The Little Engines That Could: Modeling the Performance of World Wide Web Search Engines

Eric T. Bradlow and David C. Schmittlein*

Abstract

This research examines the ability of six popular web search engines, individually and collectively, to locate URLs containing common marketing/management phrases. We propose and validate a model for search engine performance that is able to represent key patterns of coverage and overlap among the engines.

The model enables us to estimate the typical additional benefit of using multiple search engines - depending on the particular set of engines being considered. It also provides an estimate of the number of URLs not found by any of the engines. For a typical marketing/management phrase we estimate that the "best" search engine locates about 50% of the URLs, and all six engines together find about 90% of the total.

The model is also used to examine how properties of a URL and characteristics of a phrase being sought affect the probability that a given search engine will find a given URL. For example, we find that the number of links within a web page increases the prospect that each of the six search engines will find it. Finally, we summarize the relationship between major structural characteristics of a search engine and its performance in locating relevant web pages.

KEY WORDS: World Wide Web, probability models, hierarchical Bayes, capture/recapture, marketing information.

*Eric T. Bradlow is Assistant Professor of Marketing and Statistics and David C. Schmittlein is Ira A. Lipman Professor, The Wharton School of the University of Pennsylvania, Philadelphia, PA.

1 Introduction

The World Wide Web (WWW) is important to managers in three rather different respects. First, managers use it to engage in electronic commercial transactions as sellers or as buyers (Alba et al. 1997; Hoffman, Kalsbeek, and Novak 1996). Second, they use it to disseminate information to customers or gather information as (business) customers, this function including both web advertising (acquiring new customers) and after-sales support to retain customers (Bakos 1997; Burke 1996; Hoffman and Novak 1996). Third, the web is emerging as a rich source of managerial information that assists in decision-making, e.g. competitive intelligence, demographic information, market trends and forecasts, general economic information, sources of external expertise or training, innovative managerial tools, tactics and strategies, and regulatory and other governmental information. Providers of such information include news organizations, governments, educational institutions, corporations, and nonprofit organizations (e.g. via press releases), etc.. Web search engines are commonly used to help locate this kind of information, and it is this performance of such engines that interests us here.

Search engine performance has begun to attract attention by both researchers and managers. Selberg and Etzioni (1996) studied the nature of search queries and results from those searches using various popular search engines, for the period July through September 1995. In a more recent and comprehensive study of engine performance published in *Science*, Lawrence and Giles (1998) examined the URLs returned for a large number of queries during December 1997. They were particularly interested in the relative number of URLs returned by the various different search engines and in estimating the number of URLs *not* found by any (or all) search engines. Coverage of those findings in *The Wall Street Journal* (1998) showed both the managerial interest in this performance and also the controversy generated by the findings. With significant advertising rev-

enue at stake, those responsible for the various search engines are sensitive to assessments of their relative performance. Indeed, such assessments have loomed large in the business press discussion of the vast sums paid to acquire search engine sites.

In this study we will offer the following contributions. First, we present and validate a model for the performance of multiple web search engines in finding URLs. We also analyze some natural, relatively simple models (Rasch-type ability/difficulty models, and the capture/recapture model used by Lawrence and Giles) and find that they fail to represent key aspects of search engine performance (which the proposed model does contain). Second, we analyze the performance of six popular web search engines in finding marketing/management phrases. Selberg and Etzioni (1996) studied all queries submitted to MetaCrawler, and Lawrence and Giles (1998) examined queries from the scientists at the NEC Research Institute in Princeton. Neither focused on management information. Third, we show how some characteristics of marketing/management phrases and of URLs affect search engine performance. We also highlight the association between structural characteristics of a search engine (e.g. size of universe covered, depth of search) and that engine's search success. Fourth, our empirical model application allows us to do more than just "rate the search engines", enabling us to describe the distinctive patterns of overlap and distinctiveness among them. Finally, for these kinds of management phrase searches, we are able to estimate the number of URLs *not* found by individual search engines, and indeed by the collection of all these engines. We also can calculate the incremental benefit in adding a particular search engine's results to those already found by one or more engines. The next section offers a description of the search process and search outcomes, some summary statistics regarding search engine performance, and a conceptualization of factors thought to affect that performance. The subsequent sections develop our model, validate it empirically, and use it to draw several kinds of substantive conclusions.

2 Searching the Web for Marketing Information

A simple example will help illustrate the research issues of interest. In October 1998, we queried each of six popular web search engines to find documents containing the phrase "mere exposure effect". Alta Vista found 99 documents. Northern Light located 83 documents, though of course many of these duplicated the ones from Alta Vista. HotBot found fewer URLs (49), but some had not been discovered by *either* Alta Vista or Northern Light. Finally, engines Infoseek, Excite, and Lycos found fewer documents (22, 21, and 9 respectively) but again some new pages were included. Together, all six engines located 172 documents; so even the "best" search engine (for this phrase) found less than 60% of this total (i.e. Alta Vista's 99 out of 172).

We should acknowledge at the outset that this study will not attempt to assess the relative "value" of the various individual sites found, and indeed one might well be skeptical of any mechanism that claimed to do so. Different searchers will no doubt have different interests or needs. Rather, thinking about this simple example leads directly to the five research questions that we do mean to address:

1. Search Engine Performance Across Phrases

Would the search-result pattern above hold up for other marketing phrases? "Mere exposure effect" is relatively new to marketing, and is more associated with academic research than with current marketing management practice. Perhaps some engines would do better for longer-established phrases, or those more prevalent among practitioners. Certainly, since web crawlers proceed from document to document via the links provided, search engines may end up covering relatively separate, disparate parts of the space of URLs. Considering the marketing phrase searched above, such a propensity can be exacerbated by, for instance, the inclination of academic sites to link to other academic sites (via connection to coauthors, references, etc.).

2. Factors Affecting Discovery of URLs

In the search example above, several URLs were found by almost all of the engines, while others were located by only one. For a given search phrase, what makes some URLs "easy" to locate? Obviously, in light of the web crawler process mentioned above, the more sites that link to a URL, the easier finding that URL will be. Of course, this measure is essentially impossible to observe. It is also not directly controllable by a site that *wishes* to be found. Instead we focus on two factors that are observable and (within limits) controllable: namely, the number of links *on* a URL (to other documents), and the domain type (.com, .edu, .org, etc.). The former should be related to URL discoverability because it indicates sophistication and connectedness of the document, and may also stimulate reciprocal linkage (i.e. a site linked from this document electing to provide a link back). The latter factor (domain type) may matter through a propensity for sites to link within (rather than across) these types.

3. Search Engine Structural Characteristics

Although search engines' operating details are proprietary, they are known to differ with respect to some basic characteristics. We will summarize the apparent relationship between such structural properties and the engines' search performance.

4. Overlap and Sequential Search

We are also interested in the way that patterns of overlap among the search engines determine their incremental benefit when combined. In our example above imagine that Alta Vista was the search engine used first. Would using a second engine be expected to add substantially to the number of documents found? What about a third? How many engines are needed to find the "lion's share" of relevant documents? Which particular engine would be expected to add the most to, say, Alta

Vista's results? The proposed model will allow us to answer each of these questions.

5. How Much Information Did We Miss?

Using all six search engines we found cumulatively, 172 documents mentioning "mere exposure effect". But how many documents did we fail to find? Note that any single URL's search results for our six engines can be summarized by a binary six-vector, where the i^{th} element is a "1" if search engine i found the URL in question, and is a "0" if it failed to find the URL. There are of course $2^6 = 64$ such patterns, and for each phrase searched we can create the full frequency count among these 64 patterns - except for one. The number of URLs associated with the (0,0,0,0,0,0) vector is of course not available in the data, since this represents the number of URLs missed by all six search engines. Of course, after creating a model that represents well the engines' web coverage and overlap (by fitting the 63 patterns above) we will use it to forecast the frequency of this 64^{th} pattern - as it indicates the size of the remaining "undiscovered" part of the web.

To build a model that would address these five issues, we proceeded through four steps to build an appropriate database.

Step 1: Marketing Phrases for Search

The marketing phrases searched needed to be diverse enough to represent an interesting universe, and also vary specifically on the factors thought to affect search engine performance (i.e. the substantive research question # 1 above). Accordingly, phrases were selected via three criteria:

1. they are relatively central to marketing thought, appearing in popular reference works (Bennett et. al. 1995; Clemente 1992),

2. they are specific enough so that a web search need not be refined further to be potentially useful (e.g. "marketing management" was found on 44,432 web pages by Alta Vista - too

many to be helpful without more detail), and

3. they span the two phrase dimensions discussed earlier: managerial versus academic; and newer versus older. Five phrases were selected in each cell of the resulting 2×2 design matrix, leading to 20 phrases overall.

Step 2: Phrase Search Via Search Engine

The six search engines examined in this study (Alta Vista, HotBot, Excite, Infoseek, Northern Light, Lycos) are the most popular based on user awareness, popular press mentions, and inclusion in previous studies and in metasearch programs (*PC Magazine Online* 1998; Beatty 1998). Note that while Yahoo! is often mentioned by users as a "search engine", the search engine provided on the site was at the time of our study powered by Inktomi, the same search engine used by HotBot. The 20 phrases were searched using each of the six engines during October 1998. During the search two properties of each located URL were recorded: the number of links contained on that document (0-5, 6-10, or 10+), and the domain type for the URL (.com, .edu, .org, or "other") indicating whether the site was commercial, academic, an organization, or other (the latter including sites outside the US - e.g. non-U.S. academic sites). This URL information will allow us to address the substantive research question #2 above.

Step 3: Integrate Search Results

As noted earlier the search result, for any located URL, can be summarized in a binary 6-vector. Meaningfully comparing search results across different search engines requires, however, substantial care. The same document may be reached by different alphanumeric strings, requiring that the documents themselves be accessed and checked, both for similarity across engines' results and for duplication within an engine's URLs. URLs were also checked to verify that they were active, and

that they in fact contained the phrase in question. (Both Excite and Infoseek use heuristics that may return URLs similar - but not identical - to the phrase searched. These instances were deleted from the dataset.)

Step 4: Search Engine Characteristics

As in research issue #3, we want the ability to link search engine performance to the characteristics of the six search engines. Since the number of search engines is small, it would not be useful to formally incorporate these engine characteristics into the model itself - but we will be able to investigate an association between overall search performance and structural properties of the engines. The key properties of interest are size (the total number of pages indexed by the engine) and several binary indicators of search engine capability. The latter includes Depth (whether an engine searches an entire site without limit or not), Frame Support (ability to follow frame links), Image Maps (ability to follow image maps), and Learns Frequency (whether an engine estimates the frequency with which a page's content changes, and uses that information to determine the frequency of site visits). Other search engine characteristics would be interesting to include (such as number of pages crawled per day) but do not appear to be reliably measured and available (Sullivan 1998). The search engine features above were taken from the Search Engine Watch site (Sullivan 1998) and were measured for each engine as of August 4, 1998.

Table 1 shows the 20 marketing phrases searched, their categorization regarding newness and academic/managerial, and the total number of URLs found by each search engine for each phrase. Note that this table is not the complete data, but rather a summary. For each of the 1588 located URLs, the data used in our model-development is a binary 6-vector together with the two URL characteristics (# of links, domain type) and two phrase characteristics (as above).

As a further summary for each search engine, Table 2 shows how the URLs found are distributed

across phrase and URL characteristics. The table entries provide the proportion of all URLs found (by any engine) having a particular characteristic that were actually located by the engine in question. For instance, Alta Vista located 52.1% of all managerial-phrase URLs that were found by any of the six engines. It did a little better (53.5%) finding academic-phrase URLs. Relative to the engine's baseline level of performance across all phrases, Infoseek had the greatest skew toward locating academic-phrase URLs (.163 academic versus .125 managerial) and Northern Light had the greatest inclination toward managerial-phrase URLs (.462 academic versus .529 managerial). Overall, Alta Vista had the best performance in finding academic-phrase URLs while Northern Light has the greatest success finding marketing-managerial phrase URLs. Analogous conclusions for other phrase/URL characteristics are available via Table 2. Table 3 provides the structural characteristics of the six search engines.

Before developing our model of search engine performance, it is useful to note what would happen if search outcomes for any given phrase were independent across search engine - i.e. if each URL had some probability of being located (possibly engine-specific) and one engine's finding the URL told us nothing about the probability that any other engine would find it. In such a situation, the substantive research questions #1 and #2 (effect of URL and phrase characteristics) could be addressed by a separate simple model (e.g. logistic regression) for each search engine, and research question #4 (overlap between engines) would have a very simple answer for any set of engines. The independence assumption is also the linchpin of the most careful model published so far for search engine performance (Lawrence and Giles 1998). Accordingly, we begin by considering this assumption in detail.

3 Are Search Engine Outcomes Independent?

The simplest, and arguably most natural starting point for representing the URLs found by multiple web search engines is the model used by Lawrence and Giles (1998) to estimate the size of the web (for science search queries). It is based on two assumptions. First, for any given search phrase j, it imagines that any given search engine i finds any one of the URLs containing that phrase independently of its finding other such URLs, and with some probability p_{ij}. Second, the model assumes that the probability p_{ij} that search engine i finds any particular URL containing phrase j does not depend on the set of URLs found by any *other* search engine. This pair of independence assumptions means that the frequency of URL finds for any given phrase, for the set of search engines, follows the multinomial distribution.

For a single URL containing phrase j, the data regarding web search engine outcomes can be written simply as the binary six-vector $(y_{1jk}, y_{2jk}, y_{3jk}, y_{4jk}, y_{5jk}, y_{6jk})$ where $y_{ijk} = 1$ if the kth URL for phrase j is found by search engine i, and is 0 otherwise. For URL k and phrase j the likelihood function is

$$\mathrm{L}(y_{1jk}, y_{2jk}, y_{3jk}, y_{4jk}, y_{5jk}, y_{6jk}) \quad = \quad \prod_{i=1}^{6} p_{ij}^{y_{ijk}} (1 - p_{ij})^{1 - y_{ijk}} \tag{1}$$

where p_{ij} is the probability that engine i finds any given URL containing phrase j. Since the URLs are exchangeable by assumption in this model, the likelihood for the data for phrase j is the product of (1) across all URLs. (Of course in practice a partial likelihood will be used, since the (0,0,0,0,0,0) vector will be missing.)

This independent multinomial model has much to recommend it. It is parsimonious: each search engine i (for each phrase j at least) can be summarized by a single quantity - it's search success

probability p_{ij}. The model can provide an estimate of the number of URLs not found. After any number of search engines have been used, the expected number of *new* URLs from another search engine h is simply $(N_j - m)p_{hj}$ where m is the cumulative number of URLs already found and N_j is the (unknown) number of URLs containing phrase j.

Lawrence and Giles expressed concern about the independence assumption and that concern was well founded. We report in Table 4 the value for -2logL for this model, and the associated BIC statistic. Four particular versions of the independent multinomial model were evaluated: (1) constant p for all engines and phrases, (2) different p for each engine but constant across phrases, (3) different p for each phrase but constant across engines, and (4) different p for each engine and phrase. A simple chi-square test on the value of -2logL rejects each of these four models. Naturally, with over 1500 observations the power of such a test is very high, and may not in itself present a strong case for substantial interdependence in search results across engines. Instead, two other considerations will argue for a model that relaxes the independence assumption. First, we will see later that these goodness-of-fit indicators can be improved substantially via a spatial interdependence model. Second, we note that the BIC criterion (which penalizes highly parameterized models for data overfitting) actually prefers, among independence models, the one where location probabilities differ only by search engine (and not by phrase). In other words, search is characterized simply by six p_i-values, one for each search engine.

The relative magnitude of these p_i-values across engines is simply the total URL count by engine at the bottom of Table 1. It is easy to show that an estimate of the number of URLs found by all engines in any 3-engine set (denoted 1,2, t for convenience) under this model is:

$$n_{12t} = \frac{n_{12}^2}{n_1 n_2} n_t. \tag{2}$$

Taking, for instance, Alta Vista and HotBot as engines "1" and "2", the actual 3-way overlap n_{12t} with each of the remaining four engines, and the overlap predicted by the independence model via (2), are :

Set of Search Engines	Actual # URLs	Predicted # URLs
Alta Vista, HotBot, Excite	50	22.4
Alta Vista, HotBot, Infoseek	37	22.7
Alta Vista, HotBot, Northern Light	100	77.5
Alta Vista, HotBot, Lycos	19	8.3

In short, looking across our 20 marketing phrases, the independence model substantially underpredicts the actual overlap in search outcomes for these triplets of search engines. These positive residuals suggest that two search engines with high coverage (Alta Vista and HotBot) are inclined to subsume the other four engines. This suggests the use of Rasch-type ability/difficulty models (Rasch 1966; Andersen 1973), whereby the probability that a given URL is located is a function of both a "difficulty" parameter associated with that URL and an "ability" parameter associated with the search engine. In this kind of model the "easy" URLs will tend to be found by all search engines and the "hard" URLs only by the search engines that find many URLs overall. In other words Alta Vista and HotBot will overlap somewhat, but the other search engines will overlap even more so with this pair (and hence produce positive residuals above), since the URLs they find will tend to be the "easy" ones already located by the high coverage engines. Of course, other search engine triplets could show different discrepancies than those observed above. Our main point is the observation that independence does not appear to be a solidly supported assumption, and a model where spatial location of search engines determine patterns of overlap may have value.

4 A General Proximity Model

In this section, we provide initially a heuristic description of our modeling approach for WWW data. This non-formal description is useful to describe our intuition, why we expect this class of models to improve on simpler ones described earlier, and the expected limitations and subsequent improvement in fit as our models become more complex. Needed notation and formal models are presented after.

4.1 Heuristic and Graphical Descriptions

We posit a general class of models for the ability of WWW search engines based on the proximity ("distance") from a specific engine to a given URL, and the "reach" of an engine. Our basic model suggests that when an engine and URL are proximate, the engine is likely to find that URL, and unlikely when not. In particular, each engine and URL are hypothesized to "sit" at an unknown location in D-dimensional space. A URL's location is modeled to be centered around a mean location determined by both its phrase and covariates specific to the phrase and URL (e.g. type of phrase, URL domain extension, etc... as described in Section 2). Then, from an engine's location, it "throws out a net" and probabilistically captures URLs within its reach. That is, there is a monotonically decreasing relationship between distance from engine to URL and the probability a URL is found by a given engine. Pushing this analogy farther, inferences of interest under the model are then derived from: (a) the location of each engine in the space (that is, do "weaker" engines find just a subset of those URLs found by the better engines, which would follow if all engines were located at the same place, or do engines "carve" out their own locations in the space), (b) the size of the net for each engine (in our model this is the ability of the engine), (c) the shape of the net (are the underlying dimensions related), (d) the number of underlying dimensions D adequate to

model the data, (e) the effects, if any, of phrase and URL covariates on URL's locations and hence their probability of being found, and (f) an exponent determining how fast the probability of an engine finding a URL drops off as a function of their proximity. In this research, we considered three specific cases of the general proximity model.

As a point of reference for describing the proximity models, consider the graphical representation of the independent multinomial model described in Section 3 and shown in Figure 1 panel A. The horizontal line represents the ($D = 1$ dimensional) space of URL locations, and the various search engines differ in the degree to which they (probabilistically) cover this space, beginning at the origin. The graph can be interpreted as having each engine stand at the origin and throw out a line, capturing as many URLs ("fish") as possible. Since engines with longer fishing lines (i.e. more ability) reach out farther from the origin, they are likely to "catch" *more* URLs, although, *which* URLs the better engine (engine 1) finds is unrelated to the *specific* URLs found by the weaker engine (engine 2). That is, via the independence assumption of the multinomial model, it is as if the URLs randomly redistributed their locations in the time elapsed between the search by engine 1 and that by engine 2.

As an alternative to this independence assumption, we will examine a $D = 1$ dimensional proximity model, depicted in Figure 1 panel B and denoted "Model 1" below. Here, each engine is again located at the origin and casts its probabilistic coverage of the line according to its engine's "ability". But unlike the independence model above, here the URL locations remain fixed. Accordingly, some URLs really are more difficult to locate (i.e. those labeled "D" and "E" in panel B) than others (e.g. "A" and "B") as they lie far from the origin. As a result, it is unlikely that the search engines with lesser ability will find URLs not found by the better engines. As suggested earlier in Section 2 (and confirmed below in Section 5.1) this feature of Model 1 does not fit the data particularly well. (Even the weakest engine Lycos finds URLs not found by other engines).

This suggested the extension of Model 1 in two ways under our general proximity model structure. First, in Model 2 (Figure 1 panel (C)) we extend to $D = 2$ dimensions yet leave all of the engine locations at the origin. A more general version considered in Model 3 (Figure 1 panel (D)) also allows the engine locations to vary - i.e. as suggested earlier, a search engine may "stake out" a distinctive part of the URL space not well covered by other engines. As shown below, the results indicate that Model 3 is necessary to provide an adequate fit to the pattern of web search results for marketing information.

4.2 Model Notation, Development, and Computational Approach

We consider the case described in Section 2 where each of $i = 1, \ldots, I$ search engines are utilized on the WWW to locate URLs for each of $j = 1, \ldots, J$ phrases. Let K_j denote the total number of distinct URLs found for the j-th phrase (by any of the engines) and y_{ijk} a binary outcome where $y_{ijk} = 1$, $k = 1, \ldots, K_j$, if the k-th URL for the j-th phrase is found by engine i, and 0 otherwise (as in (1), page 10). The collection of all outcomes y_{ijk} is denoted Y. In addition, for each URL we obtain covariate vector $x_{jk} = (x_{jk1}, \ldots, x_{jkP})$ in order to identify known characteristics of phrases and/or URLs that may make them harder or easier to find. The collection of all covariates is denoted X.

We posit a proximity model for $p_{ijk} = \text{Prob}(y_{ijk} = 1)$ defined as a function of the following engine and URL specific parameters. Let $\theta_i^t = (\theta_{i1}, \ldots, \theta_{iD})$ and $\gamma_{jk}^t = (\gamma_{jk1}, \ldots, \gamma_{jkD})$ denote the location of the i-th engine and k-th URL for phrase j in D-dimensional space. Additionally define Σ_i, a $D \times D$ dimensional scaling matrix for engine i, and $d_{ijk} = d(\theta_i, \gamma_{jk}) = (\theta_i - \gamma_{jk})^t \Sigma_i^{-1} (\theta_i - \gamma_{jk})$, a squared Mahalanobis distance between engine i and the k-th URL for phrase j. Thus, the diagonal elements of Σ_i are the abilities ("reach") and the off-diagonal elements indicate the covariation of

abilities for engine i in the D dimensions.

We assert a model for p_{ijk} as a function of d_{ijk} given by

$$p_{ijk} = \frac{1}{1 + d_{ijk}^u} \tag{3}$$

where (as described above) u defines the rate at which the probability an engine finds a given URL drops off. In general, spatial/distance models have been utilized in other marketing contexts, especially brand choice (Elrod 1988; Kamakura and Srivastava 1984). We note that (3) is equivalent to $\text{logit}(p_{ijk}) = -u \cdot \log(\text{d}_{\text{ijk}})$, a logistic link where u is the slope of regressor $\log(d_{ijk})$. Assuming conditional independence of engines, phrases, and URLs within phrase this yields a product Bernoulli likelihood for the unknown parameters $\Omega_1 = (\theta_1, \ldots, \theta_I, \gamma_{11}, \ldots, \gamma_{JK_J}, \Sigma_1, \ldots, \Sigma_I, u)$ equal to

$$p(Y|\Omega_1) = \prod_i \prod_j \prod_k (\frac{1}{1 + d_{ijk}^u})^{y_{ijk}} (\frac{d_{ijk}^u}{1 + d_{ijk}^u})^{1 - y_{ijk}} \tag{4}$$

As commonalities are likely to exist among the engines, the phrases, and the URLs, we extend the model for Y given in (4) to include a set of prior distributions for Ω_1 allowing for the sharing of information across units. The choice of priors for the components of Ω_1 were made in the following manner. Since the six engines that we consider here represent the engines of interest, we treat the engine specific parameters as fixed effects and put non-informative priors on θ_i, Σ_i, $i = 1, \ldots, I$. A non-informative prior is also adopted for u reflecting our lack of knowledge regarding this parameter. In contrast, it is of interest to summarize the location of phrase j for which we may regard γ_{jk}, $k = 1, \ldots, K_j$ as a random sample of URLs drawn from a population distribution. By convention and for computational convenience, we put a hierarchical multivariate normal-Inverse Wishart prior structure on the URL locations:

$$
\begin{aligned}
\gamma_{jk} &\sim \text{MVN}_D(\alpha_j + \beta x_{jk}, \Lambda_j) \\
\alpha_j &\sim \text{MVN}_D(\bar{\alpha}, \Sigma_\alpha)) \\
\Lambda_j &\sim W_\nu^{-1}(S)
\end{aligned} \tag{5}
$$

where $\text{MVN}_D(x, y)$ denotes a D-dimensional multivariate normal distribution with mean vector x and covariance matrix y, $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jD})$ the mean location of phrase j, β a $D \times P$ dimensional coefficient matrix where β_{dp} is the slope for the $p-$th covariate in dimension d, $\bar{\alpha} = (\bar{\alpha}_1, \ldots, \bar{\alpha}_D)$ the population mean of the phrase locations, Λ_j and Σ_α are $D \times D$-dimensional covariance matrices for phrase j and the population of phrase means, and $W_\kappa^{-1}(Q)$ denotes an Inverse-Wishart distribution with κ degrees of freedom and scale matrix Q. The values of ν and S were chosen as uninformative allowing the data to fully specify the values of Λ_j. As well, a non-informative prior distribution was utilized for β. We denote the prior level model parameters by $\Omega_2 = (\alpha_1, \ldots, \alpha_J, \beta, \Lambda_1, \ldots, \Lambda_J, \bar{\alpha}, \Sigma_\alpha)$ and the distribution on the prior level parameters by $p(\Omega_1|\Omega_2)$.

Inferences for the model parameters Ω_1 and Ω_2 were derived by obtaining samples from the marginal posterior distributions $p(\Omega_1|Y, X)$ and $p(\Omega_2|Y, X)$ using a Markov chain Monte Carlo (MCMC) sampler (Gelfand *et al.* 1990, Rossi *et al.* 1996). For each of Model 1, Model 2, and Model 3, we report results obtained by running three independent chains of the sampler for 3000 draws from overdispersed starting positions, discarding the initial 500 draws of each chain after determining convergence (Gelman and Rubin (1992)), and estimating the quantities of interest using the remaining 7500 draws. Further details of the computational methods are provided in the Appendix.

5 Results

5.1 Model 1: One-dimensional ability/difficulty model

We first considered a simple special case of the general proximity model defined by (3), (4), and (5) which consisted of a $D = 1$ dimensional model with all engines located at the origin $\theta_1 = \ldots \theta_I = 0$. To identify the model, we set as a reference point $\Sigma_1 = 1$, the ability of Alta Vista indexed as $i = 1$, and set the rate factor $u = 0.5$. This model, in which each engine ("examinee") has a unidimensional ability Σ_i and each URL has a unidimensional location γ_{jk} ("test item difficulty") is similar in spirit to the Rasch (1960) model commonly used in the modeling of educational testing data.

Model 1 was applied to the set of 20 phrases and 1588 URLs described in Section 2. A summary table of results for engine abilities, presented as Σ_i is given in column 2 of Table 5. The ordering of engine abilities suggested (Alta Vista, Northern Light, Hot Bot, Excite = Infoseek, Lycos) is unambiguous in all comparisons (true for all 7500 obtained draws) except for the comparison (a) Alta Vista > Northern Light, $p = 0.78$ and (b) Excite $\geq$ Infoseek, $p = 0.48$. The results from Models 2 and 3, better fitting models, described later will further refine these relations.

Inferences under Model 1 regarding the effects of the phrase and URL covariates (i) domain extension: .edu, .com, .org, other, (ii) # of Links on the URL page: 0-5, 6-10, 10+, (iii) Type of Phrase: Managerial/Academic, (iv) Age of Phrase: Newer/Older, and (v) the interaction between (iii) and (iv), on the mean of URL locations γ_{jk} and hence p_{ijk} are given in column 2 of Tables 6 and 7. In Table 6, we report the posterior median, standard error, and probability of the effect being greater than 0 for each covariate. Table 7 gives the adjusted phrase mean for URLs with a given covariate level. To interpret these findings, we note that all engines for Model 1 are located

at the origin and hence any positive coefficient suggests that the covariate level makes URLs of that type harder to find and vice-versa. We observe strong evidence that URLs with fewer links are harder to find than those with the most number of links (10+), and modest evidence that URLs having domain extensions .edu, or .org are slightly easier to find. Other inferences related to the parameters were: (a) there was no significant difference in the phrase locations (posterior median of $\Sigma_\alpha = 0.001$) which is consistent with the stable hit rates for each engine by phrase reported in Table 1 and the log-likelihoods for various models reported later in Table 10, and (b) URL variances Σ_j were inversely related to the number of URLs found for each phrase ($r = -0.85$).

A more detailed and informative look at the performance of Model 1 is presented in columns 3-6 of Table 8. Here we consider the number of URLs showing each of the $2^6 = 64$ possible engine-hit patterns. The table provides the observed number n_{obs} for each pattern (excluding (0,0,0,0,0,0)), as well as the 2.5%, 50%, and 97.5% percentiles for the predicted frequency of that pattern. Some interesting residuals are evident. First we note that Model 1 tends to underpredict the number of unique URLs found by each engine as seen in the unique engine-hit patterns 32, 48, 56, 60, and 62 (pattern 63 is slightly over-predicted). Secondly and related to the underprediction in the number of uniques, Model 1 also tends to overpredict the number of URLs found by exactly two engines as seen in patterns 16, 24, 28, 30, 44, 46, 47, 52, 54, 59, and 61 (patterns 31, 55, and 58 are adequately fit, and pattern 40 is underpredicted). These results were not surprising as suggested in our heuristic description of Model 1 since each engine is located at the origin and casting its "fishing line" in the same direction.

One further inference that can be derived from the model is an estimate of the number of URLs that are missed by all six of the engines. This question has managerial relevance from two perspectives. One, a manager searching for URLs on a specific topic may wish to know the fraction of those related URLs he or she is likely to find by using these six engines. A second issue relates to

the owner of a URL wanting his or her webpage to be found. Under the model, we can compute the posterior distribution of the number of URLs not found, K, by noting that

$$\begin{aligned} \text{P(all engines miss a URL)} &= \prod_i (1 - p_{ijk}) \Rightarrow \\ \text{P(at least one finds it)} &= 1 - \prod_i (1 - p_{ijk}) \Rightarrow \\ n_{obs} &= (1 - \prod_i (1 - p_{ijk})) * K \Rightarrow \\ K &= \frac{n_{obs}}{1 - \prod_i (1 - p_{ijk})} \end{aligned} \tag{6}$$

These results are shown in pattern 64 and suggest that the 95% posterior interval for the number of missing URLs for the 20 phrases is (253.94, 330.30) with posterior median 283.43. This indicates that Model 1 predicts $283.43/(1588 + 283.43) \approx 15\%$ of the URLs are missed by using all 6 engines.

5.2 Model 2 and Model 3 Results

We considered two additional special cases of the general proximity model to improve on Model 1. Model 2 consisted of a $D = 2$ dimensional version of Model 1 where each engine was located at the origin ($\theta_{11} = \theta_{12} = \ldots \theta_{I1} = \theta_{I2} = 0$). As a scale identifiability constraint we set Σ_{11}, the ability of Alta Vista on dimension 1, equal to 1. By definition, the addition of a second dimension would improve the fit over Model 1; however, we suspected that locating each engine at the origin, as per a pure ability/difficulty model, would still provide an inadequate fit to the data. In Model 3, we generalize Model 2 to allow individual search engines to carve out a distinctive portion of (2-dimensional) URL space, i.e. the engine locations $(\theta_{i1}, \theta_{i2})$ were allowed to vary. In fitting Model 3, we set $\theta_{11} = \theta_{12} = 0$, $\theta_{21} = 0$, restricted $\theta_{31} > 0$, and put $\Sigma_{11} = 1$ as shift, y-axis rotation, x-axis rotation, and scale identifiability constraints respectively.

Model 2 and 3 results for engine abilities $\Sigma_{i11}, \Sigma_{i22}$ and the correlation between dimensions

ρ_{i12} is given in Table 5 columns 3-8. A graphical representation of the engine performances for Model 3 is given in Figure 2, panels A and B. The results suggest that there are indeed two unique dimensions in which each engine operates. Model 2 findings give the ordering in dimension 1 of Northern Light, Alta Vista, HotBot, Excite, Infoseek, and Lycos whereas dimension 2 results give the ordering Alta Vista, Northern Light, HotBot, Infoseek, Excite, and Lycos. This is consistent with the Model 1 findings of an ambiguous ordering of Alta Vista versus Northern Light and Excite versus Infoseek. However, we note that the total "area" of the space covered by Northern Light is superior to that of Alta Vista as its posterior median abilities (2.670, 1.020) suggest greater coverage than Alta Vista's of (1.000, 1.760). These findings are replicated in Model 3 in which Northern Light is far superior to Alta Vista on dimension 1 (3.720 versus 1.000) and almost equal on dimension 2 (1.870 versus 1.960). This is suggested by Northern Light's high number of unique finds (pattern 62), indicating its location far from the other engines, but still high hit rate 785/1588 (i.e. high ability to "compensate" for a distant location). The remaining ordering of engines for Model 3 are similar to those described for Model 2.

The engine locations for Model 3 are given in Table 9 (also seen in Figure 2) and suggest that the engines do carve out different locations in the 2 dimensional space. Northern Light, and HotBot are located the farthest distance from Alta Vista indicating their abilities to have unique finds. Infoseek and Lycos are located "half-way" between Northern Light and Alta Vista, and in a sense are "maximizing" their ability to find URLs that happen not to be found by either of the two best performing engines. Excite's location near Alta Vista suggests, as described more fully in Section 6.3, that the additional benefit of using Excite if Alta Vista has already been used is less than that for Infoseek, despite the fact that they are "equally able" engines.

The effects of the phrase and URL covariates on dimensions 1 and 2 for Models 2 and 3 are given in Tables 6 and 7. The posterior probabilities of the effects being greater than 0 (Table 6

columns 4,6,8,10) indicate that in fact domain extension, # of links, and type and age of phrase do have a significant impact on the mean phrase location for a given URL. To interpret their effects on the probability that a given URL is found, consider Table 7 which gives the coordinates of the mean phrase location for a URL with each of the given covariate attribute levels, and that of a URL with each covariate level at the baseline condition. Since under Model 2, all engines are located at the origin, and the mean phrase under the baseline condition is at (0.130,-0.080) any covariate level that brings the phrase mean closer to the origin will increase the probability a URL is found and vice-versa. The results indicate that less than 10+ links and managerial phrases move the mean farther from the origin and hence lower the probability these URLs are found. The domain extension .com, .org, and the interaction of new and managerial phrase condition move the mean phrase locations closer to the origin. The remaining covariate levels have results which depends on the ability of a given engine in each dimension. The covariate-effect results for Model 3 generally need to be examined separately for each search engine as the locations of the engines vary. This examination is straightforward, using the phrase/URL locations from Table 7 and the search engine locations from Table 9. The results for Model 3 do however indicate one consistent finding across search engines: the 0-5 and 6-10 Link conditions move the mean phrase locations further away from the locations of the engines decreasing the predicted probability they are found. For the remaining cases, the results depend on the covariate and the specific engine.

A more detailed analysis for Models 2 (columns 7-9) and 3 (columns 10-12) of the 2^6 engine-hit patterns with observed counts n_{obs} and 2.5%, 50%, 97.5% quantiles is provided in Table 8. We observe a significant improvement in Model 3 fit for the uniques (patterns 32, 48, 56, 60, 62, 63) relative to Models 1 and 2. We also note that for 14 of the 15 engine pairs (excluding pattern 59) the 95% interval for Model 3 contains the observed find count compared to 3 out of 15 for Model 1 and 12 out of 15 for Model 2. An estimate of the fraction of URLs not found is also obtained

in pattern 64. The estimates under Model 2 (174.756/(1588+174.756)≈10% and that for Model 3 (192.704/(1588+192.704)≈11% are consistent with each other and suggest that these six engines as a whole, for these 20 phrases, cover a significant proportion of the total URLs. A global comparison of model fit is presented next.

5.3 Model Comparison and Cross-Validation

A global goodness-of-fit comparison was performed for each of Models 1 - 3 against the following reasonable and simple "strawman" independence models for the data: (1) constant p for all engines and phrases, (2) different p for each engine but constant across phrases, (3) different p for each phrase but constant across engines, and (4) different p for each engine and phrase. These are of course the special cases of the independent multinomial model discussed earlier. Table 10 presents the number of parameters, -2 * loglikelihood (-2logL) and the BIC criterion for each model. The results for Models 1 - 3 are the mean value of -2logL and BIC averaged across all 7500 MCMC draws.

The relative performance among these seven models is assessed by comparing the difference in -2logL across a pair of models with the χ^2 distribution having degrees of freedom equal to the difference in number of model parameters, or by using the BIC criterion. (This chi-square test is strictly appropriate among the four independence models, and among Models 1-3, since these 2 groups each represent a set of nested models. It is also defensible in comparing Model 3 with the four independence models, since the latter group are a limiting case of Model 3. In comparing Models 1 and 2 with the independence models, the chi-square test is a heuristic.) In the end, Model 3 is selected using both criterion. Interestingly, we note that Model 1 does not defeat the simple model of constant p for each engine and phrase or a different p by engine in terms of -2logL or BIC

and only defeats two models in terms of BIC.

To assess the predictive ability of our model, we employed a version of Bayesian cross-validation (Rust and Schmittlein 1985) where we dropped out in turn each of the 1588 URLs, re-estimated the model for each reduced data set, and predicted the engine find pattern for the left our URL. To make this approach computationally feasible under a MCMC simulation structure, we employed the method of Bradlow and Zaslavsky (1997) in which case deletion of URLs is implemented by importance reweighting the parameter draws from the full data posterior distribution. Due to the conditional independence structure of the likelihood given in (4), the importance reweighting scheme is trivial and computationally cheap in that each parameter draw is reweighted for URL jk by the inverse of the contribution of the likelihood of that observation to the total likelihood, i.e. $p(y_{jk}|\Omega_1, \Omega_2)^{-1}$. The total number of predictions made under this approach 9528 (1588 URLs by 6 engines) provides an adequate basis for validation. The results of the validation experiment indicated that Models 1-3 respectively were able to predict 58%, 72% and 81% of the URL correctly (all results significant at the 0.05 level) suggesting an adequate predictive ability of the modeling approach, and a substantial preference for the search-engine spatial locations in Model 3.

6 Discussion and Conclusions

We set out to better understand the performance of popular web search engines in finding marketing phrases. This required development of a model (Model 3) able to capture distinctive patterns of overlap and coverage among the engines. Furthermore, we wanted to understand how some characteristics of the phrase being searched, and of the URL being sought, would affect search outcomes for particular engines. As discussed in Section 5.2, two phrase characteristics (newer/older and managerial/academic) and two URL characteristics (number of links, domain type) significantly

affected search engine outcomes. The effect of number-of-links happens to be consistent across engines: the more links, the more likely the document will be located. Given the disparity in web engine coverage patterns (as in Figure 2), the other substantive effects differed by engine. For instance, a search for an academic phrase (as opposed to managerial) aided Infoseek's prospect for locating URLs, but hindered that of Northern Light.

To elaborate on our empirical and model-based results we conclude by addressing four simple questions:

- What search engine "works best"?
- Why do certain search engines find more URLs than other engines?
- What are the benefits to sequential search?, and
- How much information is still unaccounted for?

6.1 What Search Engine "Works Best"?

We again acknowledge that "best" here means simply locating more URLs containing the desired reference marketing phrase. Overall, based on the Model 3 estimates in Table 8 (and consistent with the results in Table 1) we can make five simple statements concerning the "best engine question":

1. Overall, for a randomly chosen marketing phrase and URL, the search engine most likely to find it is Alta Vista. BUT,
2. Northern Light is a very close second, and in fact does slightly better than Alta Vista in finding the managerial phrases, and

3. HotBot is a very respectable third, locating a little over 50-60% as many URLs as Alta Vista or Northern Light,

4. Excite and Infoseek trail more substantially, locating 20-30% as many documents as the two leading engines, and

5. Lycos found 10-15% as many documents as the two leaders.

Of course, these findings pertain specifically to the time period of search (October 1998), the information domain of interest to us (marketing phrases), and the particular 20 phrases selected. With respect to this last restriction, however, we note that the variation in mean locations across phrases (after accounting for our covariates) was very small. (The variance across phrases in the baseline mean phrase location $(\bar{\alpha}_1, \bar{\alpha}_2)$ from Table 7, is only .0027 for $\bar{\alpha}_1$ and 0.002 for $\bar{\alpha}_2$.) That is, another set of 20 phrases drawn at random from our marketing-phrase universe would have essentially no chance to change our findings. We next consider possible explanations for the differential performance of the engines.

6.2 Why Do Certain Engines Find More URLs?

Research question 3 in Section 2 asked how structural characteristics of search engines would affect the observed search results. Recall that some fundamental measures of this sort were provided in Table 3. Since the number of popular search engines (here 6) is small relative to the information in this table, it was not desirable to embed these features formally in our URL-location model. Armed, however, with overall performance statistics engine-by-engine we can conduct an exploratory analysis linking search engine properties to overall search effectiveness.

Of course, the factor that looms largest in such an analysis is search engine size - i.e., the total

number of web pages indexed by that engine. Not only would it be extraordinary if "size did not matter", but it could be well argued that "size is everything", i.e. that the number of URLs found by search engine A relative to engine B is entirely predicted by their relative sizes. This last hypothesis was essentially tested with the independence model of search outcomes, and rejected, in Section 3. In other words, our Model 3 with search engines that are somewhat distinct in the space that they cover (as in Figure 2) argues that structural characteristics beyond size may have an impact on search outcomes, and motivated us to examine the full set of characteristics in Table 3.

Accordingly, our profiling search outcomes based on engine characteristics was done in two sequential steps. The first examined the relationship between size and overall URLs found. The second looked at any deviations from a "size/total-URLs" connection to see if those deviations are associated with other engine properties from Table 3. Essentially, the factor size represents a very simple "par" model for engine performance, and we examine in step 2 engines that overperform (and underperform) relative to size.

Table 11 reports the results of these analyses. Columns (a) and (b) show clearly that our marketing phrase search outcomes are correlated substantially with search engine size ($\rho = 0.833$). They also show that size is far from the only factor that matters. Column (c) reports the ratio of URLs found to engine size. The variation in these values shows that much more is going on in search engine success than simply indexing more pages. Based on column (c) three of the search engines did substantially better in locating URLs than their size would indicate: Northern Light, Alta Vista, and Infoseek. At the other extreme, note that Lycos not only was tied for smallest size; but it also found fewer URLs relative to its size, in this study, than any of the other engines examined. Taking the overperformance of Northern Light and Alta Vista alone, one might suggest a convex relationship between size and URLs found (increasing returns to size) as opposed to a linear

one posited in column (c); but this explanation is inconsistent with HotBot's underperformance and with Infoseek's overperformance.

Instead, we sought to understand the variation in column (c) via the other search engine characteristics from Table 3. Specifically, we created a simple index of search sophistication from the characteristics Depth of Search, Frames Support, Image Maps, and Learns Frequency. For each engine, we summed the binary indicators for each of the four variables ("1" = more sophisticated search, "0" = less sophisticated) and report the resulting index in Table 11 column (d).

Our measure of sophistication does a good job of explaining which engines overperform relative to their size. The three overperforming engines in column (c) are also leaders with respect to the sophistication index - though Infoseek and HotBot were admittedly tied. Overall, the correlation between overperformance in column (c) and the sophistication index in (d) is $\rho = 0.658$, which shows that these structural properties of search engines are substantially related to engine performance, and in a way not reflected in the engine's size.

6.3 Sequential Search

One practical question of managerial interest is "which search engine should I use?" We believe that the previous two subsections summarize what our data and modeling say about that issue. Another practical question is "Now that I have used search engine **X** should I do an additional search, and if so what engine **Y** should I use?" Let's examine the first part of this question. Based on the results for Model 3 (Table 8), Alta Vista would be one's best single search engine choice and it is expected to find 48% of the marketing phrase URLs that exist. This is pretty good, but there is still plenty to find. More to the point, there is still plenty that can readily be found. Now turning to the second part of our question above, if one added a second search engine after using

Alta Vista, which should it be? Figure 2 by itself does not provide a clear answer. Instead, this figure shows that a putative case could be made for four of the other engines. HotBot's coverage does not overlap much with Alta Vista's, but Northern Light also does not overlap completely and covers a great deal of the URL space. Alternatively, Alta Vista will not actually find all URLs in its Figure 2 coverage area as indicated by the probability values 0.5 and 0.3 for the iso-probability curves, and many URLs exist to be found close to the origin. Excite and Infoseek are centered near the origin and accordingly are well-positioned to locate those residual URLs.

As it turns out, Northern Light is easily the best choice here for finding additional URLs. This can be established both by Table 8 using the actual search pattern finds (column 3) or Model 3's predicted search pattern outcomes. For our purposes it will suffice to simply tally the *incremental* URLs (not found by Alta Vista) for each of the remaining five engines. These are, in order, Northern Light (actual incremental = 443, predicted incremental using Model 3 = 468), HotBot (actual = 271, predicted = 259), Infoseek (actual = 136, predicted = 124), Excite (actual = 110, predicted = 109), and Lycos (actual =35, predicted = 42). Thus we conclude that in general it is important to consider both overall coverage ability and overlap in selecting combinations of search engines.

6.4 How Much Information Is Still Unaccounted For?

We have seen that combined search outcomes from multiple engines improves greatly on any one engine's performance. Yet, how much marketing information remains unlocated, even after using all of our six search engines? For our 20 marketing phrases, the results in Table 8 provide an answer to that question. Based on the estimate from Model 3, the fraction of total relevant URLs missed by all six search engines is just 10.8% (192.704/1786.967). So the search engines collectively find 89.2% of all URLs. Given the small variation in phrase location for our 20 marketing phrases

searched, the reader should feel confident that the search engines cover about 90% of what exists to be found for these kind of phrases.

This is quite different - and much better - than the web coverage estimated by Lawrence and Giles (1998) for their scientific-phrase searches. There, the six search engines were estimated to cover about 60% of the indexable URLs. Two explanations for the discrepancy across studies suggest themselves readily. First, the estimated number of URLs *not* found could be highly sensitive to the particular model specification selected. As we have seen, the data reject the independent multinomial model used by Lawrence and Giles to estimate this quantity, as that model does not effectively capture the patterns of overlap for sets of engines. So if we had to select one model to estimate the size of the web we would propose our Model 3 as a more appealing approach. Nonetheless, if the estimated web size is so sensitive to model specification one might well question the ability of *any* of these models to provide a reliable estimate, at least without exhaustive checking of individual assumptions of the model. Fortunately, this situation has not arisen. While we do not recommend using the independent multinomial model, its estimate of cumulative URL coverage by our six search engines (across all twenty phrases) is 89.6% - very close to the value found using our Model 3. In short, while the independent multinomial model methodology is suspect, it too indicates high coverage of marketing information by the set of search engines. Accordingly, the difference between our results and those of Lawrence and Giles does not stem from hypersensitivity to model assumptions.

This brings us to the second explanation for the difference; namely, that these kinds of marketing/management documents are relatively easy to locate. While we cannot prove this, it is a reasonable hypothesis. Parts of the web are of course much more "active" than others, with respect to both availability of hyperlinks from one document to another, and the degree of use of those links. This interconnectedness is the key to a search engine's performance. Documents containing

our marketing research and marketing management phrases may well be relatively active in this respect. That is, other web documents may be particularly likely to link to the commercial sites (e.g. company descriptions of a marketing technique), educational sites (e.g. university research available on the web) or organizations' sites (e.g. sponsored research) that contain the information. While our results do not say that web-based marketing information providers can simply count on search engines bringing multitudes to their location, they do indicate that much of the marketing information currently on the web can be located readily - if one uses multiple search engines.

We hope that this paper has provided some useful data, and some insight, concerning use of web search engines to find managerial information. Our proposed (and validated) spatial coverage model provides both a "snapshot summary" of the search engines vis-a-vis each other (as in Figure 2), and also yields predictions regarding cumulative performance of engine combinations. Of course the search engines themselves will evolve, and patterns of coverage and overlap can change accordingly. Nonetheless, our model framework should continue to provide a basis for summarizing these patterns. Furthermore, we have shown that certain characteristics of search engines, search phrases, and URL locations, affect the probability that a given engine will locate a given URL containing a given phrase. The marketing information base on the web is evolving - and expanding - very rapidly. For many purposes it has (and will continue to) outstrip the ability of managed directories, lists, and the like to provide focused useful direction - or even to keep up with change. The web search engines are well positioned to meet this challenge in the future, and currently they collectively - if not individually - can do so for marketing information.

Appendix

Inferences for parameters Ω_1 and Ω_2 are obtained from the marginal posterior distributions

$$p(\Omega_1|Y) \propto \int p(Y|\Omega_1)p(\Omega_1|\Omega_2)d\Omega_2, \quad \text{and} \tag{7}$$

$$p(\Omega_2|Y) \propto \int p(Y|\Omega_1)p(\Omega_1|\Omega_2)p(\Omega_2)d\Omega_1 \tag{8}$$

defined by the likelihood and priors given in (4) and (5). The non-conjugate likelihood and prior structure prevent closed-form integration of (7) and (8). The approach taken here to solve these intractable integrals is iterative simulation via a Markov chain Monte Carlo (MCMC) sampler. This approach states that under certain regularity conditions, samples from (7) and (8) may be obtained by repeatedly sampling values $\Omega_1^{(t+1)}$ from the conditional distribution $p(\Omega_1|Y, \Omega_2^{(t)})$ and $\Omega_2^{(t+1)}$ from $p(\Omega_2|Y, \Omega_1^{(t+1)})$ until convergence, and treating draws thereafter as draws from the desired marginal posterior distributions.

Unfortunately, for our model, the conditional distributions $p(\Omega_1|Y, \Omega_2^{(t)})$ necessary to straightforwardly implement an MCMC sampler can't be sampled from directly. We note that the conditional distribution of $p(\Omega_2|Y, \Omega_1^{(t+1)})$ can be sampled directly due to the conjugate multivariate normal - Inverse Wishart prior structure chosen for Ω_1. To sample $\Omega_1^{(t+1)}$ from $p(\Omega_1|Y, \Omega_2^{(t)})$ we implemented a Metropolis-Hastings jumping algorithm (Hastings, 1970) where for each parameter that was unconstrained, we utilized a symmetric Gaussian jumping distribution with mean at the previously drawn value $\Omega_1^{(t)}$, and variance set to provide a high acceptance rate for the draws. For those parameters constrained to the positive real line (variances, u, and θ_{31} in Model 3) we utilized a Gamma distribution kernel with shape parameter $k(\Omega_1^{(t)})^2$ and scale parameter $k\Omega_1^{(t)}$ which has mean equal to the previous draw $\Omega_1^{(t)}$ and variance $1/k$. The value of k was set differently for each

parameter to obtain an adequate acceptance rate.

Three independent streams for each of the three models was running using overdispersed starting values obtained from an initial run of the sampler. Computing time for Models 1-3 was 3, 12, and 14 seconds per iteration on an HP7000 workstation using Fortran 77 code.

References

Alba, Joseph, John Lynch, Barton Weitz, Chris Janiszewski, Richard Lutz, Alan Sawyer, and Stacy Wood (1997), "Interactive Home Shopping: Incentives for Consumers, Retailers, and Manufactures to Participate in Electronic Marketplaces", *Journal of Marketing*, 61 (July), 38-53.

Andersen, Erling B. (1973), "Conditional Inference for Multiple-Choice Questionnaires", *British Journal of Mathematical and Statistical Psychology*, 26, 31-44.

Bakos, Yannis (1997), "Reducing Buyer Search Costs: Implications for Electronic Marketplaces", *Management Science*, 43(12), 1676-1692.

Beatty, Sally (1998), "NBC Puts Its Firepower Behind Snap!", *The Wall Street Journal*, September 15, B8.

Bennett, Peter D. (1995) (ed.), *Dictionary of Marketing Terms* (2nd edition). Lincolnwood, IL: NTC Business Books.

Bradlow, E.T. and Zaslavsky, A. M. (1997), "Case Influence Analysis in Bayesian Inference", *Journal of Computational and Graphical Statistics*, September 1997, Vol. 6, 314-331.

Burke, Ray (1996), "Virtual Shopping: Breakthrough in Marketing Research", *Harvard Business Review*, 74(2) (March april), 120-131.

Clemente, Mark N, (1992), *The Marketing Glossary*, New York, AMACOM (American Management Association)

Elrod, Terry (1988), "Choice Map: Inferring a Product-Market Map from Panel Data", *Marketing Science*, 7, 21-40.

Feller, William (1968), *An Introduction to Probability Theory and Its Applications*, 3^{rd} edition. New York: Wiley.

Gelfand, Alan E., Susan E. Hills, Amy Racine-Poon, and Adrian F.M. Smith (1990), "Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling", *Journal of the American Statistical Association*, Vol. 85, 972-985.

Gelman, Andrew and Donald B. Rubin (1992), "Inference from Iterative Simulation using Multiple Sequences", *Statistical Science*, Vol. 7, 457-511.

Hastings, W.K. (1970), "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika*, 57, 97-109.

Hoffman, Donna L, William D. Kalsbeek, and Thomas P. Novak (1996), "Internet and Web Use in the United States: Baselines for Commercial Development", *Communications of the ACM*, 39 December, 36-46.

Hoffman, Donna L. and Thomas P. Novak (1996), "Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations", *Journal of Marketing*, 60(July), 50-68.

Kamakura, Wagner A. and Rajendra K. Srivastava (1984), "Predicting Choice Shares Under Conditions of Brand Interdependence", *Journal of Marketing Research*, 21, 420-434.

Lawrence, Steve and C. Lee Giles (1998), "Searching the World Wide Web", *Science*, 280 (3), 98-100.

PC Magazine (1998), "Web Search Sites: Metasearch Sites", December 1.

Rasch, G. (1960), *Probablistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Nielson and Lydiche (for Danmarks Paedagogiske Institut).

– "An Item Analysis Which Takes Individual Differences Into Account" (1966), *British Journal of Mathematical and Statistical Psychology*, 19 (Part 1), 49-57.

Ritter, Christian, and Martin A. Tanner (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler, *Journal of the American Statistical Association*, Vol. 87, 861-868.

Rossi, Peter E., Robert E. McCulloch, and Greg M. Allenby (1996), "The Value of Purchase History Data in Target Marketing", *Marketing Science*, Vol. 15, 4, 321-340.

Rust, Roland T. and David C. Schmittlein (1985), "A Bayesian Cross-Validated Likelihood Method for Comparing Alternative Specifications of Quantitative Models", *Marketing Science*, 4(1), 20-40.

Selberg, Erik and Oren Etzioni (1996), "Multi-Engine Search and Comparison Using the MetaCrawler", in *Proceedings of the Fourth International World Wibe Web Conference*, Boston, MA, p.195.

Sullivan, Danny (1998), "Search Engine Watch: Search Engines Feature Chart", webdocument address http://searchenginewatch.interest.com/webmasters/features.html.

The Wall Street Journal (1998), "Web's Vastness Foils Even Best Search Engines", April 3, B1,B5.

Table 1: Number of URLs Found By Search Engine and Marketing Phrase. Manag. = 1 indicates a managerial phrase, 0 an Academic phrase. Newer = 1 a newer phrase, 0 an older phrase.

			Search Engine*						**Total:**
Phrase	Manag.	Newer	**AV**	**HB**	**EX**	**IS**	**NL**	**LY**	**6 Engines**
flanker brand	1	1	9	9	1	6	5	0	19
umbrella branding	1	1	38	21	7	4	51	0	76
second mover advantage	1	1	8	9	4	2	20	1	26
professional respondents	1	1	41	19	12	7	31	0	62
audience fragmentation	1	1	106	59	37	36	120	14	215
category development index	1	0	18	11	0	2	19	2	29
modified rebuy	1	0	40	23	5	3	33	1	78
perceived value pricing	1	0	19	14	4	6	21	5	38
simulated test market	1	0	25	15	8	15	35	7	66
unaided recall	1	0	92	45	29	14	67	13	150
low involvement learning	0	1	10	11	5	7	13	4	22
elimination by aspects	0	1	61	35	21	8	57	4	114
mere exposure effect	0	1	99	49	21	22	83	9	172
preference map	0	1	29	21	10	35	41	1	101
decision calculus	0	1	74	54	28	32	55	14	134
multiattribute attitude models	0	0	17	6	2	1	26	2	37
Reilly's law	0	0	27	13	6	3	20	0	40
wheel of retailing	0	0	68	28	12	10	44	2	113
beta binomial model	0	0	39	13	9	10	33	4	64
diffusion of innovation model	0	0	20	13	6	7	11	2	32
Total			840	468	227	230	785	85	1588

* AV=Alta Vista, HB=HotBot, EX=Excite, IS=Infoseek, NL = Northern Light, LY=Lycos

Table 2: Search engine results by phrase age, phrase type, URL # of links, and domain extension.

	Age		Type		Links			Domain			
Engine	New	Old	Manag.	Acad.	0-5	6-10	10+	.edu	com	.org	other
AV	0.504	0.564	0.521	0.535	0.523	0.545	0.548	0.495	0.557	0.644	0.530
HB	0.304	0.280	0.297	0.293	0.284	0.288	0.328	0.312	0.269	0.328	0.288
Ex	0.155	0.125	0.140	0.144	0.146	0.138	0.137	0.140	0.142	0.164	0.143
IS	0.169	0.109	0.125	0.163	0.135	0.155	0.167	0.153	0.110	0.205	0.147
NL	0.506	0.478	0.529	0.462	0.481	0.551	0.505	0.502	0.526	0.521	0.458
LY	0.050	0.058	0.056	0.050	0.044	0.080	0.066	0.056	0.088	0.041	0.026

Table 3: Structural Characteristics of Search Engines*

	Search Engine					
Characteristic	AV	HB	EX	IS	NL	LY
Size (million pages)	140	110	55	30	80	30
Depth of Search	No Limit	No Limit	No Limit	Sample	No Limit	Sample
Frames Support	Yes	No	No	No	Yes	No
Image Maps	Yes	No	No	Yes	Yes	No
Learns Frequency	Yes	Yes	No	Yes	No	No
* Source: Search Engine Watch (Sullivan 1998)						

Table 4: Global goodness-of-fit for independence models. Reported are -2 * Log-Likelihood, and the BIC criterion.

Model	# Parameters	-2*LL	BIC
Constant p	1	11236.72	11245.88
Different p by engine	6	9602.83	9657.80
Different p by phrase	20	11197.58	11380.82
Different p engine by phrase	120	9276.17	10375.60

Table 5: Posterior median engine abilities on Dimensions 1 and 2 (Σ_{11}, Σ_{22}), correlation (ρ_{12}), and distance factor u for Models 1, 2, and 3. Posterior standard errors are in parenthesis.

	Model 1	Model 2			Model 3		
Engine	Dim 1	Dim 1	Dim 2	ρ	Dim 1	Dim 2	ρ
AV	1.000 (–)	1.000 (–)	1.760 (.41)	0.005 (.01)	1.000 (–)	1.960 (.40)	0.020 (.01)
HB	0.128 (.03)	0.157 (.01)	0.199 (.00)	0.006 (.01)	0.074 (.01)	0.761 (.01)	0.020 (.00)
Ex	0.017 (.00)	0.040 (.01)	0.024 (.01)	-0.586 (.01)	0.020 (.01)	0.060 (.01)	-0.630 (.01)
IS	0.017 (.00)	0.010 (.01)	0.040 (.01)	0.593 (.01)	0.055 (.01)	0.052 (.01)	0.640 (.01)
NL	0.945 (.08)	2.670 (.48)	1.020 (.40)	-0.007 (.09)	3.720 (.52)	1.870 (.45)	-0.003 (.00)
LY	0.001 (.00)	.0004 (.00)	0.001 (.00)	0.000 (.00)	.0003 (.00)	.0008 (.00)	0.000 (.00)
u	0.500 (-)		0.369 (.02)			0.354 (.015)	

Table 6: Phrase and URL covariate slopes (β) for Models 1, 2, and 3. Reported are the posterior medians (standard deviations) and posterior probability of the effect being greater than 0. Int. is the interaction between managerial and newer.

	Model 1		Model 2				Model 3			
Cov.	Dim 1		Dim 1		Dim 2		Dim 1		Dim 2	
.edu	-0.052 (.06)	0.177	-0.104 (.05)	0.050	-0.145 (.07)	0.008	-0.208 (.06)	0.000	0.017 (.04)	0.653
.com	0.018 (.07)	0.610	-0.223 (.09)	0.003	0.008 (.08)	0.423	-0.213 (.09)	0.003	-0.182 (.04)	0.000
.org	-0.090 (.12)	0.117	-0.134 (.11)	0.005	0.020 (.10)	0.633	-0.222 (.10)	0.018	0.057 (.05)	0.998
0L-5L	0.099 (.05)	0.957	-0.280 (.06)	0.005	-0.030 (.08)	0.470	0.138 (.06)	0.990	-0.146 (.03)	0.000
6L-10L	0.136 (.09)	0.947	0.090 (.10)	0.733	-0.017 (.09)	0.360	0.042 (.09)	0.673	-0.132 (.05)	0.005
Man.	0.067 (.11)	0.733	0.198 (.11)	0.990	0.370 (.12)	1.000	-0.114 (.10)	0.148	0.148 (.10)	0.913
Newer	0.106 (.11)	0.807	0.137 (.08)	0.930	0.082 (.15)	0.635	-0.230 (.08)	0.000	0.280 (.04)	1.000
Int.	-0.112 (.15)	0.237	-0.457 (.15)	0.000	-0.388 (.22)	0.010	0.026 (.13)	0.560	-0.374 (.10)	0.000

Table 7: Effect of phrase and URL covariates x_{jk} on the mean phrase location. $(\bar{\alpha}_1, \bar{\alpha}_2)$ is the mean phrase location with all covariates at baseline levels. $(\mu_1, \mu_2) = (\bar{\alpha}_1 + \beta_1 x_{jk}, \bar{\alpha}_2 + \beta_2 x_{jk})$ are the new coordinates including the covariate effects. Model 1: $\bar{\alpha} = 1.020$, Model 2: $(\bar{\alpha}_1, \bar{\alpha}_2) = (0.130, -0.080)$, Model 3: $(\bar{\alpha}_1, \bar{\alpha}_2) = (0.225, 0.068)$

	Model 1	Model 2	Model 3
x_{jk}	μ_1	(μ_1, μ_2)	(μ_1, μ_2)
.edu	0.968	(0.026, -0.225)	(0.017, 0.085)
.com	1.038	(-0.093, -0.072)	(0.012, -0.114)
.org	1.011	(0.004, -0.060)	(0.003, 0.125)
0L-5L	1.119	(-0.105, -0.110)	(0.363, -0.078)
6L-10L	1.156	(0.220, -0.097)	(0.268, -0.064)
Man.	1.087	(0.328, 0.290)	(0.111, 0.216)
Newer	1.126	(0.267, 0.000)	(-0.005, 0.348)
New + Man.	1.081	(-0.008, -0.016)	(-0.113, 0.112)

Table 8: Table of Web Engine Patterns and 95% confidence intervals for Models 1-3

#	Pattern	n_{obs}	Model1			Model 2			Model 3		
			2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
1	111111	2	0.020	0.032	0.052	0.019	0.113	0.825	0.017	0.104	0.623
2	111110	3	0.573	0.901	1.262	0.453	2.009	9.105	0.385	1.796	11.195
3	111101	1	0.024	0.041	0.060	0.016	0.096	1.009	0.021	0.115	0.843
4	111100	2	0.809	1.120	1.415	0.372	2.006	11.473	0.467	2.067	9.218
5	111011	3	0.175	0.284	0.450	0.153	0.805	3.599	0.155	0.616	3.553
6	111010	21	5.641	7.762	10.218	3.647	13.798	38.908	3.794	10.159	46.521
7	111001	0	0.225	0.353	0.509	0.148	0.740	4.358	0.162	0.706	4.677
8	111000	18	7.946	9.418	11.372	3.569	13.459	55.313	4.115	10.819	43.796
9	110111	3	0.179	0.289	0.425	0.190	0.625	6.061	0.213	0.788	3.075
10	110110	16	5.911	7.831	9.669	5.174	11.833	44.923	4.465	13.090	38.602
11	110101	1	0.246	0.357	0.509	0.121	0.619	4.927	0.263	0.794	3.828
12	110100	9	8.188	9.518	11.848	3.874	11.461	45.196	4.783	13.902	42.832
13	110011	7	1.721	2.556	3.427	1.243	4.902	20.107	1.478	4.284	16.198
14	110010	45	56.218	68.057	76.816	39.428	89.492	227.409	33.747	80.826	149.755
15	110001	2	2.220	3.086	4.035	1.198	4.844	18.233	1.495	5.245	19.813
16	110000	64	77.743	82.908	94.274	36.453	83.855	203.724	38.321	89.130	174.093
17	101111	3	0.071	0.111	0.176	0.045	0.220	1.527	0.053	0.252	1.727
18	101110	5	2.090	2.957	3.983	0.807	4.139	19.556	0.944	4.334	25.004
19	101101	0	0.089	0.139	0.212	0.040	0.195	1.360	0.057	0.264	1.962
20	101100	4	2.947	3.648	4.577	0.751	3.936	20.847	1.097	4.850	20.581
21	101011	4	0.631	0.973	1.431	0.253	1.671	7.866	0.400	1.569	8.107
22	101010	25	20.300	25.408	32.607	6.517	27.677	91.151	9.312	26.298	123.853
23	101001	1	0.825	1.210	1.782	0.371	1.481	7.023	0.423	1.669	9.798
24	101000	25	28.132	31.949	36.674	5.151	27.985	109.350	11.617	28.970	106.308
25	100111	2	0.681	0.971	1.415	0.412	1.388	11.372	0.539	1.893	8.642
26	100110	20	21.908	26.016	31.929	8.630	24.931	92.002	13.482	32.703	98.609
27	100101	4	0.883	1.176	1.710	0.343	1.290	8.354	0.545	2.091	9.668
28	100100	19	28.950	32.011	37.390	6.627	24.896	115.699	12.236	35.924	129.168
29	100011	9	6.364	8.431	11.262	3.053	10.262	41.608	3.704	12.416	38.137
30	100010	174	206.952	225.715	244.697	89.316	194.995	394.634	102.703	209.021	379.458
31	100001	8	7.843	10.398	14.566	2.678	9.751	30.043	4.200	13.789	42.763
32	100000	340	259.575	279.628	303.729	47.639	186.828	347.876	197.822	290.697	378.992
33	011111	2	0.024	0.040	0.059	0.019	0.115	0.716	0.017	0.091	0.435
34	011110	1	0.747	1.097	1.440	0.688	2.015	7.967	0.386	1.451	8.018
35	011101	0	0.031	0.050	0.075	0.014	0.107	0.703	0.019	0.096	0.587
36	011100	3	1.025	1.331	1.661	0.479	1.970	11.391	0.414	1.675	7.726
37	011011	0	0.217	0.348	0.512	0.146	0.822	3.502	0.101	0.566	3.300
38	011010	9	7.311	9.215	11.637	4.358	15.006	51.108	3.175	9.224	33.472
39	011001	0	0.289	0.424	0.631	0.176	0.718	2.900	0.098	0.600	4.482
40	011000	24	9.909	11.462	13.131	3.415	13.234	60.630	2.877	10.249	36.014
41	010111	3	0.235	0.351	0.495	0.190	0.752	4.655	0.229	0.723	2.877
42	010110	11	7.567	9.521	11.487	4.499	12.636	45.393	3.630	11.113	40.041
43	010101	2	0.303	0.432	0.624	0.147	0.666	3.678	0.140	0.799	3.564
44	010100	8	9.834	11.614	14.324	3.495	11.751	58.152	2.893	13.000	40.329
45	010011	3	2.199	3.041	3.941	1.410	5.082	22.546	1.281	4.111	14.191
46	010010	54	73.643	81.158	91.576	31.988	99.461	192.606	26.370	76.380	144.021
47	010001	2	2.703	3.684	5.172	3.182	5.192	17.191	1.211	4.632	19.091
48	010000	149	88.824	100.612	113.904	28.300	94.545	244.392	68.209	124.176	164.069
49	001111	1	0.087	0.135	0.201	0.056	0.245	1.740	0.046	0.236	1.440
50	001110	4	2.732	3.597	4.634	1.248	4.310	21.954	0.908	4.136	22.126
51	001101	0	0.111	0.166	0.261	0.043	0.204	1.232	0.054	0.245	1.673
52	001100	2	3.636	4.416	5.447	0.787	3.936	19.224	1.057	4.550	21.405
53	001011	0	0.798	1.180	1.700	0.294	1.680	9.401	0.260	1.500	8.731
54	001010	16	26.524	31.352	37.340	8.416	31.783	127.606	6.588	26.053	85.965
55	001001	1	0.986	1.417	2.314	0.304	1.759	6.933	0.320	1.580	10.571
56	001000	47	31.994	37.984	46.043	5.575	31.091	116.979	6.935	47.154	90.273
57	000111	5	0.873	1.159	1.651	0.426	1.675	9.338	0.496	1.734	7.448
58	000110	36	27.684	31.448	36.390	7.405	28.992	86.199	9.828	29.417	98.250
59	000101	0	1.058	1.401	2.210	0.370	1.423	6.813	0.405	1.862	9.071
60	000100	58	33.064	39.282	45.388	8.765	25.982	87.026	8.498	52.983	114.146
61	000011	7	7.793	10.255	13.397	8.992	51.819	36.697	3.479	11.115	38.810
62	000010	291	260.756	273.615	289.516	80.482	213.347	416.143	111.628	289.659	340.592
63	000001	9	9.430	12.453	18.961	2.735	10.615	27.031	2.946	11.953	40.908
64	000000	NA	253.938	283.431	330.301	44.592	174.756	327.190	82.149	192.704	354.000

Table 9: 2.5%, 50%, and 97.5% posterior percentiles for Model 3 engine locations (θ_1, θ_2).

	Dimension 1			Dimension 2		
Engine	2.5%	50%	97.5%	2.5%	50%	97.5%
AV	0.000	0.000	0.000	0.000	0.000	0.000
HB	0.000	0.000	0.000	-2.100	-1.820	-1.340
EX	0.150	0.265	0.626	-0.321	0.010	0.187
IS	-1.080	-0.800	-0.350	-0.110	0.330	0.600
NL	-2.140	-1.640	-1.377	-0.316	-0.004	0.184
LY	-0.949	-0.799	-0.645	-0.091	0.162	0.418

Table 10: Global goodness-of-fit for various models. Reported are -2 * Log-Likelihood, and the BIC criterion.

Model	# Parameters	-2*LL	BIC
Constant p	1	11236.72	11245.88
Different p by engine	6	9602.83	9657.80
Different p by phrase	20	11197.58	11380.82
Different p engine by phrase	120	9276.17	10375.60
Model 1	63	10643.00	11220.21
Model 2	140	8152.45	9435.13
Model 3	149	8020.71	9385.85

Table 11: The Relation Between Search Engine Performance and Search Engine Structural Characteristics

	(a)	(b)	(c)	(d)
Engine	Total URLs Found	Size (millions)	URLs/Size	Sophistication Index*
AV	840	140	6.0	4
NL	785	80	9.8	3
HB	468	110	4.3	2
IS	230	30	7.7	2
EX	227	55	4.1	0
LY	85	30	2.8	0

* Sum of indicators for high performance in Depth of Search, Frames Support, Image Maps, and Learns Frequency from Table 3