	21.3	13.5
mamma.com	search	5.6	2.8	5.6	2.9	7.1	2.6
mcafee.com	software	3.0	16.0	3.3	17.5	3.0	14.6
msn.com	portal	51.8	45.9	52.9	46.1	52.8	44.9
msnbc.com	news	10.2	4.7	12.1	4.9	9.9	4.5
nascar.com	sports	2.0	10.7	1.2	10.5	2.2	12.6
neopets.com	entertainment	1.4	90.9	1.3	83.7	1.9	97.5
netflix.com	movies	12.1	2.0	11.3	2.3	12.0	2.2
netscape.com	portal	12.1	14.2	13.3	14.7	13.1	14.5
nytimes.com	news	4.6	4.9	6.9	6.7	4.6	5.2
reunion.com	people search	8.2	1.8	7.3	1.9	8.5	1.8
sportsline.com	sports	4.0	27.8	5.2	21.6	4.1	31.2
travelzoo.com	travel	8.7	1.5	8.5	1.5	9.8	1.5
tripod.com	hosting	9.8	3.8	9.7	4.3	11.2	4.2
weather.com	weather	8.1	5.3	8.8	4.5	8.4	5.5
weatherbug.com	weather	14.4	12.1	12.3	11.4	13.6	10.6
webmd.com	health service	3.1	3.7	3.2	3.8	3.2	3.8
webshots.com	software	4.3	10.5	3.6	12.4	4.1	10.1
x10.com	electronics	2.8	1.6	2.1	1.8	3.0	1.6
yahoo.com	portal	63.2	49.3	62.7	47.2	64.8	49.7

month of September 2002, resulting in a panel subset of 10,000 machines. We chose September because it is not affected by summer holiday behavior and it has 30 days. Later we will use a randomly selected validation sample of 10,000 different machines from November 2002 which also has 30 days, thereby keeping the estimation and validation periods the same in length. Our intention is to fit the model to the September data, then use it to predict website ad campaign reach and frequency distributions for November. For the full data set, over 130,000 different websites are visited by the panelists each day but only 13% of these are visited by more than 10 panelists, with around one-half of sites visited just once. Therefore, we select just the 45 most visited websites, which are listed alphabetically in Table 2 and receive substantial Web traffic. Savage and Waldman (2004) report that the top 45 websites attract almost 80% of Internet advertising revenue which makes them especially relevant to our application.

Table 2 gives the reach for each website for the full subset of 10,000 households and for two demographic subsets, those where the most educated household member has a college degree and for households with children. It can be seen that yahoo.com has the highest reach by some way, even among this group of popular websites. There is not much variation in its reach across the demographic groups, showing its broad appeal. Contrast this with the children's site, neopets.com, which has a reach of 1.4% among all households but its reach is nearly 40% higher, at 1.9%, in households with children. In addition to reach, Table 2 gives the average frequency, in other words, the mean number of page impressions for the month of September among those reached by the site. neopets.com is a good example of a website with low reach but high average frequency, meaning its users consume a high number of page impressions.

4.2. Validation Method and Data

Previous models used to estimate advertising campaign EDs in traditional or online media have used the estimation data set to also evaluate the fit of the model (Chandon 1986; Danaher 1991, 1992; Leckenby and Kishi 1984; Leckenby and Hong 1998; Rust 1986). In the past, media model researchers have used part of the data such as first-insertion reach for each vehicle and bivariate reach across all pairs of vehicles to fit their models, then predicted the entire ED from this subset of the data. This is necessitated by media audience suppliers such as SMRB reporting only this limited amount of data in the print industry (Rust 1986). In our application to online media, we have the entire data set at hand so there is no such limitation. In this situation, it is possible to obtain an empirical estimate of the ED by counting the number of households with none, just one, just two, etc., page impressions to a website or a combination of websites, seemingly making it unnecessary to use a model. However, we show later that our model is able to outperform an empirical estimate of the ED.

Media planners typically use data from an estimation period to predict the audience size to a campaign for a future time period. In the case of television, the time lag might be as much as six months to a year from the time of prediction to the actual airing of commercials (Katz 2003). Internet advertising lead times are much shorter, however. Due to the six-month duration and large size of the available comScore data set, we are able to simulate a more challenging prediction environment than has been possible for traditional media. A rigorous test of ours and other models is to use an estimation data set for model fitting, but then predict the reach and ED for a different group of households in a future time period. As mentioned above, we use a random sample of 10,000 panelists for the month of September 2002 for estimation, then pick a different random sample of 10,000 panelists for November 2002. This is more in keeping with what happens in media planning, where *sample* data are employed to predict future media audiences for the downstream *population*. In our case, the September data comprise the estimation sample, while the November validation sample of different people can be viewed as the "population."

As explained in §2.1, the way that Internet campaigns work is that a buyer purchases a fixed number of page/ad impressions from an online publisher. Visitors to the publisher's site are served up pages (with the embedded ads) in time order until the purchase quota is expended. Since users of the site have different visit behavior, the likelihood is that different visitors "consume" a different number of page/ad impressions over the duration of the campaign. This can create problems for advertisers, who may be trying to achieve a reach target but then find that purchasing a large number of impressions results in a small group of people being served the majority of the ad impressions (Chandler-Pepelnjak 2004), resulting in a shortfall in the actual reach delivered.[10]

Consider an example from our data set in which the total page impressions for travelzoo.com in September is 1,293 for the 10,000 panelists. In November, the total page impressions increase markedly to 2,129 for the same panel size of 10,000 but different people. The respective observed reach values for these two time periods are 8.7% and 12.4%. To enable us to use the full 30 days of validation data, we assume that an advertiser purchases 2,129 impressions on travelzoo.com. We then obtain the empirically derived reach and ED values for the validation data. Hence, in the case of travelzoo.com we use the estimation data for model fitting, which has only 1,293 page impressions, and somehow extend this to 2,129 page impressions to get an accurate prediction of reach in the validation sample. This is detailed in §5.1. To adjust our model for different page impression totals, it turns out that the NBD has a very convenient mathematical property that allows us to make this extension using just a simple rescaling of one of its parameters, as we now demonstrate. This property is not shared, for example, by the betabinomial distribution, which is a very common media model in everyday practice (Leckenby and Kishi 1982) and has been previously used by Leckenby and Hong (1998) for online campaigns.

4.3. Parameter Modification for Varying Page Impressions

We now show how the α parameter of the NBD and the mixing functions of the Sarmanov model require

[10] A practical solution to this problem is to impose "frequency capping," but this depends on the use of cookies.

only a straightforward modification to allow for different page impressions across the estimation and validation data sets.

For a single website, let X be the number of page impressions served in the time interval $[0, T_E]$ for the estimation data set. For this total number of page impressions in the estimation sample, denote the sample mean number of page impressions per panelist as $\bar{X}_E$. Our NBD model parameterizes this mean number of page impressions per person as λ, with the NBD model assuming that $X \mid \lambda \sim Poi(\lambda)$ and that $\lambda \sim gamma(r, \alpha)$.

In the prediction data set, with time interval $[0, T_P]$, the mean number of page impressions per person is likely to change either due to more or less website activity or the fact that the prediction sample are completely different panelists. We parameterize the new mean as $\delta\lambda$, that is, we simply multiply the mean from the estimation data set by δ, with $\delta > 0$. Due to the additive property of the Poisson distribution, it is not difficult to prove that for $X \mid \delta\lambda \sim Poi(\delta\lambda)$ and $\lambda \sim gamma(r, \alpha)$, the unconditional distribution of X is $NBD(r, \alpha/\delta)$. Hence, the r parameter is the same across the NBD models applied to each respective data set but α changes to α/δ in the prediction data set. Lilien et al. (1992, p. 34) demonstrate the same property of the NBD when it is applied to different observation periods of unequal duration.

We denote the sample mean number of page impressions in the prediction period as $\bar{X}_P$, and then we can equate the respective sample means to their parametric means under the NBD models. This gives $r/\alpha = \bar{X}_E$ and $r/(\alpha/\delta) = \bar{X}_P$, which implies that $\delta = \bar{X}_P/\bar{X}_E$. For example, in our data set the number of page impressions for amazon.com in the November prediction month is about double that of the estimation month of September. Therefore, when applying the NBD to the November data, all that is required is to use the same parameter estimates for September but divide the α parameter by two. A similar adjustment is required for the mixing functions[11] in Equation (3), where the α parameter is replaced by α/δ.

5. Test of the Multivariate NBD Model

In this section, we describe how the estimation data can be used to derive a simple empirical prediction of the ED. This empirical model is then compared with the newly developed models for ad campaigns ranging from 1 to 15 websites. We also show how the NBD model reveals how the reach for a website increases over the duration of a campaign.

5.1. Empirical Model for ED Prediction

The standard method for evaluating estimation accuracy for reach and frequency models is to compare the observed ("true") ED to that estimated by a model (Chandon 1986; Danaher 1988, 1989, 1991, 1992; Leckenby and Hong 1998; Leckenby and Kishi 1984; Rust 1986; Rust and Leone 1984). Observed EDs are easily obtained from individual-level data by counting the total number of exposures each person has to the vehicles which comprise the campaign and then aggregating over individuals (Leckenby and Hong 1998).

As explained in §4.2, the availability of only partial information in traditional media data sets has required that models be developed to "fill in the gaps" left by the partial data. However, for Internet campaigns, panel data suppliers like comScore and Nielsen/Netratings provide enough data to enable empirical EDs to be developed, seemingly making models redundant. As explained above, a model *is* required for Internet campaigns to predict the ED for a future time period when the visit behavior of Web users may not be the same as for the estimation period, as we now illustrate.

Returning to our earlier three-website example, Table 1 shows that the total page impressions for each site is higher in November than for September. That is, across two periods of identical duration (30 days), the "consumption" of pages is higher in the second period. Suppose an advertiser purchases 104,970, 5,768, and 2,129 page impressions, respectively, on the sites aol.com, netflix.com, and travelzoo.com. Advertising planning is based on the estimation period in September but the actual campaign goes "live" in the November validation period. Hence, it is more appropriate to predict the ED based on the November total page impressions. Table 3 gives two observed distributions, respectively, for the estimation and validation periods. The two distributions are clearly not the same even though they both span 30 days, with the November ED placing greater weight on higher page impressions. One of the reasons the two distributions are not the same is that the page impressions are not equivalent over the two time periods. They are also based on different panelists. To get a comparable estimate of the ED for the two periods, there should be an identical number of page impressions in the estimation and validation data. This is easily achieved by appending October data to the estimation data set for the original 10,000 estimation panelists. For instance, for aol.com, Table 1 shows that by the end of September some 97,701 page impressions have been consumed by the estimation panelists. For that same group of panelists, we now run into October, accumulating enough additional page impressions until exactly 104,970 page

[11] We make no modification of the ω parameters because they are measures of association and we find it is reasonable to assume they are constant across the September and November time periods.

Table 3 Comparison of the Observed and Estimated EDs for the Example Online Schedule in Table 1

Page impressions	Observed distribution, %		Model			
	Estimation	Validation	MNBD	ANBD	Independent	Empirical
0	55.7	49.9	51.6	51.6	46.5	54.5
1	10.7	8.8	10.1	7.8	12.4	11.7
2	4.4	5.8	5.2	4.4	6.7	4.6
3	2.7	3.6	3.6	3.2	4.4	2.7
4	1.9	3.1	2.7	2.5	3.1	2.0
5	1.6	2.5	2.1	2.0	2.4	1.6
6	1.4	1.8	1.8	1.7	1.9	1.3
7	1.1	1.5	1.5	1.5	1.5	1.2
8	1.0	1.0	1.3	1.3	1.3	1.0
9	0.9	1.0	1.1	1.2	1.1	0.9
10	0.8	0.8	1.0	1.1	0.9	0.6
11+	17.8	20.2	18.0	21.7	17.8	17.9
Number of parameters	—	—	10	12*	6	0**
RER, %	—	—	3.4	3.4	6.7	9.2
EPOR	—	—	12.0	14.7	20.9	27.8

*The ANBD has 2 parameters, as given in Equation (9), plus 10 further parameters that are used to estimate nonreach, which is obtained from the MNBD. **The empirical-based exposure distribution has no parameters.

impressions have been "consumed." In the case of aol.com, we had to append the first two days in October to achieve the required total number of impressions. The same "over run" is done for the other two websites so that they too have the same total page impressions as for the validation period.[12] From this augmented September/October estimation data, we are able to obtain an empirical estimate of the future ED. Indeed, this appears to be the method used by Nielsen/Netratings; The description of their reach and frequency planning tool, WebRF, states that "WebRF uses actual respondent-level data acquired by measuring users' online activity, so there's no modeling—just real results" (www.netratings.com). The last column in Table 3 shows this empirical distribution for our example campaign. Comparing it with the ED from the validation period shows that it overestimates the zeroth and first exposure levels, meaning it underestimates reach. The "true" reach is 50.1% whereas the empirical method estimates it at 45.5%. Contrast this with the MNBD model, which gives a much closer estimate of reach at 48.4%.

5.2. Test Design

Predicted EDs are obtained as shown earlier for the MNBD (Equation 8) and the ANBD (Equation 9). We contrast these new models with three other possible methods for predicting EDs. The first is the empirical ED just described and the second is a naïve independent model, where we assume that there is no association in exposures across different websites. This allows us to estimate a multivariate model that is the product of the marginal distributions. Essentially, this is the MNBD with all the bivariate and trivariate ω parameters set to zero. Park and Fader (2004) also use a model assuming independence as a benchmark for their bivariate model which allowed for correlation in website visit timing. The third model is Li et al's. (2002) multivariate discretized Tobit model for page views. Although their model allows for the inclusion of demographic variables, we do not make use of this possibility as none of the other models incorporate demographics. This makes the models consistent in their use of data.

The empirical model described in §5.1 appears to be the basis of Nielsen's proprietary WebRF model. Another proprietary model has been developed by Atlas DMT for forecasting the reach and frequency for Internet campaigns. However, it requires both user-centric panel data and data on the number of ads served to a visitor using cookies placed on a user's PC. As we do not have the cookie information, we can not reproduce this method, which is based on simulation (Chandler-Pepelnjak 2004). However, we can at least compare the accuracy of our models with that of the Atlas DMT simulation method, as Chandler-Pepelnjak (2004) reports that its average prediction error for reach is 20%. We do not compare our models with those examined by Leckenby and Hong (1998) since, as noted above, models like the BBD are not adaptable to the predictive setting that is germane to Internet media planning.

The test schedules in our model comparison range from $m = 1$ to 15. For the single-vehicle schedules ($m = 1$), all 45 websites in Table 2 are estimated with univariate NBDs as given by Equation (1). For all

[12] If it happened that the page impressions for November were less than for September, then we would truncate the September impressions to less than 30 days to get a match with November.

Table 4 **Average Errors for Alternative Models Across Different Demographic Groups**

		Demographic group					
		All households, $n = 10{,}000$		College educated, $n = 2{,}865$		Households with children, $n = 4{,}420$	
Number of websites	Model	RER, %	EPOR, %	RER, %	EPOR, %	RER, %	EPOR, %
1 (45 schedules)	NBD	10.6 (0.74)*	28.3 (1.27)	12.6 (0.97)	35.9 (1.90)	9.8 (0.86)	28.3 (1.49)
	Tobit	14.1 (0.88)	57.4 (1.39)	16.0 (0.99)	60.6 (1.57)	15.1 (0.96)	57.9 (1.35)
	Empirical	13.3 (0.79)	30.9 (1.62)	14.8 (1.0)	42.1 (2.20)	12.6 (0.87)	33.2 (1.75)
2–8 (1,400 schedules)	MNBD	5.9 (0.11)	19.7 (0.23)	5.0 (0.13)	20.5 (0.25)	5.0 (0.13)	19.9 (0.24)
	ANBD	5.9 (0.11)	21.0 (0.29)	5.0 (0.13)	21.0 (0.31)	5.0 (0.13)	20.6 (0.29)
	Independent	9.6 (0.16)	23.4 (0.26)	11.4 (0.19)	28.2 (0.29)	11.4 (0.27)	27.2 (0.28)
	Tobit	9.2 (0.17)	83.3 (0.43)	11.1 (0.19)	84.6 (0.43)	10.8 (0.19)	84.8 (0.46)
	Empirical	8.5 (0.15)	20.5 (0.35)	7.8 (0.17)	23.5 (0.35)	7.9 (0.15)	22.0 (0.35)
9–15 (1,400 schedules)	ANBD	3.2 (0.07)	11.3 (0.11)	4.2 (0.10)	13.4 (0.12)	3.6 (0.09)	12.6 (0.11)
	Empirical	6.3 (0.06)	13.4 (0.15)	5.5 (0.07)	15.0 (0.14)	5.8 (0.05)	14.4 (0.14)

*Standard error in parentheses.

other values of m, we randomly select 200 advertising schedules. We find that when the number of websites exceeds $m = 8$, models that require estimation of the full joint distribution such as the MNBD and multivariate Tobit become computationally very slow. For this reason, we fit only the ANBD and the empirical methods for schedules with $m = 9$ through 15.

5.3. Definition of Prediction Errors

Denote $f_x = f(X = x)$ and $\hat{f}_x$, respectively, as the observed and predicted probabilities of the ED for the validation data of November, $x = 0, 1, 2, \ldots$. The two measures of error employed are the same as used previously by Danaher (1988, 1989, 1991), Leckenby and Hong (1998), and Leckenby and Kishi (1984): relative error in reach (RER), where

$$\text{RER} = \frac{|\hat{f}_0 - f_0|}{1 - f_0},$$

and error in the exposure probabilities over schedule reach (EPOR), where

$$\text{EPOR} = \frac{\sum_{x=0}^{20} |\hat{f}_x - f_x|}{1 - f_0}.$$

We limited the EPOR range of exposures to 20 as beyond that almost all observed exposure frequencies are zero. An example of these fit statistics is given in Table 3, showing that for the Internet advertising campaign comprised of the three websites given in Table 1, the MNBD model gives the smallest values of RER and EPOR followed, respectively, by the ANBD, independence, and empirical methods.

5.4. Results of Model Comparison

The average prediction errors for the MNBD, ANBD, independent, Tobit, and empirical models are given in Table 4. Results are shown for the entire panel of 10,000 households plus the demographic subgroups of households where the highest education attainment is a college degree and households with children.

5.4.1. Single-Website Schedules. For single-website schedules, the MNBD and independent models reduce to the NBD so only the NBD is reported. The RER and EPOR error measures are both lower for the NBD than for the Tobit and empirical models. The standard errors for the RER and EPOR measures confirm that the NBD gives a statistically significantly better fit than the next-best model, at the 5% level, for almost all the comparisons.[13] As an illustration of the how the NBD predicts accurately, consider the website travelzoo.com in Table 1. Its reach across the

[13] The only nonsignificant differences are for EPOR in the "all households" group and RER in the "college-educated" group when comparing the NBD and empirical models.

30 days of September is 8.7% for 1,293 page impressions. For the validation period, the number of page impressions increases to 2,129, resulting in a much higher reach of 12.4%. Using the adjustment method described in §4.3, the δ value of 1.65 gives an adjusted α value of $1.053/1.65 = 0.6397$. Using Equation (1), the estimate of reach in the validation period is $1-(0.6397/(1+0.6397))^{0.1362} = 12.0\%$, very close to the actual reach value of 12.4% in the validation data. Contrast this with the empirical model, which gives a much lower reach estimate of only 8.8%. For this example, and in general across all 45 websites, the adjustment of the NBD to different page impression totals is superior to simply using raw data to compile a prediction of the downstream ED.

5.4.2. Reach Velocity and Acceleration. As described above, since Internet users rather than Web publishers largely determine the delivery of page impressions, online campaigns require an understanding of the *rate* at which users are exposed to the ads. For example, suppose that two websites both have a cumulative reach of 10% after 30 days but the first site achieves 9.5% reach after just 5 days, while the second site takes 25 days to attain 9.5%. It is apparent that page impressions served after the fifth day on the first website are delivering mostly repeat impressions to visitors previously exposed to the campaign, while the second site is still reaching new visitors after the fifth day.

Knowledge of the rate at which page impressions are "consumed" by site users is important when planning how to handle responses to an ad campaign. Consider, for instance, a campaign of 100,000 banner ad impressions for a discount travel package where customers must call an 800 number. If most of the impressions are delivered in 3 days, more call center staff are required than if the 100,000 impressions are spread evenly over 15 days.

To describe the rate at which reach builds over time for Internet ad campaigns, we introduce the concept of *reach velocity*. Using Equation (1), under the NBD model for a single website, the reach is defined as

$$\text{reach} = 1-\Pr(X=0\mid r,\alpha) = 1-\left(\frac{\alpha}{1+\alpha}\right)^{r}.$$

The NBD model is fit using D days of estimation data ($D = 30$ in §4.1). Recall from §4.3 that the NBD has the flexibility to accommodate different time periods with a straightforward modification of the α parameter (Lilien et al. 1992, p. 34). All that is required is to replace α with $\alpha/(t/D)$, where t is time measured in days. Denote the reach at time t as $R(t)$, which is

$$R(t) = 1-\left(\frac{D\alpha}{t+D\alpha}\right)^{r}.$$

Reach velocity is just the rate of change of reach, being

$$\text{reach\_velocity} = \frac{dR(t)}{dt} = \frac{r}{D\alpha}\left(\frac{D\alpha}{t+D\alpha}\right)^{r+1}.$$

Since $r > 0$ and $\alpha > 0$, reach velocity is always positive and, therefore, reach increases with time as expected. However, because we expect diminishing returns in reach, the rate of increase in velocity, i.e., reach acceleration, should be decreasing. This is confirmed by taking the second derivative of reach to obtain

$$\text{reach\_acceleration} = \frac{dR^2(t)}{dt^2} = -\frac{r(r+1)}{(D\alpha)^2}\left(\frac{D\alpha}{t+D\alpha}\right)^{r+2},$$

which is clearly always negative in sign.

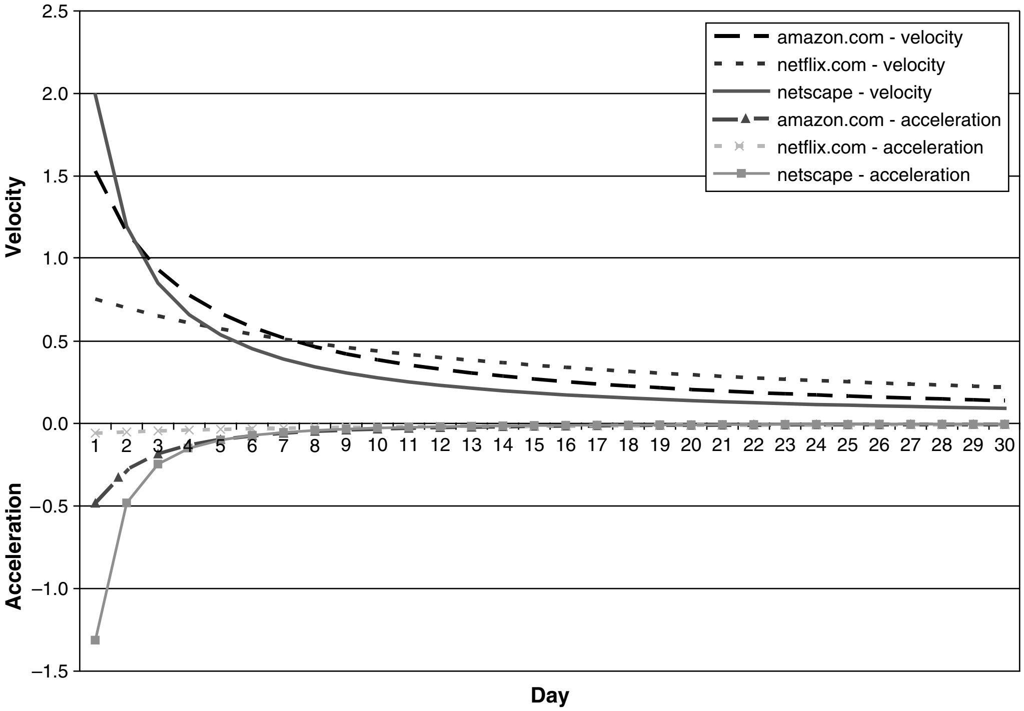
We illustrate the concepts of reach velocity and acceleration with an example from our data. It happens that the three websites amazon.com, netflix.com, and netscape.com have very similar reach values after 30 days, being 12.8%, 12.1%, and 12.1%, respectively. Despite this similarity, the growth rates in reach for these three sites are very different. We calculate the reach velocity and acceleration curves using the appropriate parameter estimates for each website and show these curves in Figure 1. It can be seen that at day one, netscape has the highest reach velocity and netflix has the lowest. netscape's higher velocity over netflix stops at day five, after which netflix's reach velocity always exceeds that of netscape. In fact, netflix's reach velocity is almost linear while the other two sites have more of a hyperbolic curve. This is also evident in the reach acceleration curves, where netflix is close to zero while amazon and netscape begin with large negative acceleration that rapidly converges to zero.

Figure 1 illustrates that both amazon and netscape quickly achieve close to their final reach while netflix has more of a linear growth toward its ultimate reach of 12.1%. The average frequency after 30 days for amazon, netflix, and netscape are 5.2, 2.0, and 14.2, respectively, showing that netscape in particular levels out in reach after about 5 days, beyond which subsequent page impressions are delivered primarily to those already reached. This results in a big buildup in frequency for netscape, with possibly excessive frequency delivered for the same reach that can be obtained for amazon and netflix.

5.4.3. Medium-Sized Schedules. For small- to medium-sized schedules ($m = 2$ to 8), Table 4 shows that the MNBD and the ANBD are the most accurate models for estimating reach, with RER values of only 5.9% on average.[14] This is significantly lower

[14] Recall that the ANBD always agrees with the MNBD reach estimate because it uses the MNBD reach estimate as input to its parameter estimation by the method of means and zeros.

Figure 1 Reach Velocity and Acceleration for Amazon, Netflix, and Netscape



at the 5% level than for the independent, Tobit, and empirical models and is also the case for the other two demographic groups. The same pattern emerges when examining the EPOR measure, which tests prediction accuracy over the entire ED. Again, all differences are significant except among the MNBD, ANBD, and empirical models in the "all households" demographic. Overall, the MNBD is the most accurate, followed by the ANBD, then the empirical, independent, and Tobit models. The finding that the independent model performs poorly underscores the observation by Park and Fader (2004) that independence among websites (be it for visit times or page impressions) is an unreasonable assumption. While the Tobit model gives reasonable predictions for reach, it predicts very badly for the remaining part of the ED as evidenced by its high EPOR values.

It is encouraging to see that even though the NBD approximation to the MNBD is not as accurate as the full multivariate model, it still performs as well or better than the empirical method. However, the ANBD can not completely replace the MNBD since it relies on the MNBD for its parameter estimation.

5.4.4. Large Schedules. For large schedules ($m = 9$ to 15), only the ANBD and empirical models are computationally feasible as they do not require the initial estimation of a full joint distribution. The ANBD again significantly outperforms the empirical model across all demographics. Notice also that the average RER and EPOR values for both models are somewhat lower than for small/medium schedules. This is consistent with Leckenby and Hong's (1998) observation that observed Internet EDs become smoother for a larger number of websites, much the same as for television which also has many channels (Rust and Klompmaker 1981). A smoother observed distribution is easier to model with the ANBD which is itself a smooth distribution, whereas the MNBD can capture the lumpiness in observed EDs which is more prevalent when $m = 2$ to 6.

We mentioned in §5.2 that there are two known proprietary models for forecasting Internet EDs. From the description on Nielsen/Netratings's website, the WebRF model appears to be similar to what we have called the empirical model in Table 4. This model is a very robust and challenging benchmark, based on individual-level data that can be filtered to accommodate alternative demographic targets as is commonplace in advertising. Table 4 shows that the RER errors for the empirical model are reasonably consistent across the demographic groups for $m = 1$ to 8, but the EPOR errors are higher for smaller sample sizes. For instance, the smallest demographic group, homes where someone is college educated, has the higher EPOR value across all schedule sizes. Hence, it appears that the empirical model is vulnerable to small sample sizes. In contrast, the MNBD and ANBD models are less vulnerable to smaller sample sizes, particularly for small- to medium-sized schedules.

The other proprietary model in everyday use is Atlas DMT's simulation method. We are not able to replicate this model but we can compare its reach performance with that of the models in Table 4. Chandler-Pepelnjak (2004) reports that the average RER value for Atlas DMT's model is 20%. Table 4 shows that none of the average RER values for any of the 4 models are higher than 20%, the closest being 16% for the Tobit model in college-educated households. In particular, the MNBD and ANBD models always have RER values much lower than 20%, averaging 5% over all schedule sizes.

6. Conclusion

The two new models developed for website page views and applied to predicting Internet EDs, the MNBD and its approximation ANBD, appear to have much merit. The MNBD, in particular, is derived from an elegant class of multivariate models introduced by Sarmanov (1966), extended by Lee (1996), and introduced to the marketing literature by Park and Fader (2004). An attractive feature of the Sarmanov model that makes it especially suitable for Internet audience prediction is that it allows each marginal distribution to have a negative binomial distribution while simultaneously accounting for association among websites. Even though there are millions of websites, when popular sites or sites with similar purpose are considered, there is frequently a high overlap that reduces reach but increases frequency. Knowledge of these dynamics is critical when advertising in any media. For instance, a model which assumes independence among websites performed very poorly at predicting Internet reach and the ED.

Previous efforts to model Internet EDs have adapted models used for traditional media, in particular, the BBD (Leckenby and Hong 1998). However, we show that such models do not capture a fundamental difference between exposure to online and offline ads. This makes the NBD, rather than the BBD, an appropriate model for each website because visitors to a site can receive anything from none to several thousand impressions over a time period. An additional feature of the NBD that makes it ideal for Internet EDs is its ability to adjust itself for different page impression totals. This expansion/contraction property is not shared by any of the models used for traditional media.

Of the two new models, the MNBD gives the best prediction accuracy but becomes computationally slow for a large number of websites. When advertising schedules have eight or fewer websites, the MNBD model is recommended. Beyond that number of websites, the ANBD gives slightly less accuracy but is much faster. Nevertheless, both new models perform better than existing proprietary models. It now remains to test these models on other Internet data sets to ensure their general applicability.

The validation method we employ has not been used in prior media modeling efforts and sets a more demanding standard for media model testing. In the past, models have used the same data for validation as for parameter estimation. This has been necessitated by a lack of data over a long time period and the use of only partial information from the estimation data set for model fitting, typically limited to univariate and bivariate information. The comScore data used here has no such limitation, however, and has revealed shortcomings in the use of empirical distributions when they are used for prediction. Such practice is commonplace in television, for example, where individual-level peoplemeter records are used to predict ratings some way into the future (three to six months). Empirical EDs are also used by proprietary Internet reach and frequency prediction software. We find that even though the empirically-based predictions are very good, they suffer from what might be termed "empirical overfitting." That is, they are limited to the visit behavior of estimation-period panelists, which ends up being somewhat different from that observed among the "population" in the validation period as little as one month later. Such overfitting is well-documented for econometric models used in forecasting applications. It is encouraging to see that our model does not suffer from overfitting to the same extent as the empirical model and, therefore, the MNBD must be capturing some fundamental website visit processes without being badly affected by the vagaries of particular panelists at particular time periods.

Such empirical overfitting could also prove problematic in CRM efforts, for example, where a sample of customer data is used to find patterns in purchase behavior that may not be as prevalent in the population of customers or potential customers. Further work is required to better understand what can and can not be reliably inferred from customer databases used for downstream prediction.

It is worth noting that the multivariate count model developed here can also be applied to situations other than Internet page views. For example, our generalization of the NBD could be used to model purchases across several categories of packaged goods. Finally, although not explicitly addressed here, the MNBD and ANBD models can be used as the basis for optimal media planning in much the same way that the BBD has been used for traditional media (see, for example, Little and Lodish 1969, Rust 1986). In this study, we also introduce additional factors to the usual media scheduling optimization problem, namely a variable number of impressions served to each website user and different rates of increase in reach even when the final reach achieved is the same. Incorporating these additional dimensions to media planning introduces further computational challenges and should be a fertile area for future research.

Acknowledgments

The author thanks comScore Media Metrix for supplying the data used in this study.

References

Anscombe, F. J. 1950. Sampling theory of the negative binomial and logarithmic distributions. *Biometrika* **37**(3/4) 358–382.

Bhat, S., M. Bevans, S. Sengupta. 2002. Measuring users' Web activity to evaluate and enhance advertising effectiveness. *J. Advertising* **31**(3, Fall) 97–106.

Cannon, H. M. 2001. Addressing new media with conventional media planning. *J. Interactive Advertising* **1**(2) http://jiad.org/vol1/no2/cannon/index.html.

Chandler-Pepelnjak, J. 2004. Forecasting reach, frequency and GRPs on the Internet. Atlas Digital Marketing Technologies, www.atlasdmt.com.

Chandon, J.-L. J. 1986. *A Comparative Study of Media Exposure Models.* Garland, New York.

Chintagunta, P. K., S. Haldar. 1998. Investigating purchase timing behavior in two related product categories. *J. Marketing Res.* **35**(February) 43–53.

Coffey, S. 2001. Internet audience measurement: A practitioner's view. *J. Interactive Advertising* **1**(2) http://jiad.org/vol1/no2/coffey/index.html.

Danaher, P. J. 1988. A log-linear model for predicting magazine audiences. *J. Marketing Res.* **25**(4, November) 356–362.

Danaher, P. J. 1989. An approximate log-linear model for predicting magazine audiences. *J. Marketing Res.* **26**(4, November) 473–479.

Danaher, P. J. 1991. A canonical expansion model for multivariate media exposure distributions: A generalization of the duplication of viewing law. *J. Marketing Res.* **28**(3, August) 361–367.

Danaher, P. J. 1992. Some statistical modeling problems in the advertising industry: A look at media exposure distributions. *Amer. Statistician* **46**(4) 254–260.

Ehrenberg, A. S. C. 1988. *Repeat Buying: Facts, Theory and Applications*, 2nd ed. Charles Griffin and Company Limited, London.

Ehrenberg, A. S. C., J. Wakshlag. 1987. Repeat-viewing with peoplemeters. *J. Advertising Res.* (February) 9–13.

Gal-Or, E., M. Gal-Or. 2005. Customized advertising via a common media distributor. *Marketing Sci.* **24**(2, Spring) 241–253.

Goodhardt, G. J., A. S. C. Ehrenberg, C. Chatfield. 1984. The Dirichlet: A comprehensive model of buying behavior. *J. Roy. Statist. Soc. A* **147**(5) 621–655.

Huang, C.-Y., C.-S. Lin. 2006. Modeling the audience's banner ad exposure for Internet advertising planning. *J. Advertising* **35**(2) 23–37.

IAB. 2006. IAB/PwC release full year 2005 Internet ad revenue figures. http://www.iab.net/news/pr_2006_04_20.asp. (April 20, 2006).

Iyer, G., D. Soberman, J. M. Villas-Boas. 2005. The targeting of advertising. *Marketing Sci.* **24**(3, Summer) 461–476.

Katz, H. 2003. *The Media Handbook: A Complete Guide to Advertising, Media Selection, Planning, Research and Buying*, 2nd ed. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Krugman, D. M., R. T. Rust. 1993. The impact of cable and VCR penetration on network viewing: Assessing the decade. *J. Advertising Res.* **33**(January) 67–73.

Leckenby, J. D., J. Hong. 1998. Using reach/frequency for Web media planning. *J. Advertising Res.* **38**(January) 7–20.

Leckenby, J. D., S. Kishi. 1982. How media directors view reach/frequency estimation. *J. Advertising Res.* **22**(June) 64–69.

Leckenby, J. D., S. Kishi. 1984. The Dirichlet-multinomial distribution as a magazine exposure model. *J. Marketing Res.* **21** 100–106.

Lee, M.-L. T. 1996. Properties and applications of the Sarmanov family of bivariate distributions. *Comm. Statist.: Theory Methods* **25**(6) 1207–1222.

Li, S., J. C. Liechty, A. L. Montgomery. 2002. Modeling category viewership of Web users with multivariate count models. Working Paper 2003-E25, Carnegie Mellon Graduate School of Industrial Administration, Pittsburgh, PA.

Lilien, G. L., P. Kotler, K. S. Moorthy. 1992. *Marketing Models*. Prentice Hall, Englewood Cliffs, NJ.

Little, J. D. C., L. M. Lodish. 1969. A media planning calculus. *Oper. Res.* **17**(1) 1–35.

Meskauskas, J. 2003. Reach and frequency—Back in the spotlight. iMedia Connection, www.imediaconnection.com. (November 5).

Metheringham, R. A. 1964. Measuring the net cumulative coverage of a print campaign. *J. Advertising Res.* **4**(December) 23–28.

Moe, W. W., P. S. Fader. 2004a. Capturing evolving visit behavior in clickstream data. *J. Interactive Marketing* **18**(1) 5–19.

Moe, W. W., P. S. Fader. 2004b. Dynamic conversion behavior at e-commerce sites. *Management Sci.* **50**(3) 326–335.

Morrison, D. G. 1979. Purchase intentions and purchasing behavior. *J. Marketing* **43**(Spring) 65–74.

Morrison, D. G., D. C. Schmittlein. 1988. Generalizing the NBD model for customer purchases: What are the implications and is it worth the effort? *J. Bus. Econom. Statist.* **6**(2, April) 145–159.

Park, Y.-H., P. S. Fader. 2004. Modeling browsing behavior at multiple websites. *Marketing Sci.* **23**(3, Summer) 280–303.

Rust, R. T. 1986. *Advertising Media Models: A Practical Guide*. Lexington Books, Lexington, MA.

Rust, R. T., J. E. Klompmaker. 1981. Improving the estimation procedure for the beta binomial TV exposure model. *J. Marketing Res.* **18**(November) 442–448.

Rust, R. T., R. P. Leone. 1984. The mixed-media Dirichlet-multinomial distribution: A model for evaluating television-magazine advertising schedules. *J. Marketing Res.* **21**(February) 89–99.

Sabavala, D. J., D. G. Morrison. 1977. A model of TV show loyalty. *J. Advertising Res.* **17**(6) 35–43.

Sarmanov, O. V. 1966. Generalized normal correlations and two-dimensional Frechet classes. *Doklady (Soviet Math.)* **168** 596–599.

Savage, S. J., D. M. Waldman. 2004. United States demand for Internet access. *Rev. Network Econom.* **3**(3, September) 228–246.

Smith, D. L. 2003. Online reach and frequency: An update. http://www.mediasmithinc.com/white/msn/msn042003.html. (April).

Wood, L. 1998. Internet ad buys—What reach and frequency do they deliver? *J. Advertising Res.* **38**(January) 21–28.