Information Systems Research
Vol. 23, No. 3, Part 1 of 2, September 2012, pp. 698–720
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.1110.0385

From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data

Zhiqiang (Eric) Zheng
School of Management, University of Texas at Dallas, Dallas, Texas 75080, ericz@utdallas.edu

Peter Fader
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, faderp@wharton.upenn.edu

Balaji Padmanabhan
College of Business, University of South Florida, Tampa, Florida 33620, bp@usf.edu

Managers routinely seek to understand firm performance relative to the competitors. Recently, competitive intelligence (CI) has emerged as an important area within business intelligence (BI) where the emphasis is on understanding and measuring a firm's external competitive environment. A requirement of such systems is the availability of the rich data about a firm's competitors, which is typically hard to acquire. This paper proposes a method to incorporate competitive intelligence in BI systems by using less granular and aggregate data, which is usually easier to acquire. We motivate, develop, and validate an approach to infer key competitive measures about customer activities without requiring detailed cross-firm data. Instead, our method derives these competitive measures for online firms from simple "site-centric" data that are commonly available, augmented with aggregate data summaries that may be obtained from syndicated data providers. Based on data provided by comScore Networks, we show empirically that our method performs well in inferring several key diagnostic competitive measures—the *penetration*, *market share*, and the *share of wallet*—for various online retailers.

Key words: business intelligence; competitive intelligence; competitive measures; probability models; NBD/Dirichlet

History: Vallabh Sambamurthy, Senior Editor; Siva Viswanathan, Associate Editor. This paper was received August 8, 2009, and was with the author 11 months for 2 revisions. Published online in *Articles in Advance* November 3, 2011.

1. Introduction

In the early part of the 20th century, Arthur C. Nielsen revolutionized the field of marketing research by inventing and commercializing the concept of "market share." Before ACNielsen Inc. began its store census process it was virtually impossible for firms to obtain timely, complete, and accurate market intelligence about competing brands. Today it is well recognized that knowledge of the overall competitive landscape is important for any business, and in response to this, there are many firms that specialize in the task of collecting and disseminating such information in various industries. Likewise, there are dozens of different kinds of measures that research firms (and their clients) use to characterize the competitive landscape (Davis 2007, Farris et al. 2006).

Recognizing the significance of competitive measures, a trend in the business intelligence (BI) field is the increasing importance given to *competitive intelligence* (CI), i.e., the information that a firm knows about its external competitive environment (Kahaner 1998). Although current BI dashboards are versatile and can pull data from different sources, most of the information in these dashboards is typically about the internal environment of the firm. Boulding et al. (2005, p. 161) consider this myopic view to be one of the pitfalls of current CRM practice; they suggest that successful implementation of CRM requires firms to incorporate knowledge about competition and competitive reaction into their CRM processes. Hence, methods that can provide useful competitive intelligence will be vital for the design of next-generation BI dashboards. As one example, Google Trends recently started providing some competitive intelligence documenting the volume of search queries across different competitors.

Current BI dashboards often fall short of providing CI capabilities, largely due to the fact that detailed information on competitors is hard to obtain. For most firms, these competitive measures are obtained solely through third-party data providers such as ACNielsen and other industry-specific syndicated data sources. Although such data can be integrated into BI systems to provide CI, it comes at the cost of

being expensive to acquire, and is also often based on historical data (rather than real time). In contrast to this approach, research in marketing (Park and Fader 2004) and information systems (Padmanabhan et al. 2001, 2006) has shown that firms may benefit from data augmentation, where they might augment their own internal data (which has been referred to as *site-centric data*[1] in the ecommerce context) with limited amounts of external data to achieve similar goals. This paper follows the same approach in spirit, and develops a method to infer competitive measures using augmented site-centric data.

Our focus is mainly on competitive measures that capture customer visits and purchasing behavior across competitors in e-commerce. Our model builds on the rich repeat-buying literature in marketing that developed and studied the classic Dirichlet model (Ehrenberg 1959, Goodhart et al. 1984) and its various extensions. The Dirichlet model assumes that each customer makes two independent purchasing decisions: she first decides on the total number of purchases within a product category, and then makes a choice on which competing firm's product to purchase. These two behavioral processes are captured through two probabilistic mixture models: the negative binomial (for category incidence) and the Dirichlet-multinomial (for brand choice). Taken together, the overall modeling framework is known as the "NBD/Dirichlet" (hereafter referred to as Dirichlet) in this literature (Winkelmann 2008). This research dates back to Ehrenberg (1959) with key updates (Goodhardt et al. 1984, Schmittlein et al. 1985, Ehrenberg 1988, Fader and Schmittlein 1993, Uncles et al. 1995) over the years. Numerous studies have documented the success of the Dirichlet model (Sharp 2010). It has been said that the Dirichlet model may be the best-known example of an empirical generalization in marketing, with the possible exception of the Bass model (Uncles et al. 1995).

A defining feature of the Dirichlet model is that for each customer it is necessary to know the number of transactions conducted with each of the competing firms in the market. We refer to this as *the complete information* requirement, which in reality is hard to fulfill because a firm needs to know the purchases their customers make across all competing firms. This requires the firm either to find a way to convince its competitors to share customer data or to convince their customers in the market to disclose their private purchasing data to the firm. Neither is an easy task. More often than not, a firm only has data on its own customers (i.e., site-centric data), and therefore is unable to implement the NBD/Dirichlet model.

Is there a middle ground where firms can obtain some data about their competitors, but not at the individual customer level? There is some recent work (e.g., Fader et al. 2007, Yang et al. 2005) that examines the possibility of sharing *summaries* of data instead of detailed customer transactions. Sharing such aggregate data rather than individual transactions can be a practical approach for retailers, and in some cases may be the only way to obtain competitive intelligence. At the same time, however, for many firms it is not enough to rely on aggregate summaries alone—they would like to make the best use of their internal data in conjunction with these external aggregate summaries in order to obtain the most complete picture of the competitive environment. These desires, constraints, and concerns bring us to the main point of this paper: *can firms combine their own customer-level data with commonly available aggregate summary statistics to infer important competitive measures*?

In this paper we show that this can be done by an implementation of the Dirichlet model in the limited-information scenario. Towards this end, we develop a realistic model termed as the *L*imited *I*nformation *N*BD/*D*irichlet (*LIND*) that improvises[2] the standard Dirchlet model. Our model aims to capture some of the power of the Dirichlet model, but with far less information, specifically with individual firms having access only to their own data plus the aggregate numbers (e.g., market share) from the other firms in the industry.

A key strength of the Dirichlet family of models is that they capture individual-level customer behavior and then derive the distributions of various aggregate statistics of interest, such as penetration, frequency, market share (MS), and share of wallet (SoW).[3] The

[1] The term *site-centric data* was first introduced by Padmanabhan et al. (2001) in the e-commerce context to refer to the data captured by an individual firm (site) on its customers' transactions with the firm. The counterpart, *user-centric data*, refers to the user-level data that capture these customers' transactions across firms (sites).

[2] Note that this objective differs fundamentally from a stream of papers that aim to *improve* the Dirichlet model by relaxing the model assumptions and incorporating additional marketing-mix variables such as price and promotion (e.g., Danaher et al. 2003, Bhattacharya 1997.) We thank the AE for making this important observation.

[3] Penetration is defined as the percentage of customers who transacted with the focal store. Frequency is defined here as the average number of purchases among the buyers of a product. Share of wallet is the percentage of purchases made to a specific store among those customers who actually transacted with the store (Uncles et al. 1995.) As an example, suppose that the total number of purchases to all online apparel stores is 1 million, and that a focal store, e.g., L.L.Bean, observes 100,000 of them. The market share for this retailer is 10%. Next, suppose that L.L. Bean's customers accounted for 300,000 of the total purchases to any apparel site. Their SoW is therefore 33%. SoW is also referred to as share of category requirements (SCR) in different settings. Please refer to the electronic companion for a brief definition of these marketing terminologies. The electronic companion is available as part of the online version at http://dx.doi.org/10.1287/isre.1110.0385.

performance of the Dirichlet model is often evaluated in terms of how well it infers these aggregate statistics, i.e., competitive measures (Ehrenberg et al. 2004, Bhattacharya 1997). We focus particularly on estimating penetration, the market share, and share of wallet for the full set of competing firms in a given market. Our results from data on various online retailers show that the limited-information model performs almost as well as the full-information NBD/Dirichlet model, with far less data, in inferring these key competitive measures.

Our research addresses a fundamental, and largely new, business problem and has implications to both the marketing and information systems fields. Our motivations, as well as implications of our results, relate to effective design of CI systems, data sufficiency, information sharing, and privacy—all areas in which the information systems community has strong interests. Our primary contribution to the marketing field is essentially a new model that works in a limited-information scenario that is commonly faced by marketers. Specifically, the method presented here and the associated empirical results are particularly important to information systems theory and practice for the following reasons:

- Our method provides a solution for CI when detailed transactional data on competitors are not available. This approach can be readily integrated into the design of BI dashboards due to the relatively simple inputs required from a data perspective.
- The competitive measures derived here can form the basis for sense-and-respond BI capabilities. Specifically, on a strategic level knowledge of the competitive landscape may clearly inform initiatives such as mergers and acquisitions. At the tactical level, the Internet presents unique opportunities for firms to act in real time on competitive information. For instance, carefully designed personalized promotions and online advertising strategies can often be implemented immediately, on a customer-by-customer basis, when faced with eroding market share or other competitive threats.
- From an information privacy context, we present an approach to sharing data that does not require individual transactions to be revealed. This is particularly relevant given numerous cases of identity revelation from supposedly anonymized data, which has made many firms far more cautious when it comes to sharing transactional data outside the boundaries of the firm. This is also related to recent research in information systems and management science on privacy-preserving information sharing (Fader et al. 2007, Menon and Sarkar 2007, Yang et al. 2005).
- It raises a more general question of how much data are really needed when there are specific objectives for which these data are sought. This is related to recent research in management science on information acquisition and active learning (Zheng and Padmanabhan 2006). Our results should also be viewed in the context of recent results in the IS literature that examine the innovative use of aggregate data to improve prediction models. For instance, Umyarov and Tuzhilin (2011) show that combining aggregated movie review data (i.e., IMDB data) with individual customer data from Netflix can enhance the movie recommendation accuracy.

The rest of this paper is structured as follows. In §2 we review related work from the research areas mentioned briefly above. Section 3 provides the theoretical background and discusses the full-information NBD/Dirichlet model that can be used to infer competitive measures when complete user-centric data are available. The limited-information NBD/Dirichlet model is then presented in §4. Results from applying the limited and full-information models are then presented in §5, followed by a discussion of limitations and future work in §6. We conclude in §7.

2. Related Work and Context

Recent work in a variety of disciplines, including finance (Kallberg and Udell 2003), marketing (Park and Fader 2004), economics (Liu and Serfes 2006), and information systems (Padmanabhan et al. 2001, 2006), have demonstrated the value of using customer data across firms for a variety of important problems. Kallberg and Udell (2003) demonstrate that lenders' sharing credit information on borrowers adds value to all participating lenders. In the case of online retail, Park and Fader (2004) show that customer browsing behavior can be modeled more accurately by using cross-site data, suggesting the value of sharing information between online firms. Padmanabhan et al. (2006) quantify the benefits that may be expected from using the more complete user-centric data compared to traditional site-centric data that individual online retailers typically observe, for predicting purchases and repeat visits. However, unlike our work here, which uses aggregate information, these approaches require integrating individual customer data across firms. Often a fairly large amount of such data has to be collected by a firm before any benefits can be reaped, and thus, Padmanabhan et al. (2006) caution against acquiring too little data. Unlike the above approaches, this paper focuses on learning competitive measures at the firm level, the inference of which may not need complete individual customers' across firm data, as we demonstrate later.

As we pointed out earlier, the research on competitive measures dates back to the early part of the 20th century, when Arthur C. Nielsen revolutionized the field of marketing research with the measure of "market share." Since then, marketers have

developed a variety of competitive measures. Farris et al. (2006) enumerates more than 50 competitive measures, including market share, market penetration, purchase frequency, etc., and show how to use a "dashboard" of these metrics to gauge a firm's business from various perspectives, such as promotional strategy, advertising, distribution, customer perceptions, and competitors' power. Davis (2007) defined over a hundred such metrics that marketers need to know to understand the competitive landscape. Both lists encompass all the competitive metrics proposed in the Dirichlet literature (Goodhardt et al. 1984, Ehrenberg 1988). Besides those we introduced before (i.e., penetration, market share, frequency, and SoW), other common measures include the percentage of a firm's customers who transacted with that firm only once ("once only"), the percentage of a firm's customers who only transact with that firm in the category ("100% loyal"), and the percentage of a firm's customers who also transacted with a specified other firm ("duplication"). Our proposed model can be used to derive all these measures analytically.

Many of the CI measures are also being used in firms as KPIs (key performance indicators),[4] as opposed to the conventionally inward-looking BI measures in a typical dashboard. However, the seemingly appealing notion of incorporating CI is often hindered by the lack of necessary data. Confidentiality and privacy issues are often major obstacles to obtaining the relevant data. This has spawned the rise of *privacy-preserving data mining* (e.g., Menon and Sarkar 2007), a set of methods that enable managers to share data while concealing any specific individual-level patterns or other potentially sensitive characteristics. Still, most of these approaches assume that firms are willing to share individual-customer level data, an assumption that is increasingly unrealistic in the context of some recent high-profile disclosure cases. As one example, when a prominent online portal released some of its supposedly anonymized search query data, reporters were able to link a specific individual to many of the search queries, sparking a major outcry.[5]

As mentioned in the introduction, one key measure we attempt to infer is SoW. The SoW measure has been noted to be one of the best ways to gauge customer loyalty as well as the overall effectiveness of a customer relationship management strategy (Uncles et al. 1995, Fox and Thomas 2006, Du et al. 2007). Beyond the academic literature, practitioners also acknowledge its importance (see, e.g., a recent white paper[6] that focuses on the online apparel industry). When comparing MS and SoW, it becomes quite clear that the latter is much harder for individual firms to determine, because it requires them to have information specifically on their customers' transactions at competing stores. However, we will show how it can be inferred from readily available summary statistics combined with the firm's site-centric customer records.

There is recent work (Du et al. 2007, Fox and Thomas 2006) that directly addresses the estimation of share of wallet from augmented data or from information sharing. Du et al. (2007) consider the case of a focal firm that desires to estimate the share of wallet for each of its customers. In this setting, initially the firm observes only their customers' purchases. Their method involves obtaining detailed survey data for a sample of customers regarding their activities at competing firms, and using these acquired values to impute those unknown values for other customers. The imputed values are then used to compute each customer's SoW. However, as recent research (Zheng and Padmanabhan 2006) in management science has shown, such list augmentation approaches may require extensive complete data before the imputation models work well. Whereas the approach in Du et al. (2007, p. 102) estimates SoW well for a validation set of 10,000+ customers, these results are based on acquiring survey data for approximately 24,000 customers. In practice, acquiring detailed data for such a large percentage of customers can be prohibitive. One possibility is to augment the above approach with *active learning* methods for list augmentation (Zheng and Padmanabhan 2006), but this has not been comprehensively studied in this literature as yet. Another recent approach (Fox and Thomas 2006) considered using customers' shopper loyalty card data to predict customer-level SoW. This approach models spending at competing retailers uses multioutlet panel data, and uses this model on loyalty card data to predict expected spending at other retailers, thereby generating SoW for each customer. In their experiments they use detailed panel data on 210 households to make SoW predictions for 148 households. Hence, as in the previous work, the acquired data in these experiments is considerably large.

Compared to these approaches, our method uses substantially less data (specifically the penetration or market share for each firm in an industry) to make SoW inferences. Second, whereas the above literature only estimates SoW for a focal firm, our approach

[4] See http://en.wikipedia.org/wiki/Competitive_intelligence.

[5] Source: AOL search data scandal at http://en.wikipedia.org/wiki/AOL_search_data_scandal.

[6] This report (http://www.techexchange.com/thelibrary/ShareOf-Wallet.html) begins with the following sentence: "Maximizing share of wallet is among the most important issues facing the partners in any consumer oriented value chain."

permits a focal firm to infer the SoW for all firms in its industry. Finally, acquiring individual customer-level data across firms may raise specific privacy concerns and will therefore require effective privacy-preserving approaches such as the ideas described in Menon and Sarkar (2007). We avoid this issue by using only aggregate data about competitors, an effective way of preserving individual privacy and minimizing data-handling complexity. Aggregated data, when used appropriately, can be a feasible and valuable alternative to more granular data. For example, the model proposed in Umyarov and Tuzhilin (2011), which combines aggregated movie review data from IMDB with the individual user's movie rating, enhances Netflix's movie recommendation system significantly.

Finally, there is broader interest in making inferences using data sets that combine limited information about individuals with comprehensive aggregate data. Recent papers such as Chen and Yang (2007) and Musalem et al. (2008, 2009) offer sophisticated new approaches that enable analysts to infer individual-level choice patterns from aggregate data. Although these papers seem similar in spirit to the method we develop here, there are several key differences worth emphasizing. First, although those papers utilize aggregate data, they assume that period-by period (e.g., weekly) sales information is available for all competitors. In many industries, however, this is not the case in practice. Second, although these papers provide valuable insights about individual-level choice processes, they are not built upon a rich, well-validated tradition such as the NBD/Dirichlet. By no means does this imply that their models are inferior, but our ability here to link the model (and the resulting output measures) back to an extensive and well-validated literature is a significant positive. Finally, these alternative approaches all require computationally intensive simulation-based estimation methods for model implementation. Ours, in contrast, is quite simple to implement—even in a familiar spreadsheet environment such as Excel. This can be an advantage for practitioners grappling with the real-world problem that we focus on in this paper.

3. Theoretical Background and the NBD/Dirichlet Model

The theory that we build is based on the extensively researched area of repeat buying in the marketing literature. The basic problem is to model customers, each of whom has a variable number of transactions within a product category, and the associated decisions about which competing firm to choose each time. A key strength of these models is that they capture individual-level customer behavior and then derive the distributions of various aggregate statistics of interest. In this case, we are going to work "backwards" in order to make individual-level inferences from some aggregate statistics. First, we briefly review the traditional NBD/Dirichlet model.

3.1. The NBD Model for Category Purchasing

The negative binomial distribution (NBD) is one of the best-known "count models" (Winkelmann 2008). A complete analysis of this model is provided by Schmittlein et al. (1985). A notable aspect of the NBD model is that just simple histogram data are needed for parameter estimation. The NBD model assumes that the overall number of (category) purchases (denoted as N, with the individual subscript suppressed for ease of exposition) for any customer during a unit time period can be modeled as a Poisson process with purchase-rate parameter λ. However, each individual may have a different purchase rate. To accommodate this difference (heterogeneity) among customers, these rate parameters are assumed to be gamma distributed $f(\lambda) = (\alpha^r \lambda^{r-1} e^{-\lambda\alpha})/\Gamma(r)$ with shape parameter r and scale parameter α, where $E(\lambda) = r/\alpha$. It is common knowledge that these two assumptions combine to yield the NBD model for the aggregate number of customer purchases. In other words, the probability of observing n purchases in any fixed unit time period for the random count variable N is:

$$P(N = n \mid r, \alpha) = \int_0^\infty P(N = n \mid \lambda) f(\lambda \mid r, \alpha)\, d\lambda = \frac{\Gamma(r+n)}{\Gamma(r) n!} \left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^n. \quad (1)$$

The key feature defining the NBD model is the linearity of the conditional mean—the purchase propensity as specified in Equation (1) does not change over time, known as the stationary market assumption (Schmittlein et al. 1985). Through comprehensive empirical and simulation studies, Dunn et al. (1983, p. 256) show that the NBD assumptions are reasonable for the majority of buyers, both for brand- and store-level purchases. They conclude that "for most purposes in brand purchasing studies, the NBD tends to be accepted as robust to most observed departures from its (stationary) Poisson assumption." Morrison and Schmittlein (1988, p. 151) also hold that the NBD model is very robust and the combination of the Poisson and Gamma processes tend to work very well. Further details about the NBD as a standalone model of product purchasing can be found in Ehrenberg (1959), Morrison and Schmittlein (1988), and Schmittlein et al. (1985).

3.2. The Dirichlet-Multinomial Model for Brand Choice

Every time a customer makes a purchase in the category, she also makes a brand-choice decision. As in the case of the NBD model for category purchasing, one first makes an assumption about the individual-level choice process, and then brings in a mixing distribution to capture heterogeneity across customers. In this case, a standard multinomial distribution is used for the former, and a Dirichlet distribution for the latter. This leads to the well-known Dirichlet-multinomial (or "Dirichlet," for short) model, which was discussed in detail in Ehrenberg (1988), Goodhardt et al. (1984), and also in Fader and Schmittlein (1993).

Suppose there are k brands, let a_j be the Dirichlet parameter indicating the attraction of brand j, and let the sum of a_j be $s = \sum_{j=1}^{k} a_j$. That is, s represents the overall attractiveness of the entire category (to a random customer) and a_j/s can be interpreted as the share of attractiveness (i.e., market share) of brand j in the market. Then the probability of a customer making x_1 purchases of brand 1 and x_2 purchases of brand $2, \ldots,$ and x_k purchases of brand k, given n category purchases has been shown to be:

$$\begin{aligned} &P(X_1 = x_1, \ldots, X_k = x_k \mid n, a_1, \ldots a_k) \\ &\quad = \int_0^{1-\sum P_j} \int_0^1 \cdots \int_0^1 \prod_{j=1}^{k} P(X_j = x_j \mid n, p_j) \\ &\qquad \cdot g(p_j \mid a_1, \ldots a_k) dp_1 \ldots dp_k \\ &\quad = \binom{n}{x_{1\ldots}x_k} \frac{\Gamma(s)}{\prod_{j=1}^{k} \Gamma(a_j)} \frac{\prod_{j=1}^{k} \Gamma(a_j + x_j)}{\Gamma(s+n)}. \end{aligned} \quad (2)$$

Ehrenberg et al. (2004) point out that the Dirichlet model works because it nicely captures regularities prevalent in a variety of markets such as (1) smaller brands not only have fewer customers, but they tend to be purchased less frequently by their customers (referred to as "double jeopardy" as if smaller brands are punished twice); (2) large heterogeneity of customers, with some purchasing very few and some purchasing very frequently; and (3) much the same proportion of any particular brand's customers also bought another brand of interest, i.e., the so-called constant duplication phenomenon (p. 1310). These "lawlike" patterns have been confirmed from soup to gasoline, prescription drugs to aviation fuel, where there are large and small brands, and light and heavy buyers, in geographies as diverse as the United States, United Kingdom, Japan, Germany, and Australasia, and for over three decades (Sharp 2010).

The Dirichlet model is not without limitations. It is best applied in what Ehrenberg et al. (2004) refer to as a stable market where (1) customers do not change their pace of purchases of a product over time (a result of the stationary NBD process of category purchases) and (2) the brands are not functionally differentiated and they show no special partitioning of certain brands (the so-called nonsegmented market as a result of the multinomial choice process). Over the years, there have been attempts to extend the model to nonstable markets by considering the existence of nonstationarity (Fader and Lattin 1993), change of pace and niche markets where a brand specializes in attracting a particular group of customers (Kahn et al. 1988), a large amount of customers who make zero purchases of a brand (spike at zero), and violations of the distributional assumptions (Morrison and Schmittlein 1988) and in segmented markets (Danaher et al. 2003). In the third online appendix (of the electronic companion), we provide more-detailed explanations for all these marketing terminologies.

3.3. The NBD/Dirichlet for Full Information

In a typical application of the NBD/Dirichlet model, researchers assume that the two aforementioned behavioral processes are independent of each other (i.e., there is no linkage between category incidence and brand choice). Assuming that complete data are available to observe both processes, standard estimation approaches (e.g., maximum likelihood) can be used to obtain the two parameters for the NBD component and the k parameters for the Dirichlet component. In the full-information world, it is unnecessary for a researcher to estimate all $k + 2$ parameters simultaneously—thus, the "full-information" NBD/Dirichlet is not truly an integrated model; it is just the concatenation of two separate, independent submodels. In contrast, our "limited-information" approach, to be discussed shortly, is fully integrated and allows for the simultaneous estimation of the entire model.

Once the $k + 2$ parameter estimates are available, it is possible to derive a broad array of summary brand performance measures, including a variety of competitive indicators, such as market share, share of wallet, and the number of 100% loyal customers. In contrast to more traditional econometric approaches, model evaluation is often judged by how well a given model can capture the various diagnostic measures mentioned above (Goodhardt et al. 1984, Ehrenberg 1995, Fader and Schmittlein 1993, Uncles et al. 1995), in addition to overall model fitness. We will follow this approach in this paper, and will similarly evaluate our limited-information model based on how well the model can be used to derive these important competitive measures.

We conclude this section by reminding the reader about the overall data requirements for the typical full-information NBD/Dirichlet model. Specifically,

for each customer it is necessary to know the number of transactions conducted with each of the competing firms in the market. As noted at the outset of the paper, this is often a very difficult data structure to obtain. More frequently, firms have complete information only for their brand, and therefore are unable to implement the NBD/Dirichlet model. We begin to address this "limited-information" scenario in the next section.

3.4. The BB/NBD Model for Limited Information

Although full information on each customer's cross-brand purchasing may not be available in practice most of the time, many firms still wish to tease apart the two underlying behavioral processes (i.e., category incidence and brand choice) from each other. Although it may be hard to sort out separate NBD and Dirichlet models when the observed data confound the two processes, there is an elegant model that, in theory, allows this "teasing apart" to be accomplished *even without complete data for each customer*. This model, known as the beta-binomial/NBD (or BB/NBD) was first introduced by Schmittlein et al. (1985) and has been used in a variety of settings to try to pull apart two integrated processes that cannot be separately observed (often due to data limitations as described earlier). The BB/NBD is a special two-brand case of the NBD/Dirichlet model that lumps all competing brands into a single "other brand" composite. Here the setting is that of a random customer who first decides whether to visit the category according to a NBD process and then decides whether to purchase the focal brand according to a beta-binomial process (Schmittlein et al. 1985).

Like the NBD/Dirichlet, the BB/NBD continues to assume that category incidence is governed by an NBD submodel, but instead of using the complete k-brand Dirichlet-multinomial, the BB/NBD uses a dichotomous beta-binomial in its place. In other words, the decision of whether to purchase the focal brand is now a binomial process (as opposed to the multinomial process in the Dirichlet case) with choice propensity p. However, customers are allowed to differ from each other in their choice propensities, and this distribution of p across the population is assumed to follow a beta distribution where $g(p) = (1/B(a,b))p^{a-1}(1-p)^{b-1}$ with mixing parameters as a and b (the beta distribution is the two-alternative special case of the more-general Dirichlet distribution). This leads to the BB/NBD distribution, the detailed derivation of which is presented in the first online appendix (of the electronic companion).

$$P(X=x) = \int_0^1 \int_0^\infty p(X=x \mid \lambda, p) g(p) f(\lambda)\, d\lambda\, dp$$

$$= \frac{\Gamma(r+x)}{\Gamma(r)x!}\left(\frac{\alpha}{\alpha+1}\right)^r\left(\frac{1}{\alpha+1}\right)^x \cdot \frac{\Gamma(a+x)}{\Gamma(a)}\frac{\Gamma(a+b)}{\Gamma(a+b+x)}{}_2F_1\left(r; b+x; a+b+x; \frac{1}{\alpha+1}\right), \quad (3)$$

where X represents the observed number of purchases of the focal brand and ${}_2F_1()$ is the Gaussian hypergeometric function.[7] Note that this model needs four parameters to be estimated (r and α parameters for the NBD model of category purchase, and a and b for the beta distribution of brand choice).

The elegance of the BB/NBD model lies in its ability to make inferences about each of the model components without being able to observe them separately. A benefit of this model is that most of the summary measures obtained from the full NBD/Dirichlet can still be obtained despite the data limitations. For instance, the expected reach or penetration (i.e., the percentage of customers who made at least one purchase of the focal brand) can be derived as $\textit{Penetration} = 1 - P(X = 0) = 1 - (\alpha/(1+\alpha))^r {}_2F_1(r; b; a+b; 1/(1+\alpha))$. This and other key summary measures are derived in detail by Fader and Hardie (2000).

However, despite the appeal of the BB/NBD, its ability to really sort out the underlying processes of interest is questionable. Given that a typical BB/NBD data set is just a simple histogram (capturing the distribution of purchase frequencies for the focal brand alone), it is hard to uniquely identify each of the four parameters in a reliable manner. The empirical analyses performed by researchers such as Bickart and Schmittlein (1999) and Fader and Hardie (2000) show several problems, including: (1) a high degree of sensitivity to initial settings in the parameter estimation process; (2) a very flat likelihood surface indicating the presence of many local optima; (3) limited ability to outperform simpler specifications, such as the ordinary NBD by itself; and (4) managerial inferences that do not have a high degree of face validity. The extent of these problems would likely become even more acute when trying to make inferences on competitive summary statistics (e.g., SoW) because the data lacks any information at all about the various competing brands.

With these problems in mind, we need to augment the basic BB/NBD model by introducing some information about competitive brands. This will give the data set more "texture," and such data augmentation beyond the simple histogram can allow the reliable estimation of multiple parameters with the

[7] The Gaussian hypergeometric function ${}_2F_1()$ is discussed in detail in Johnson et al. (1992) and the Wolfram functions site at http://functions.wolfram.com/HypergeometricFunctions/Hypergeometric2F1/02/. It solves the integral ${}_2F_1(a; b; c; x) = (\Gamma(c)/(\Gamma(b)\Gamma(c-b))) \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tx)^{-a}\, dt$.

Table 1 Overview of Notation Used

Notation	Explanation
X_{ij}, x_{ij}	X_{ij} is a random variable and x_{ij} is the actual number of purchases of customer i to site j.
r, α	The parameters that capture customers' category-level purchase behavior according to an NBD process, where r is the shape parameter and α is the scale parameter.
a_j	The Dirichlet parameter a_j captures customers' multinomial choice propensity to site j.
s and b_j	Summary statistics (used for convenience) where $s = \sum_{j=1}^{k} a_j$ and $b_j = s - a_j$.
${}_2F_1()$	The Gaussian hypergeometric function detailed in Footnote 7.
M_j	The total number of customers for the focal site j.
N_i, n_i	N_i is a random variable and n_i is the total number of category purchases for customer i.

ability to make direct linkages to (and inferences about) the focal brand's competitors. The key idea is to move from the pure BB/NBD back towards a NBD/Dirichlet, but using only commonly available summary statistics instead of the complete individual-level transaction histories that it usually requires. We demonstrate how to do so in the next section.

4. The Limited-Information NBD/Dirichlet (*LIND*) Model

There are three characteristics for the data we utilize in our limited-information setting:

1. Each firm has access to its own customer data, i.e., site-centric data for an online store, but knows that there may exist customers that only purchase with its competitors in the same category (whom they may therefore never observe).

2. Each firm knows exactly one piece of aggregated information (e.g., penetration or market share) for each of the other relevant firms in the category. For example, if Amazon is the focal store to examine, then it uses as inputs the count information on all its customers (e.g., the number of purchases that each customer made at Amazon) as well as just the penetration (or market share) for each of the other leading online stores. This kind of competitive information is widely available at low cost.[8]

3. At the category level, we make a realistic assumption that we can only observe those customers who made at least one purchase at the category. That is, we assume we do not know how many customers may be interested in a category but have not purchased anything yet.

We believe that these are reasonable assumptions, supported by actual data collection practices. We make these assumptions strictly because of their realism, as opposed to mathematical convenience. Furthermore, they apply equally well to an online or offline context, and are not tied to any particular industry/sector.

4.1. The Model

We propose a model that integrates the NBD model for category purchasing and the Dirichlet-multinomial model for brand choice, yet with a considerably easier data requirement than the full-information NBD/Dirichlet relies upon. This limited-information NBD/Dirichlet (LIND) model takes one specific firm (e.g., Amazon) as the focal one and models the entire market from its perspective. We summarize the notation to be used in Table 1.

First we assume that the usual NBD process applies to category-level purchase patterns, with one difference: because we only observe customers with at least one purchase in the category (data assumption 3), we model the number of category purchases, N, as a shifted Poisson[9] with rate λ:

$$P(N = n \mid \lambda) = \frac{\lambda^{n-1} e^{-\lambda}}{(n-1)!}, \quad n = 1, 2 \ldots; \ \lambda > 0. \tag{4}$$

Then, as with the regular NBD, we use a gamma distribution with parameters r and α to characterize the differences in these rates. This yields a commonly used *Shifted NBD* (*sNBD*) model for category purchasing. Next, given the number of category purchases for a random customer, the choice of the focal brand is modeled (just as before) using a binomial distribution with rate p. Following standard assumptions (Schmittlein et al. 1985, Fader and Hardie 2000), λ and p are assumed to follow a gamma and a beta distribution, respectively.

Altogether this yields a new model that we term as *BB/sNBD* (beta-binomial/shifted negative binomial

[8] Such data are routinely reported by many research firms, both online and offline. For instance, Information Resources, Inc. has regularly reported penetration statistics for grocery categories for many years (e.g., Fader and Lodish 1990), and comScore Networks publishes online firms' reach as part of its Media Metrix service (http://www.comscore.com/metrix/).

[9] Another common way to deal with nonzero data is the truncated Poisson. We derive the truncated LIND model and present the results in Online Appendix 2 of the electronic companion as a possible alternative. Guidelines for choosing between the shifted and truncated models are also provided there.

distribution) model. The detailed derivation of this model is provided in Appendix 1. Below we directly present the marginal distribution of x.

$$P(X=x)=\left(\frac{\alpha}{\alpha+1}\right)^r\frac{b}{a+b}\cdot{}_2F_1\left(r;b+1;a+b+1;\frac{1}{\alpha+1}\right),\quad x=0 \tag{5.1}$$

$$\begin{aligned}P(X=x)&=\frac{\alpha^r}{(\alpha+1)^{(x+r-1)}}\frac{\Gamma(r+x-1)}{(x-1)!\Gamma(r)}\frac{\Gamma(a+x)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+x)}\\&\quad\cdot{}_2F_1\left(x+r-1;b;a+b+x;\frac{1}{\alpha+1}\right)\\&\quad+\frac{\alpha^r}{(\alpha+1)^{(x+r)}}\frac{\Gamma(r+x)}{x!\Gamma(r)}\frac{\Gamma(a+x)\Gamma(b+1)\Gamma(a+b)}{\Gamma(a+b+1+x)\Gamma(a)\Gamma(b)}\\&\quad\cdot{}_2F_1\left(x+r;b+1;a+b+x+1;\frac{1}{\alpha+1}\right),\quad x\geq 1\end{aligned} \tag{5.2}$$

where x represents the number of purchases of a customer to a site. Note that Equation (5.1) yields the expected penetration:

$$\begin{aligned}&\textit{Penetration}\\&=1-P(X=0)\\&=1-\left(\frac{\alpha}{\alpha+1}\right)^r\frac{b}{a+b}{}_2F_1\left(r;b+1;a+b+1;\frac{1}{\alpha+1}\right).\end{aligned} \tag{6}$$

If we only had data from a single site, the BB/sNBD would be a logical model to use, and we will examine it as a benchmark to help evaluate the performance of the LIND model. In order to go from the BB/sNBD towards LIND, we replace the beta-binomial component with the multibrand Dirichlet-multinomial choice process, so that each site j is characterized by its own attraction parameter, a_j. The focal site thus has $K+2$ parameters to estimate: two parameters (r and α) for category purchasing and K parameters $a_1, a_2, \ldots, a_k$ representing the cross-firm Dirichlet parameters. Now, as shown in Equation (6), if these parameters are known, then each firm's expected penetration can be derived analytically. Because in our setting we know the actual value of these penetrations (see data assumption 2), we introduce K constraints on the LIND model that restrict the expected penetration computed for each site to equal the actual observed values in Equation (7). We would like to note that LIND is not restricted to using only penetration as the input (aggregate measure); it is a flexible model that can accommodate other inputs (e.g., market share, frequency, etc.) in a similar fashion. We will also demonstrate how to use market share as the inputs instead later in §5.2, and provide guidelines in §6 on how to choose appropriate inputs to best estimate the competitive measures of interest.

Adding constraints turns the usual unconstrained maximum-likelihood optimization problem into a constrained optimization task, but a relatively straightforward one. This constrained optimization formulation (Equation (7)) solves the problem from the focal site's perspective. Denote x_{ij} as the number of purchases of customer i to site j. The objective function is simply the sum of the log-likelihood based on the density function shown in Equation (A5). Let M_j be the number of customers of site j. We have:

$$\begin{aligned}&\max_{r,\alpha,a_1\ldots a_k}\sum_{i=1}^{M_j}\ln[P(X_{ij}=x_{ij})]\\&s.t\begin{cases}\textit{Penetration}_1=1-\left(\frac{\alpha}{\alpha+1}\right)^r\frac{b_1}{a_1+b_1}\cdot{}_2F_1\left(r;b_1+1;a_1+b_1+1;\frac{1}{\alpha+1}\right)\\\cdots\\\textit{Penetration}_k=1-\left(\frac{\alpha}{\alpha+1}\right)^r\frac{b_k}{a_k+b_k}\cdot{}_2F_1\left(r;b_k+1;a_k+b_k+1;\frac{1}{\alpha+1}\right)\\r,\alpha,a_1\ldots a_k>0.\end{cases}\end{aligned} \tag{7}$$

The right-hand side (RHS) of each constraint is the expected penetration from Equation (6) and the left-hand side (LHS) is the observed actual penetration for each site. Note that each firm in the product category would solve formulation (7) using its own observed customer data, with a separate constraint for itself and each of its competitors.

Estimating the $K+2$ parameters may appear to be somewhat daunting because the likelihood function and constraints include the Gaussian hypergeometric function, noted to be cumbersome (Fader and Hardie 2000). However, Appendix 2 presents the *VBA* (Visual Basic for Applications) code we wrote for the ${}_2F_1$ component, which can be integrated into Excel as a customized function. With this customized function, the constrained optimization problem can be easily implemented in Excel using its built-in Solver tool.

4.2. Deriving Competitive Measures

After estimating the parameters for (7), a focal site can then use them to derive any traditional NBD/Dirichlet measures of interest in order to characterize the competitive landscape of the overall category. As noted in the introduction, two popular ones are the market share and the share of wallet for each firm. An important result here is that a focal site can derive these measures analytically for all other firms without requiring any individual customer data from them.

The expected market share for each firm is straightforward and given in (8). This is essentially the ratio of the propensity of a random customer to purchase at a specific site to the propensity to purchase at any site in the category. Let $s = \sum_{j=1}^{k} a_j$ and $b_j = s - a_j$, the market share of site j is:

$$MS_j = \frac{a_j}{s}. \tag{8}$$

Note that MS_j is often available from the same sources that provide the penetration data, so this might not be a big deal by itself. However, we include it here both for generality as well as a way to check the validity of our model. The share of wallet for any firm is shown in (9). Let x_j be the number of purchases of a random customer to site j and n be the total number of purchases to all firms in this category. The share of wallet can be estimated for any firm as $E(x_j/n \mid x_j > 0)$. Unlike the market share, the conditioning is important because share of wallet for a firm is only computed based on its own customers' across-firm purchases ($x_j > 0$). The numerator in (9) represents a random customer's number of purchases to the site of interest. Suppose the customer made n purchases to the category, with the probability of $P(n)$, i.e., the density of the shifted NBD. From the market share definition, given n, we know the number of purchases the random customers make at a specific site is $n \times a/s$. Because n is a random variable, the numerator thus represents the expectation of this number of purchases. The denominator represents the expectation of the total number of category purchases a random customer who has transacted at the site (the conditioning) makes. The probability of the random customer buying at least once at the site is $1 - P(X_j = 0 \mid n)$, and the probability of making n category purchases is $P(n)$. Integrating the expected category purchases over n, we derive the denominator as $\sum_{n=1}^{\infty} P(n)n(1 - P(x_j = 0 \mid n)$. Thus, SoW of site j becomes:

$$SoW_j = E\left(\frac{x_j}{n} \,\middle|\, x_j > 0\right) = \frac{\sum_{n=1}^{\infty} P(n)na_j/s}{\sum_{n=1}^{\infty} nP(n)(1 - P(x_j = 0 \mid n)}$$

$$= ((a_j/s)(r/\alpha + 1)) \cdot \left((a_j/s)(r/\alpha + 1) - \sum_{n=1}^{\infty} n(\Gamma(r + n - 1)\right.$$

$$/\Gamma(r)(n-1)!)(\alpha/(\alpha+1))^r(1/(\alpha+1))^{n-1}(\Gamma(s)\Gamma(b_j + n)$$

$$\left./\Gamma(b)\Gamma(s+n)\right)^{-1}. \tag{9}$$

Appendix 2 presents detailed derivations and the VBA code for customizing SoW as a function with parameters r, α, and $a_1, \ldots, a_k$ as inputs.

Whereas we focus on two principal measures (MS and SoW), there are other competitive measures (Ehrenberg 1988, Ehrenberg et al. 2004) that may also be derived. Appendix 2 also provides the derivations for "frequency," "once only," "100% loyal," and "duplication." We would like to note that several of the key formulas we derived (e.g., SoW, 100% loyal and duplication) are new to the Dirichlet literature.

5. Results

The data we use are provided by comScore Networks, made available through The Wharton School's WRDS service (wrds.wharton.upenn.edu). The raw data consist of 50,000 panelists' online visiting and purchasing activities. This "academic" data set provides detailed session-level data at the customer level for every site within a category (in contrast to the purely site-centric data to which most firms are limited). This enables us to treat each site as the "focal" site and run the model separately for each of them. This also provides us with a much more comprehensive test of the model than if we had only chosen a single site. Furthermore, after first presenting detailed results for one category (online air travel agents), we will then present a summary of results across multiple categories.

The online travel agent category is one of the early adopters and one of most successful industries in e-commerce. Determining various competitive measures is of particular value given the fierce competition in this industry. We selected the top five (based on the number of customer purchases) online travel agents in this category in the entire year of 2007—Expedia (EP), Orbitz (OB), Cheaptickets (CT), Travelocity (TL), and Priceline (PL),[10] and hereafter we use the abbreviations in parentheses to refer to these sites. Because we are interested in estimating competitive measures regarding customer purchases to the various sites, the preprocessed data representing the full-information case (used by the NBD/Dirichlet model) is a count data set where each record represents a specific customer and has six variables—the user ID and five variables representing the total number of purchases that this customer made to each of the five online agents.

5.1. Detailed Results for the Online Air Travel Industry

Table 2 provides a summary of this data including reach (the number of customers who made at least one purchase), penetration, frequency (the average number of purchases per customer who made at least one purchase at the focal site), market share, and SoW. In the spirit of the "double jeopardy" phenomenon in marketing (Goodhardt et al. 1984), we see that firms with higher reach tend to also have higher values for most of the other measures. As seen (and

[10] These sites account for 94% of total visits, 85% of unique users, and 92% of total purchases in the category.

Table 2 Some Observed Competitive Measures for Online Travel

	EP	OB	CT	TL	PL	Category
Reach	1,930	1,578	1,410	1,250	646	6,132
Penetration (%)	31.5	25.7	22.9	20.4	10.5	100
Frequency	1.40	1.33	1.26	1.41	1.37	1.51
Market share (%)	29.3	22.8	19.2	19.2	9.6	100
SoW (%)	84.0	79.9	78.0	81.4	77.6	100

emphasized) quite often in the NBD/Dirichlet literature, the cross-site penetrations have much higher variance than the purchase frequencies.

We first estimate the parameters of the full-information NBD/Dirichlet model using maximum likelihood estimation (MLE). There are $K+2$ parameters to be estimated for each focal site: the r and α parameters for the shifted NBD model of category purchase, and K parameters $(a_1, \ldots, a_k)$ for the brand-choice probabilities across these K sites. The estimated parameters of this full-information model are reported in Table 3. As the expected market shares (a/s) show, the full-information Dirichlet model's market share inferences correspond reasonably well to the actual observed market share measures reported in Table 2.

5.1.1. The LIND Results. The LIND model takes a given firm as the focal site for parameter estimation, but we repeat this estimation procedure for each firm. Thus, there are five separate sets of parameters reported in Table 4, depending on which site is considered as the focal one (in other words, the model is run separately for each one). Each set (i.e., column) consists of the parameter estimates for r, α, and a_1 through a_5 for all five sites. The order of the estimated parameters is consistent across all five sites, regardless of which one is being used as the focal site. At first glance, the magnitudes of the estimated Dirichlet parameters (a_1 to a_5) seem to vary across different focal sites, but this is largely due to inconsequential scaling effects. As we will show shortly (in Table 5), the estimates of market shares, which automatically adjust for these scaling differences (as shown in Equation (8)), are virtually identical across the five focal sites. The magnitude of r and α are adjusted accordingly to reflect the scaling. Also note

Table 3 Parameter Estimates from the Shifted NBD Model and the Dirichlet Model

NBD					
r	0.398				
α	0.788				
	EP	OB	CT	TL	PL
Dirichlet					
a	0.174	0.140	0.124	0.111	0.057
Market share (a/s) (%)	28.7	23.1	20.5	18.3	9.3

Table 4 *LIND* Parameter Estimation with Each Site (Each Column) as the Focal Site

Focal site	EP	OB	CT	TL	PL
r	0.306	0.281	0.495	0.331	0.513
α	0.570	0.590	1.248	0.577	0.933
a (EP)	0.171	0.202	0.225	0.156	0.148
a (OB)	0.139	0.164	0.183	0.126	0.120
a (CT)	0.123	0.146	0.163	0.113	0.107
a (TL)	0.109	0.129	0.144	0.099	0.094
a (PL)	0.056	0.066	0.073	0.051	0.048

that these estimated parameters are comparable to those of the full-information NBD/Dirichlet, which suggests that the proposed LIND model does a good job of recovering the underlying customer behaviors even with a drastically smaller set of input data.

However, as noted above, merely examining the closeness of the parameter estimates does not reveal the performance of individual LIND models. Thus, we further examine the overall fitness by plotting the histogram charts of the observed customer purchases versus the expected purchases from LIND in Figure 1. It is clear that each of the five models works quite well for each focal site.

Based on these parameters shown in Table 4, we then use the formulae described in §4.2 to estimate the market share and share of wallet measures for these five retailers. These results are presented in Tables 5 and 6.

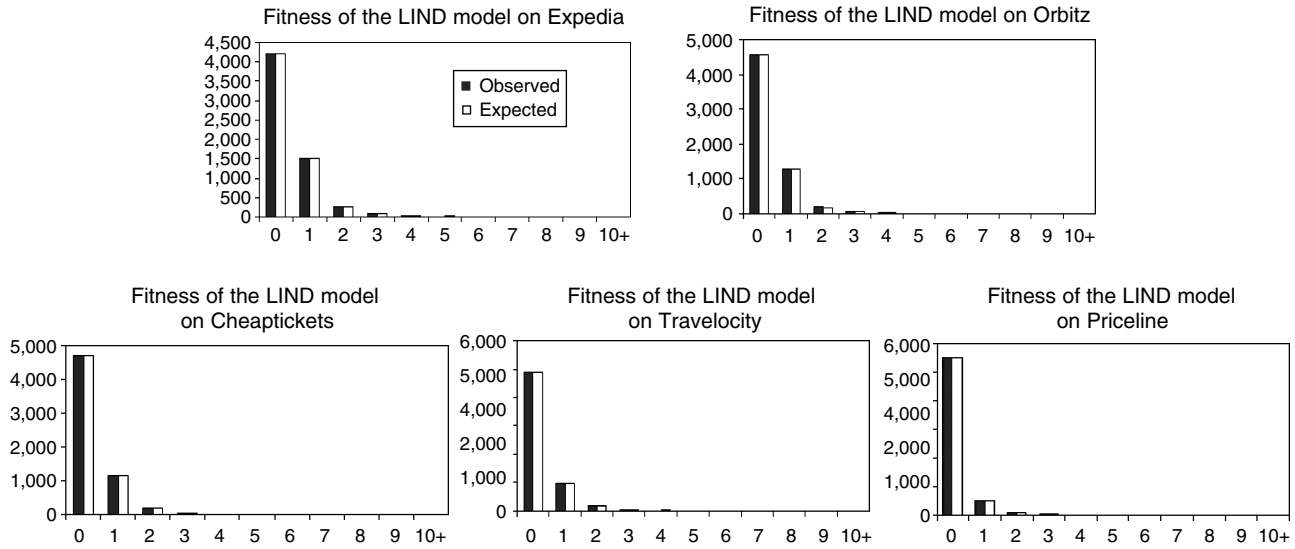
These tables should be interpreted as follows. Each column under "LIND" represents the view of the entire category from each site's perspective. For instance, the results corresponding to the column EP correspond to building a LIND model with Expedia as the focal site (i.e., estimating the model parameters from EP's data plus the constraints) and using this model to compute MS and SoW for the other sites. Hence, the results under column EP would represent Expedia's view of the world as it related to MS and SoW, and so on for the columns OB, CT, TL, and PL. The last two rows of Tables 5 and 6 compute the mean absolute deviation (MAD)[11] of the LIND results

[11] We use MAD as the measure here, following the convention of the literature in NBD/Dirichlet (e.g., Goodhardt et al. 1984, Fader and Schmittlein 1993, Uncles et al. 1995). We also tried some other measures such as average percentage deviation, but the results do not

Table 5 Market Share Results

			LIND (%)				
Brand	Observed	Dirichlet (%)	EP	OB	CT	TL	PL
EP	29.3%	28.7	28.7	28.6	28.6	28.5	28.5
OB	22.8%	23.1	23.2	23.2	23.3	23.2	23.2
CT	19.2%	20.5	20.5	20.7	20.7	20.8	20.7
TL	19.2%	18.3	18.2	18.3	18.2	18.2	18.3
PL	9.6%	9.3	9.3	9.2	9.2	9.3	9.3
MAD	Vs. observed	**0.68**	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**
MAD	Vs. Dirichlet		**0.07**	**0.07**	**0.07**	**0.07**	**0.07**

Figure 1 The Fitness of LIND: Histograms of the Observed vs. Expected



versus the observed values and the LIND results versus the Dirichlet results, respectively.

As these two tables show, the full-information Dirichlet model captures the observed MS and SoW measures well for all firms. This is consistent with prior research that applies the Dirichlet model for other retail markets (Goodhardt et al. 1984, Ehrenberg 1988). For our purpose, we use these results as an upper bound that the use of full information can achieve. For the MS estimation, LIND's results are very close to those of NBD/Dirichlet (within 0.07%), and not far from the actual (within 0.8%). For the SoW results, LIND compares well with the NBD/Dirichlet model (within 1.5%) and the observed values (within 2.0%).

In summary, for the online travel industry, the LIND model performs well compared to the full-information NBD/Dirichlet model, while at the same time using significantly less data than what the NBD/Dirichlet model uses.

5.1.2. Comparing LIND with BB/NBD and BB/sNBD. We further compare the LIND model's overall fitness statistically with those of two benchmark models: BB/NBD and BB/sNBD. We use the Bayesian Information Criteria (BIC) to compare the three models. Notice that all three models were run on the same customer data (in total 6,132 customers) for each of the five sites. Both the BB/NBD and BB/sNBD models use four parameters (r, α, a, b), whereas LIND uses seven parameters (r, α, and five a).

Table 7 presents the BIC values for the three models under each site. Based on the BIC numbers, it is clear that LIND consistently outperforms both BB/sNBD and BB/NBD; the traditional BB/NBD performs worst among the three.

5.2. Robustness Checks

In this section, we conduct two robustness checks.

5.2.1. Results Using Market Share as the Input. So far we considered using penetration to be the inputs for LIND. However, LIND is a flexible model,

change qualitatively (e.g., the LIND model estimates are still within 1% of the Dirichlet estimates). Furthermore, we only use MAD to illustrate the performance of LIND on one category because of the insufficient degrees of freedom to perform a formal statistical analysis. Later, in Table 12, we formally test the differences using multicategory data, where we do not rely on the MAD measure.

Table 6 Share of Wallet Results

			LIND (%)				
Brand	Observed (%)	Dirichlet (%)	EP	OB	CT	TL	PL
EP	84.0%	82.0	81.0	80.7	82.8	81.3	82.9
OB	79.9%	80.7	79.7	79.4	81.6	80.0	81.7
CT	78.0%	80.1	79.0	78.7	81.0	79.4	81.1
TL	81.4%	79.6	78.4	78.1	80.5	78.8	80.6
PL	77.6%	77.5	76.2	75.8	78.5	76.6	78.7
MAD	Vs. observed	**1.35**	**1.72**	**1.92**	**1.52**	**1.54**	**1.57**
MAD	Vs. Dirichlet		**1.15**	**1.48**	**0.83**	**0.81**	**0.99**

Table 7 The Overall Fitness (BIC) of the Three Models

Site	LIND	BB/sNBD	BB/NBD
EP	10,671.2	10,745.9	10,860.9
OB	9,201.2	9,272.5	9,354.5
CT	8,395.4	8,408.1	8,430.1
TL	8,229.4	8,261.3	8,291.7
PL	5,133.2	5,167.7	5,178.1

and other aggregate measures can also be utilized instead; a natural candidate is market share due to its common availability. Here we further investigate the performance of LIND when market share is used as the aggregated inputs. The only modification we need to make in this case is to change the constraints in Equation (7) to ensure the expected market share (formula of which is provided in the appendix) is equal to the actual observed market share. Then we use the estimated LIND parameters to infer penetration and SoW, the results of which are reported in Tables 8 and 9, respectively.

As these two tables show, LIND still performs well, confirming the flexibility and generality of LIND.

5.2.2. Summary Results for Multiple Categories. We also report results from applying the model to a broader set of five online retail categories: online apparel, wireless services, books, office supplies, and travel services. Among the different categories of purchases that comScore Networks identify in their data, these categories are among the most-visited retail categories in the panel and are featured by comScore among key industries[12] for online retailing. Specifically, these five categories reached 73% of unique customers and accounted for 36% of total transactions in the panel.

One caveat concerning the comScore panel data is that there are not many customers who make purchases across multiple sites in these additional categories. In Table 10, we report the percent of 100% loyal customers (those customers who only purchase at a single website) in each category. For the five categories of interest, on average only a little more than 10% of customers purchase at multiple sites. Although this is the reality of the online business, it does not provide a particularly tough test for the LIND model. Accordingly, we decided to switch from purchasing data to site-visit data for these further tests. Competitive measures of visits are important in their own right, e.g., share of visits has a direct impact on advertising revenue. Google's new tool "Google insights for search" has added "share of search" as a measure to analyze searching patterns of users across different competitors. Further, visit patterns capture prepurchase information search and consideration set formation of the consumer. It is well recognized that consideration set formation is important in understanding and explaining consumer-choice behavior (Roberts and Lattin 1997). This is especially critical for the online world because it is possible to capture what sites an individual considered before arriving at a purchase decision. Google Analytics offers a tool for sites to funnel down from site visits to purchasing. Finally, this alternative measure gives us a chance to see the robustness of the LIND model, which helps demonstrate its general applicability to different kinds of behavioral settings.

Table 10 shows that there are plenty of customers who visit multiple websites in each category (along with some interesting and believable differences across the five categories). We apply the LIND model to the visit data from comScore in exactly the same way as we did with the purchase data, using penetration as the inputs. For each category we continue to use the top five sites to maintain consistency. The comScore data show that visits to various online retail categories is highly concentrated, with the top five sites in each category accounting for more than 90% of total visits in the category. In one extreme case (wireless services) the top five sites account for 99.5% of all visits.

To compare the LIND and the Dirichlet estimates, reconsider the results shown for online travel in Tables 5 and 6. Here note that every LIND cell in these tables—and there are 25 of them corresponding to five focal sites—can be compared to the corresponding

[12] See http://www.comscore.com/solutions/is.asp.

Table 8 Penetration Results (with MS as the Input)

			LIND (%)				
Brand	Observed	Dirichlet (%)	EP	OB	CT	TL	PL
EP	31.5%	31.6	**31.7**	32.8	34.3	31.2	31.6
OB	25.7%	25.7	24.9	**25.9**	27.1	24.4	24.7
CT	23.0%	22.9	21.0	21.9	**23.0**	20.6	20.9
TL	20.4%	20.5	21.0	21.9	23.0	**20.6**	20.9
PL	10.5%	10.6	10.6	11.1	11.8	10.4	**10.6**
MAD	Vs. observed	0.07	0.76	0.94	1.63	0.88	0.74
MAD	Vs. Dirichlet		0.68	0.88	1.61	0.85	0.67

Table 9 Share of Wallet Results (with MS as the Input)

			LIND (%)				
Brand	Observed	Dirichlet (%)	EP	OB	CT	TL	PL
EP	84.0%	82.0	81.0	80.7	82.8	81.3	82.9
OB	79.9%	80.7	79.7	79.4	81.6	80.0	81.7
CT	78.0%	80.1	79.0	78.7	81.0	79.4	81.1
TL	81.4%	79.6	78.4	78.1	80.5	78.8	80.6
PL	77.6%	77.5	76.2	75.8	78.5	76.6	78.7
MAD	Vs. observed	1.35	1.67	1.53	1.50	1.21	2.03
MAD	Vs. Dirichlet		1.76	1.55	1.02	0.18	2.18

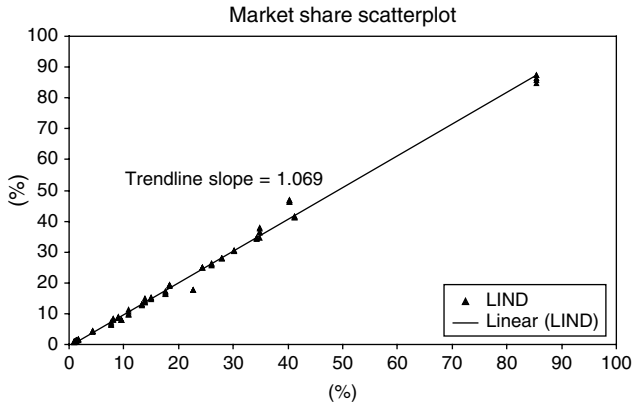
Table 10 Percentage of 100% Loyal Customers

	Apparel (%)	Book (%)	Office (%)	Travel (%)	Wireless (%)	Average
Purchases	83.8	89.2	90.2	90.0	93.6	89.4
Visits	63.9	70.1	67.1	34.2	67.3	60.5

Dirichlet estimates. Extending this to five categories, we obtain 125 specific comparisons for each measure. Figures 2 and 3 summarize the LIND/Dirichlet comparison, with the LIND values on the Y axis and a trend line plotted for comparison.

The market share estimates of the LIND model are very closely aligned with the estimates from the full-information model, and the SoW estimates also line up well, albeit with slightly higher variance.

Figure 2 Market Share Plot for All Five Categories

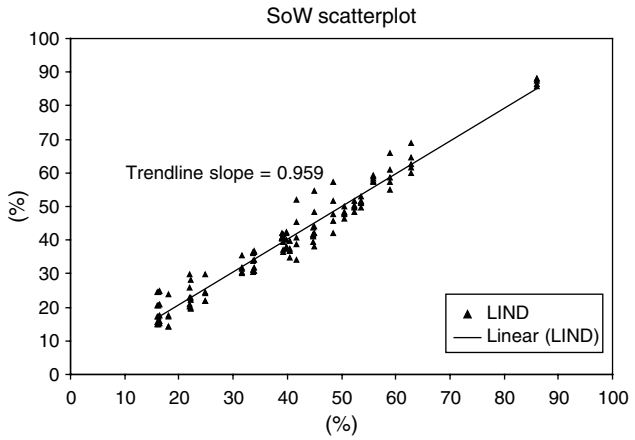


The estimated slopes are significant at 1.06 and 0.96, respectively, again close to the ideal values obtained when the LIND estimates exactly correspond to the Dirichlet estimates.[13] These results are very significant, considering the fact that the full-information Dirichlet model uses detailed customer-level cross-site data for the entire category, whereas LIND only uses one simple, often publicly available, aggregate measure for each competing firm. In the stream of work on estimating competitive measures, to our knowledge this is the first result that shows such good performance with so little data.

Next we present results that compare the Dirichlet, observed, and LIND estimates against each other. Again, we obtain 125 points where LIND can be compared against Dirichlet and observed values. For each comparison, we compute an absolute deviation representing the absolute value of the difference between

[13] Compared to the market share results, the SoW estimates exhibit higher variance. This is expected, as is evident from Equations (8) and (9) (for estimating market share and SoW, respectively). The market share variance is only a function of the variance of parameter a, whereas the SoW variance is amplified by the variance of r and α. Thus, the variability associated with these additional parameter estimates will inflate the variability for the SoW statistics.

Figure 3 SoW Plot for All Five Categories



the LIND and Dirichlet/observed values. Across the 125 points we compute a mean absolute deviation (MAD). Instead of reporting this entire table, we group these into five based on the categories and report the MAD for each category.

Grouping the results into categories also has the advantage that the Dirichlet/observed comparisons will not have "repeats." Specifically, for each category the Dirichlet estimates (and observed values) will always be the same for any focal site. For instance, referring back to Tables 5 and 6, note that the first two columns (Dirichlet and observed values) are independent of which focal site is chosen for parameter estimation.

Table 11 presents the summary of results. First, note that the market share results are particularly good. The MADs for all comparisons are low. The SoW estimates are also good, but not as close as the market share estimates. Specifically, the MAD of LIND versus Dirichlet is only 2.7%. The MAD comparison is an effective method for determining how close the estimates are in absolute terms (Goodhardt et al. 1984). To compute closeness to optimality in relative percentage terms, we compared the percentage differences between the LIND and Dirichlet estimates across all cases. For the market share estimates LIND is within 0.9% of the Dirichlet estimates on average, and for the SoW estimates it is within 2.7%.

We further use a repeated measure ANOVA to test if the results from the LIND "treatment" are significantly different from that of the Dirichlet's. Hence, the LIND and Dirichlet treatments are the two within-subjects factors each site receives. Moreover, all the sites are grouped into five categories, and thus the categories form the between-subjects factors. Table 12 presents the two within-subjects tests for market share and SoW results separately. Both tests show that there is no significant difference under the Dirichlet versus the LIND treatments for inferring market share (p-value $= 0.984$) and SoW ($p = 0.583$). In addition, LIND's market share estimates approach the observed values ($p = 0.987$), but the SoW estimates are still significantly different than the actual observed values ($p = 0.007$). This is also true even for the full-information Dirichlet model ($p = 0.006$), suggesting the difficulty of estimating SoW precisely even under the classic Dirichlet model.

6. Discussion

The model presented in this paper, LIND, is a special case of the NBD/Dirichlet, applicable to a limited data scenario that is prevalent in business. We show that by formulating this as a constrained optimization problem, the derived parameter estimates can be used to intelligently infer several useful competitive measures. Our analysis of online retail data shows that LIND is an effective model that improvises on the well-known Dirichlet model when only limited data are available. We are not aware of any studies that show how transactional data can be combined with aggregate summaries to make important inferences in such a case. This paper presents one novel approach towards this end.

This study is not without limitations. Understanding these limitations is important to know the scope of LIND and how to apply it effectively. Given its lineage, LIND inherits the strengths as well as the limitations of the Dirichlet model. We first discuss the

Table 11 Mean Difference Results and Significance Tests by Category

	Market share			SoW		
Categories	Observed vs. Dirichlet	Observed vs. LIND	Dirichlet vs. LIND	Observed vs. Dirichlet	Observed vs. LIND	Dirichlet vs. LIND
Apparel (%)	1.1	2.4	2.7	4.5	5.0	2.1
Book (%)	0.2	0.4	0.5	3.4	5.3	2.7
Office (%)	1.3	1.3	0.1	6.7	7.5	3.9
Travel (%)	1.2	2.0	0.8	1.9	3.3	2.2
Wireless (%)	1.8	1.9	0.3	7.1	7.6	2.7
Average ($n = 125$) (%)	1.1	1.6	0.9	4.7	5.7	2.7
Deviation ($n = 125$)	0.009	0.014	0.015	0.045	0.046	0.022

Table 12 Results from Repeated-Measure ANOVA

	Market share			SoW		
Categories	Observed vs. Dirichlet	Observed vs. LIND	Dirichlet vs. LIND	Observed vs. Dirichlet	Observed vs. LIND	Dirichlet vs. LIND
Mean square error	7.04E–10	6.40E–08	6.40E–08	1.57E–02	1.81E–02	1.55E–04
F-value	0.000	0.000	0.000	7.924	7.657	0.304
P-value	1.000	0.987	0.984	0.006	0.007	0.583

limitations common to the Dirichlet family of models and the conditions under which these models work well (or not). We then discuss the limitations specific to LIND.

6.1. Limitations Common to the Dirichlet Family of Models

As we discussed in §3.2, the Dirichlet model requires the market to be stable, two key conditions of which are (1) the market is stationary and (2) the market is nonsegmented (Ehrenberg et al. 2004, Goodhardt et al. 1984, Ehrenberg 1988). The resultant market exhibits the prevalent "double jeopardy" phenomenon.

First, the stationarity condition is mainly a result of the Poisson assumption for the category purchase. Poisson distribution is memoryless, i.e., a customer's purchase propensity does not change across different time periods and the last-period purchase has no influence on a customer's next purchase. In a nonstable market, a customer may change her pace of purchase across different time periods. Thus the conditional mean of the future-period purchase may no longer be a linear projection of the previous periods, the defining feature of the NBD model (Schmittlein et al. 1985). This is less of a problem when the marketer's interest is to perform diagnostic analysis on the current period of data as we do in this paper. When the endeavor moves to predicting future-period customer behavior, violation of this condition looms larger and will have an adverse effect on the Dirichlet model's (and LIND's) performance. Unfortunately, there is no standard treatment of nonstationarity in the literature. Some early attempts include Fader and Lattin (1993) and Jeuland et al. (1980). For example, Jeuland et al. (1980) replaced the stationary Poisson assumption with the Erlang-2 distribution, which processes records at every second arrival to allow for change of pace among customer purchases over time. However, the trade-off is that we pay a big statistical price for the added generality with limited improvement of these complex extensions (Morrison and Schmittlein 1988).

Second, the performance of the Dirichlet family of models is subject to the nonsegmented market condition. However, a fully disparate, segmented market may not be best treated as a whole market in the first place. We may need to redefine the relevant market to be each individual segment (i.e., applying a separate Dirichlet model to each segment). If we force these segmented markets to be one big market, the double jeopardy property may not hold. For example, in a niche market, the segment may cater to excessive loyal customers who purchase more frequently than what is expected by the Dirichlet model in a normal market. There are some attempts to extend the Dirichlet model to account for segmented markets. Danaher et al. (2003) used a latent class model approach where latent segments are modeled as a finite mixture model as a function of additional information such as price, promotion, etc. These are excellent approaches that shed light on future extensions of LIND to such markets.

Third, it is not easy to incorporate market-mix variables (e.g., price and promotion) or the individual customer-level variable (e.g., demographics) into the Dirichlet-type models (e.g., Bhattacharya 1997). This work avoids the complex task of directly integrating covariates into the Dirichlet model by using a two-step modeling procedure. First, a regular Dirichlet model is built and then competitive measures such as SCR (share of category requirement) are estimated using the Dirichlet parameters. Deviations of these estimates from the actual SCRs are then calculated. Subsequently, in the second step, a linear regression analysis is performed by regressing these deviations on the marketing-mix variables. Bhattacharya (1997, p. 431) maintains that "it is better to study and explain these deviations *outside* the Dirichlet model, rather than try to make this flexible, elegant and parsimonious model more complex." To our knowledge, a generalized Dirichlet model that directly integrates market mix or customer level of data has not been developed in the literature, making it an interesting topic for future research.

Fourth, the Dirichlet model is elegant, but rather complex, because it mixes Poisson, Gamma, multinomial, and Dirichlet distributions. It is hard to tweak the model with other alternative distributional assumptions. Even a small extension often ends up with an intractable model. Although there are some early attempts, none supersedes the NBD/Dirichlet model and it still stands as one of the most popular and successful models in the repeat-buying literature. Some notable attempts include the condensed

NBD model (Jeuland et al. 1980) with an Erlang-2 distribution (instead of Poisson); the log-normal mixing distribution (Lawrence 1980); and the generalized inverse Gaussian (Sichel 1982) to replace the Gamma mixing distribution. An analysis of the Dirichlet model's sensitivity to these distributional assumption is provided in Fader and Schmittlein (1993).

We would like to conclude this section with the argument of Ehrenberg et al. (2004, p. 1312) that "such known discrepancies (with the stable-market assumptions) seldom curtail the model's practical use. Attempts to improve or elaborate the Dirichlet have so far not resulted in major gains in either predictive power or parsimony." Just as Morrison and Schmittlein (1988, p. 152) elegantly argued that although there is only one way for a process to remain stable, there are of course an infinite variety of possible nonstable behaviors. Thus, not very much can be said in general about the effects of instability. It is beyond the scope of a single paper to thoroughly examine these generalizations. In future research we plan to extend LIND to such a market.

Moreover, even in nonstable markets, the Dirichlet model still serves as an important benchmark model for managers to detect whether there is deviation from the norm predicted by the model and where the deviations come from. For example, Fader and Schimittlein (1993) use the Dirichlet benchmark to identify niche-market brands—those exhibit higher loyalty than predicted by the Dirichlet. Kahn et al. (1988, p. 385) assert that although deviations from norms are plausible, "they are difficult to apply without a relevant benchmark for comparison."

6.2. LIND-Specific Issues

LIND can be conceived as a realistic version of the Dirichlet model for the limited data scenario. In terms of implementing LIND, the user needs to first choose appropriate known inputs to infer the unknown outputs (i.e., competitive measures) of interest. Not all inputs are equally useful. We identify three conditions to ensure the proper selection of an input.

First, the chosen input should be indicative of the market. Measures reflecting the attractiveness of the brand, such as penetration and market share, are informative of the structure of the market. These measures tend to have higher variation across brands (e.g., the penetration of the biggest brand Expedia is almost three times that of Priceline, the smallest one in our data). Measures on the intensity of customer purchase (e.g., frequency) and the loyalty of customers (e.g., SoW, 100% loyalty and duplication) tend to vary less and hence are less indicative of the market structure (Ehrenberg et al. 2004). For example, the frequency of Expedia is 1.4 versus 1.37 for Priceline; SoW of expedia is 84% versus 77% for Priceline. Thus, inputs such as frequency and SoW, which do not vary much across brands, may not serve as good candidate inputs. This is echoed in Ehrenberg (2000, p. 188) that "it is penetration, not the average purchase level (i.e., frequency) that determines the sales level of a brand . . . ."

Second, inputs that are expected to have high correlation with the outputs are preferred. Not surprisingly, our results indicate that the higher the correlation, the better the model performs. For example, in the travel industry, the (Spearman) correlation between penetration and market share is 0.97, that between market share and SoW is 0.82, and that between penetration and SoW is 0.7. The results show that LIND works best when using penetration to estimate market share, followed by using market share to estimate SoW and then using penetration to estimate SoW.

Third, in a stable market, the rank order of the outputs is generally expected to follow that of the inputs. In a stable market exhibiting "double jeopardy," the rank order of several key measures such as penetration and frequency are expected to be exactly the same. This is described in the seminal paper of Goodhardt et al. (1984, p. 623) that "this is known as a Double Jeopardy pattern (for two brands X and Z): Z not only has fewer buyers than X, but they also buy it (slightly) less often." The double-jeopardy rule is further summarized in Ehrenberg (2000, p. 186) as $w(1-b)=c$, where w is the frequency of a brand (among all customers in the market), b is the penetration of the brand, and c is a constant. This implies that the larger the brand (in terms of penetration), the larger the frequency and, in turn, the larger the market share (because it is a function of $w \times b$). In other words, market share is expected to exhibit the same rank order as penetration in the Dirichlet world.

The same expected rank order of penetration, market share, and SoW (the three measures demonstrated in this paper) can be shown analytically. Everything else being equal, a bigger brand in LIND (also Dirichlet) will yield a higher value in parameter a_j (an indication of customer purchase propensity to brand j). It is easy to verify that the following measures would be a monotonic-increasing function of a_j.

(i)

$$MarketShare = \frac{a_j}{\sum_{j=1}^{k} a_j} = \frac{a_j}{s},$$

clearly the first derivative with respect to a_j is positive.

(ii)

$$Penetration_j = 1 - P(X_j = 0) = 1 - \left(\frac{\alpha}{\alpha+1}\right)^r \frac{b_j}{s} {}_2F_1\left(r, b_j+1, s+1, \frac{1}{\alpha+1}\right),$$

because ${}_2F_1(a_j)$ is a monotonic-decreasing function of a_j, and b_j/s is also decreasing as a function of a_j, penetration increases as a_j increases.

(iii)

$$SoW = ((a_j/s)(r/\alpha+1)) \cdot \Bigg((a_j/s)(r/\alpha+1) - \sum_{n=1}^{\infty} (\Gamma(r+n-1))/\Gamma(r)(n-1)!) \cdot (\alpha/(\alpha+1))^r (1/(\alpha+1))^{n-1} \cdot n(\Gamma(s)\Gamma(b_j+n)/\Gamma(b)\Gamma(s+n)) \Bigg)^{-1}$$

is also an increasing function of a_j. Let

$$c1 = \left(\frac{r}{\alpha}+1\right),$$

$$c2 = \sum_{n=1}^{\infty} \frac{\Gamma(r+n-1)}{\Gamma(r)(n-1)!} \left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^{n-1} \cdot n \frac{\Gamma(s)\Gamma(b_j+n)}{\Gamma(b)\Gamma(s+n)}.$$

Then $SoW'(a_j) = [C1/(C1 - C2(1 + b_j/a_j))]' > 0$.

In sum, all these three measures exhibit the same rank order. The ideal market for the Dirichlet world is a triple-jeopardy world: a large brand not only attracts more purchases per customer, but more-loyal customers as well.

7. Conclusion

In this paper we addressed the problem of estimating important competitive measures from a single focal site's point of view. Given a realistic assumption that each firm has its own data and is able to obtain the penetration or market share numbers of its leading competitors, we develop a constrained optimization model based on a well-established theory of consumer behavior. Solving this problem from each focal site's perspective provides parameter estimates for the proposed LIND model. We then show how these parameter estimates can be used to derive analytical expressions for the market share and share of wallet. Testing our approach on various online retail categories shows that the LIND model performs surprisingly well considering how little information it uses. In summary, in this paper we developed a method for making competitive inferences that is (i) practical, in that it relies on sharing simple aggregate data; (ii) effective, in that it can enable the determination of useful competitive measures such as market share and share of wallet; and (iii) explainable in that it is based on a well-studied theory.

The problem of making inferences about these competitive measures is one that is important in any industry. Although our application domain here was online retail, our methods are certainly not restricted to it, and can be applied in more traditional markets as well. Difficulties in obtaining cross-firm panel data are widespread, as are sources of aggregate competitive measures such as SoW. The emergent competitive intelligence field has emphasized the importance of competitive measures. We expect to see the next-generation dashboard to incorporate some of these measures. Google's new tool "Google insights for search," released in 2008, has added competitive intelligence analysis, focusing on the searching patterns of users across different competitors (http://www.google.com/insights/search/#). Their concept of "share of search" is what we demonstrate in the robustness check section. Microsoft's "adCenter Labs" also provides some Web analytics functions that track a firm's performance relative to its competitors. The social media analytics tool release in 2010 has added the functionalities to analyze the share of buzz in the social media across different brands (http://www.sas.com/software/customer-intelligence/social-media-analytics/).

As one more example, the Lundberg Survey is an authoritative source for information about gasoline prices and competitive measures in the petroleum industry, used heavily by investors and industry experts to understand and forecast purchasing patterns in the industry. Data collection is extremely complex, based on a laborious biweekly sampling of 7000 of the 133,000 gas stations in the United States. However, even this survey cannot obtain SoW estimates because that would require linking customer-level information across gas stations. Thus, the method presented in this paper is of particular managerial relevance given the prevalence of such scenarios in this and many other industries.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/isre.1110.0385.

Appendix A. Deriving the BB/sNBD Model

A.1. Deriving the Conditional Probability $P(X = x \mid \lambda, p)$

Based on the data assumption (see §4) that firms only observe those users making at least one purchase to the category, we model the category purchase N as a shifted Poisson with parameter λ as the purchase rate of a random customer in the unit time interval.

$$P(N = n \mid \lambda) = \frac{\lambda^{n-1} e^{-\lambda}}{(n-1)!}, \quad n = 1, 2 \ldots; \ \lambda > 0 \qquad \text{(A1)}$$

Given the category purchase n of a random customer, the choice of a focal brand is modeled as a binomial distribution with rate p.

$$P(X=x \mid n,p)=\binom{n}{x}p^x(1-p)^{n-x},\quad x=0,1,\ldots n;\ 0\le p\le 1. \tag{A2}$$

It is assumed that an individual's choice rate p is independent of the purchase rate λ. Thus, $P(X=x \mid \lambda,p)$, the conditional probability of a customer making a purchase to the focal site given λ and p, is a BB/sNBD (beta-binomial/shifted negative binomial distribution) model.

First consider the special case where $x=0$. Because firms share the aggregated data—the number of customers they reached—each firm knows the number of customers who purchased the category, but not the focal firm. We have

$$\begin{aligned}P(x=0\mid\lambda,p)&=\sum_{n=1}^{\infty}\frac{\lambda^{n-1}e^{-\lambda}}{(n-1)!}(1-p)^n\\&=e^{-\lambda p}(1-p).\end{aligned} \tag{A3}$$

Notice that the sum in (3) starts with $n=1$ (because the data are truncated at $n\ge 1$) instead of $n=x$. In the more general case when $x\ge 1$, we have

$$\begin{aligned}P(x\mid\lambda,p)&=\sum_{n=x}^{\infty}\frac{\lambda^{n-1}e^{-\lambda}}{(n-1)!}\binom{n}{x}p^x(1-p)^{n-x}\\&=\frac{e^{-\lambda}p^x}{x!}\sum_{n=x}^{\infty}\frac{\lambda^{n-1}}{(n-1)!}\frac{n!(1-p)^{n-x}}{(n-x)!}\\&=\frac{e^{-\lambda}p^x\lambda^{x-1}}{x!}\sum_{n=x}^{\infty}\frac{n\lambda^{n-x}(1-p)^{n-x}}{(n-x)!}.\end{aligned} \tag{A4}$$

Let $y=n-x,$. Equation (A4) is transformed into

$$\begin{aligned}P(x\mid\lambda,p)&=\frac{e^{-\lambda}p^x\lambda^{x-1}}{x!}\sum_{y=0}^{\infty}\frac{(x+y)\lambda^y(1-p)^y}{y!}\\&=\frac{e^{-\lambda p}p^x\lambda^{x-1}}{x!}\sum_{y=0}^{\infty}(x+y)\frac{[\lambda(1-p)]^y e^{-\lambda(1-p)}}{y!}.\end{aligned}$$

Notice that

$$\sum_{y=0}^{\infty}\frac{[\lambda(1-p)]^y e^{-\lambda(1-p)}}{y!}=1\quad\text{and}$$

$$\sum_{y=0}^{\infty}y\frac{[\lambda(1-p)]^y e^{-\lambda(1-p)}}{y!}=E(y)=\lambda(1-p),$$

thus, when $x\ge 1$,

$$\begin{aligned}P(x\mid\lambda,p)&=\frac{e^{-\lambda p}p^x\lambda^{x-1}}{x!}[x+\lambda(1-p)]\\&=\delta_{x>0}\frac{e^{-\lambda p}p^x\lambda^{x-1}}{(x-1)!}+\frac{e^{-\lambda p}p^x\lambda^x(1-p)}{x!}\end{aligned} \tag{A5}$$

where $\delta_{x>0}=1$ if $x>0$ and 0 otherwise.

A.2. Deriving the Marginal Probability $P(X=x)$

To capture consumer heterogeneity, the distribution of the purchase frequency λ across the population is assumed to be distributed as gamma with *pdf*

$$f(\lambda)=\frac{\alpha^r\lambda^{r-1}e^{-\lambda\alpha}}{\Gamma(r)},\quad \lambda>0. \tag{A6}$$

Similarly, the distribution of consumer choice p across the population is distributed as beta with *pdf*

$$g(p)=\frac{1}{\mathrm{B}(a,b)}p^{a-1}(1-p)^{b-1},\quad 0\le p\le 1 \tag{A7}$$

where $\mathrm{B}(a,b)$ is the beta function.

Further, the following derivation uses the Euler's integral representation of the Gaussian hypergeometric function ${}_2F_1(\cdot)$ as explained in Footnote 4.

$$\begin{aligned}&\int_0^1 t^a(1-t)^b(u+vt)^{-c}\,dt\\&=\begin{cases}\mathrm{B}(a+1,b+1)u^{-c}{}_2F_1\left(c,a+1,a+b+2,-\dfrac{v}{u}\right),\\\quad |v|<u\\[2ex]\mathrm{B}(a+1,b+1)(u+v)^{-c}{}_2F_1\left(c,b+1,a+b+2,\dfrac{v}{u+v}\right),\\\quad |v|\ge u\end{cases}\end{aligned} \tag{A8}$$

Further, the Gaussian hypergeometric function converges when the last argument is no greater than 0 and diverges otherwise.

A.2.1. Case 1: $X=0$. In this case, from (3), the marginal probability becomes

$$\begin{aligned}p(X=0)&=\int_0^1\int_0^{\infty}e^{-\lambda p}(1-p)\frac{\lambda^{r-1}\alpha^r e^{-\lambda\alpha}}{\Gamma(r)}\frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)}\,d\lambda\,dp\\&=\frac{\alpha^r}{\Gamma(r)\mathrm{B}(a,b)}\int_0^1 p^{a-1}(1-p)^b\left[\int_0^{\infty}e^{-\lambda(p+\alpha)}\lambda^{r-1}d\lambda\right]dp\\&=\frac{\alpha^r}{\Gamma(r)\mathrm{B}(a,b)}\int_0^1 p^{a-1}(1-p)^b(p+\alpha)^{-r}\,dp.\end{aligned}$$

Further necessary transformation is needed to ensure the convergence of the Gaussian hypergeometric function as noted above. Let $q=1-p$, which implies $dp=-dq$, we have

$$\begin{aligned}&\int_0^1 p^{a-1}(1-p)^b(p+\alpha)^{-r}\,dp\\&\quad=-\int_1^0(1-q)^{a-1}(q)^b(1-q+\alpha)^{-r}\,dq\\&\quad=\int_0^1(1-q)^{a-1}(q)^b(1+\alpha-q)^{-r}\,dq.\end{aligned}$$

It is clear that $(1+\alpha)>1$, and based on the Euler's integral given in (8), we have

$$\begin{aligned}&p(X=0)\\&=\frac{\alpha^r\Gamma(r)}{\Gamma(r)\mathrm{B}(a,b)}\frac{\mathrm{B}(a+x,b+1)}{(\alpha+1)^r}{}_2F_1\left(r,b+1,a+b+1,\frac{1}{\alpha+1}\right)\\&=\left(\frac{\alpha}{\alpha+1}\right)^r\frac{b}{a+b}\times{}_2F_1\left(r,b+1,a+b+1,\frac{1}{\alpha+1}\right).\end{aligned} \tag{A9}$$

A.2.2. Case 2: $X > 0$**.** In the more general case when $X > 0$,

$$P(X=x) = \int_0^1 \int_0^\infty p(X=x \mid \lambda, p) f(\lambda) g(p)\, d\lambda\, dp$$
$$= \int_0^1 \int_0^\infty \left[\delta_{x>0} \frac{e^{-\lambda p} p^x \lambda^{x-1}}{(x-1)!} + \frac{e^{-\lambda p} p^x \lambda^x (1-p)}{x!} \right] \cdot \frac{\lambda^{r-1} \alpha^r e^{-\lambda\alpha}}{\Gamma(r)} \frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)}\, d\lambda\, dp. \quad \text{(A10)}$$

For convenience, we separate the following derivation for Equation (10) into two terms. Term 1 consists of

$$\int_0^1 \int_0^\infty \delta_{x>0} \frac{e^{-\lambda p} p^x \lambda^{x-1}}{(x-1)!} \frac{\lambda^{r-1} \alpha^r e^{-\lambda\alpha}}{\Gamma(r)} \frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)}\, d\lambda\, dp,$$

and term 2 consists of

$$\int_0^1 \int_0^\infty \frac{e^{-\lambda p} p^x \lambda^x (1-p)}{x!} \frac{\lambda^{r-1} \alpha^r e^{-\lambda\alpha}}{\Gamma(r)} \frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)}\, d\lambda\, dp.$$

Term* 1 *derivation.

Term 1
$$= \frac{\alpha^r}{(x-1)!\Gamma(r)\mathrm{B}(a,b)} \int_0^1 p^{(x+a-1)}(1-p)^{b-1} \cdot \left[\int_0^\infty e^{-\lambda(p+\alpha)} \lambda^{(x+r-2)}\, d\lambda \right] dp$$
$$= \frac{\alpha^r}{(x-1)!\Gamma(r)\mathrm{B}(a,b)} \int_0^1 p^{(x+a-1)}(1-p)^{b-1} \cdot [-(p+\alpha)^{-(x+r-1)} \Gamma(r+x-1, (p+\alpha)\lambda)]_0^\infty]\, dp$$
$$= \frac{\alpha^r \Gamma(r+x-1)}{(x-1)!\Gamma(r)\mathrm{B}(a,b)} \int_0^1 p^{(x+a-1)}(1-p)^{b-1}(p+\alpha)^{-(r+x-1)}\, dp$$

where $\Gamma(a,x) = \int_x^\infty t^{a-1} e^{-t} dt$ is the incomplete Gamma function. Again let $q = 1 - p$, which implies $dp = -dq$; we then have

$$\int_0^1 p^{(x+a-1)}(1-p)^{b-1}(p+\alpha)^{-(r+x-1)}\, dp = -\int_1^0 (1-q)^{(x+a-1)} q^{b-1} (1+\alpha-q)^{-(r+x-1)}\, dq = \int_0^1 (1-q)^{(x+a-1)} q^{b-1} (1+\alpha-q)^{-(r+x-1)}\, dq.$$

Clearly, we have $(1+\alpha) > 1$, and based on (8) the above integral yields

$$\int_1^0 (1-q)^{(x+a-1)} q^{b-1} (1+\alpha-q)^{-(r+x-1)}\, dq = \mathrm{B}(b, x+a)(1+\alpha)_2^{-(r+x-1)} F_1\left(x+r-1, b, a+b+x, \frac{1}{\alpha+1} \right),$$

and thus

$$\textit{Term } 1 = \frac{\alpha^r \Gamma(r+x-1)}{(x-1)!\Gamma(r)\mathrm{B}(a,b)} \frac{\mathrm{B}(a+x,b)}{(\alpha+1)^{(x+r-1)}} \cdot {}_2F_1\left(x+r-1, b, a+b+x, \frac{1}{\alpha+1} \right)$$
$$= \frac{\alpha^r}{(\alpha+1)^{(x+r-1)}} \frac{\Gamma(r+x-1)}{(x-1)!\Gamma(r)} \frac{\Gamma(a+x)\Gamma(b+a)}{\Gamma(a)\Gamma(a+b+x)} \cdot {}_2F_1\left(x+r-1, b, a+b+x, \frac{1}{\alpha+1} \right). \quad \text{(A11)}$$

Term* 2 *derivation.

Term 2
$$= \int_0^1 \int_0^\infty \frac{e^{-\lambda p} p^x \lambda^x (1-p)}{(x)!} \frac{\lambda^{r-1} \alpha^r e^{-\lambda\alpha}}{\Gamma(r)} \frac{p^{a-1}(1-p)^{b-1}}{\mathrm{B}(a,b)} d\lambda\, dp$$
$$= \frac{\alpha^r}{x!\Gamma(r)\mathrm{B}(a,b)} \int_0^1 p^{a+x-1}(1-p)^b \left[\int_0^\infty e^{-\lambda(p+\alpha)} \lambda^{x+r-1} d\lambda \right] dp$$
$$= \frac{\alpha^r \Gamma(r+x)}{x!\Gamma(r)\mathrm{B}(a,b)} \int_0^1 p^{a+x-1}(1-p)^b (p+\alpha)^{-(x+r)}\, dp$$
$$= \frac{\alpha^r \Gamma(r+x)}{x!\Gamma(r)\mathrm{B}(a,b)} \frac{B(a+x, b+1)}{(\alpha+1)^{(x+r)}} \cdot {}_2F_1\left(x+r, b+1, a+b+x+1, \frac{1}{\alpha+1} \right)$$
$$= \frac{\alpha^r}{(\alpha+1)^{(x+r)}} \frac{\Gamma(r+x)}{x!\Gamma(r)} \frac{\Gamma(a+x)\Gamma(b+1)\Gamma(a+b)}{\Gamma(a+b+1+x)\Gamma(a)\Gamma(b)} \cdot {}_2F_1\left(x+r, b+1, a+b+x+1, \frac{1}{\alpha+1} \right). \quad \text{(A12)}$$

The final result of the marginal probability $P(X=x)$ is a combination of the two terms, which is

$$P(X=x) = \frac{\alpha^r}{(\alpha+1)^{(x+r-1)}} \frac{\Gamma(r+x-1)}{(x-1)!\Gamma(r)} \frac{\Gamma(a+x)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+x)} \cdot {}_2F_1\left(x+r-1, b, a+b+x, \frac{1}{\alpha+1} \right) + \frac{\alpha^r}{(\alpha+1)^{(x+r)}} \frac{\Gamma(r+x)}{x!\Gamma(r)} \frac{\Gamma(a+x)\Gamma(b+1)\Gamma(a+b)}{\Gamma(a+b+1+x)\Gamma(a)\Gamma(b)} \cdot {}_2F_1\left(x+r, b+1, a+b+x+1, \frac{1}{\alpha+1} \right). \quad \text{(A13)}$$

This is the shifted BB/sNBD model, which is the baseline model of LIND.

Appendix B. Derivation of the Various Competitive Measures

Here we derive the various competitive measures for the LIND model. The notations used here follow the definitions in Table 1.

B.1. Market Share

Market share is equal to the ratio of the focal site's purchases to the category purchases, and thus:

$$\textit{MarketShare} = \frac{a_j}{\sum_{j=1}^k a_j} = \frac{a_j}{s}.$$

B.2. Share of Wallet (SoW)

SoW is equal to the ratio of the total purchases at a focal brand over the total category purchases of those customers who made purchase at the focal site. Suppose the category purchase n is known, then the total purchases at a focal is simply $n \times a_i/s$, where $s = \sum_j a_j$. Moreover, $P(n)$ follows a shifted NBD process. Hence, the numerator of the ratio is simply $\sum_{n=1}^\infty P(n) \times n \times a_j/s$. The probability of customers who made at least one purchase at the focal site is $1 - P(x_j = 0 \mid n)$, where $P(x/n)$ follows a binomial process. The denominator of the ratio thus is $\sum_{n=1}^\infty P(n) \times n \times (1 - P(x_j = 0 \mid n))$.

$$SoW = E\left(\frac{x_j}{n} / x_j > 0\right)$$

$$= \frac{\sum_{n=1}^{\infty} P(n) n a_j / s}{\sum_{n=1}^{\infty} P(n) n (1 - P(x_j = 0 \mid n))}$$

$$= \frac{\sum_{n=1}^{\infty} \frac{\Gamma(r+n-1)}{\Gamma(r)(n-1)!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^{n-1} n a_j / s}{\sum_{n=1}^{\infty} \frac{\Gamma(r+n-1)}{\Gamma(r)(n-1)!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^{n-1} n\left(1 - \frac{\mathrm{B}(a_j+0, b_j+n+0)}{\mathrm{B}(a_j, b_j)}\right)}$$

$$= \frac{\frac{a_j}{s}\left(\frac{r}{\alpha}+1\right)}{\frac{a_j}{s}\left(\frac{r}{\alpha}+1\right) - \sum_{n=1}^{\infty} \frac{\Gamma(r+n-1)}{\Gamma(r)(n-1)!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^{n-1} n \frac{\Gamma(s)\Gamma(b_j+n)}{\Gamma(b)\Gamma(s+n)}}.$$

Solving the above expression analytically is nontrivial. We use a numerical approach instead that integrates over n from 1 to a sufficiently large number (determining what large number is sufficient depends on the desirable precision one needs. In our implementation, we use 100, which reaches a precision 10^{-8}). This approach can be easily implemented in Excel using a VBA Macro. The following VBA Macro defines a customized function SoW(a, b, r, α) in excel.

```
  Public Function SoW(a As Double, b As Double, r As
Double, alpha As Double) As Double
  Dim n, x As Integer
  Dim numerator, denominator, pn, px0 As Double
    Temp, pn_sum, numerator, denominator = 0
  For n = 1 To 100
  With Application
    pn = Exp(.GammaLn(r + n-1) - .GammaLn(r)) /
.Fact(n-1) * (alpha / (alpha + 1))  r * (1 / (alpha +
1))  (n-1)
    numerator = numerator + pn * n * a / (a + b)
    px0 = Exp(.GammaLn(b + n) + .GammaLn(a + b) -
.GammaLn(a + b + n) - .GammaLn(b))
  denominator = denominator + pn * n * (1-px0)
    End With
  Next n
  SoW = numerator / denominator
  End Function
```

B.3. Penetration

Site j's expected penetration is equal to 1 minus the expected percentage of customers who did not purchase the site and thus

$$Penetration_j = 1 - P(X_j = 0)$$

$$= 1 - \left(\frac{\alpha}{\alpha+1}\right)^r \frac{b_j}{s}\, {}_2F_1\left(r, b_j+1, s+1, \frac{1}{\alpha+1}\right).$$

B.4. Frequency

Expected frequency of all customers with respect to site j is given below. Notice that the numerator is the same as the numerator in the SoW expression above, which represents a random customer's number of purchases to the site of interest. The denominator represents the probability that a random customer (in the category) would purchase the focal site.

$$Frequency = E(X_j)$$

$$= \sum_{n=1}^{\infty} P(n) n a_j / s$$

$$= \sum_{n=1}^{\infty} \frac{\Gamma(r+n-1)}{\Gamma(r)(n-1)!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^{n-1} n a_j / s.$$

Letting $n = m+1$, we can transform the above summation into

$$E(x_j) = \sum_{m=0}^{\infty} \frac{\Gamma(r+m)}{\Gamma(r)m!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^m (m+1) a_j / s$$

$$= \frac{a_j}{s} \times \left[\sum_{m=0}^{\infty} \frac{\Gamma(r+m)}{\Gamma(r)m!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^m m + \sum_{m=0}^{\infty} \frac{\Gamma(r+m)}{\Gamma(r)m!}\left(\frac{\alpha}{\alpha+1}\right)^r \left(\frac{1}{\alpha+1}\right)^m\right]$$

$$= \frac{a_j}{s}\left(\frac{r}{\alpha}+1\right).$$

Note that the frequency of buyers of site j (those customers who made at least one purchase) is simply the above frequency divided by penetration. That is,

$$E(x_j \mid x_j > 0) = \frac{E(x_j)}{Penetration}$$

$$= \frac{(a_j/s)(r/\alpha+1)}{1-(\alpha/(\alpha+1))^r (b_j/s)\, {}_2F_1(r, b_j+1, s+1, 1/(\alpha+1))}$$

B.5. Once Only

This represents the marginal probability that a random customer purchased the focal site exactly once.

$$P(X_j = 1)$$

$$= \frac{\alpha^r}{(\alpha+1)^r} \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+1)}\, {}_2F_1\left(r, b, a+b+1, \frac{1}{\alpha+1}\right) + \frac{\alpha^r}{(\alpha+1)^{(1+r)}} \frac{\Gamma(r+1)}{\Gamma(r)} \frac{\Gamma(a+1)\Gamma(b+1)\Gamma(a+b)}{\Gamma(a+b+2)\Gamma(a)\Gamma(b)} \cdot {}_2F_1\left(1+r, b+1, a+b+2, \frac{1}{\alpha+1}\right)$$

$$= \left(\frac{\alpha}{\alpha+1}\right)^r \frac{a}{a+b}\, {}_2F_1\left(r, b, a+b+1, \frac{1}{\alpha+1}\right) + \left(\frac{\alpha}{\alpha+1}\right)^r \frac{r}{\alpha+1} \frac{ab}{(a+b)(a+b+1)} \cdot {}_2F_1\left(1+r, b+1, a+b+2, \frac{1}{\alpha+1}\right).$$

B.6. 100% Loyal

This measure represents the expected percentage of customers that only purchase the focal site in the category. This

is equivalent to saying that the purchases to all other brands are zero.

$$
\begin{aligned}
&P((X_j = n \mid (N = n, X_j > 0))\\
&= \frac{P(X_j = n \mid N = n, X_j > 0)}{P(X_j > 0)}\\
&= \frac{P(X_j = n \mid N = n) - P(X_j = n \mid N = n, X_j = 0)}{P(X_j > 0)}\\
&= \frac{P(X_j = n \mid N = n) - 0}{P(X_j > 0)}\\
&= \frac{\sum_{n=1}^{\infty} \frac{\Gamma(r+n)}{n!}\left(\frac{1}{\alpha+1}\right)^n \frac{\Gamma(s)\Gamma(a_j+n)}{\Gamma(s+n)\Gamma(a_j)}}{\sum_{n=1}^{\infty} \frac{\Gamma(r+n)}{n!}\left(\frac{1}{\alpha+1}\right)^n (1 - \frac{\Gamma(s)\Gamma(b_j+n)}{\Gamma(s+n)\Gamma(b_j)})}.
\end{aligned}
$$

Duplication Between Two Brands X and Y: Customers Who Bought X Also Bought Y

$$
\begin{aligned}
Duplication(X, Y) &= P(Y > 0 \mid X > 0)\\
&= \frac{P(Y > 0, X > 0)}{P(X > 0)}\\
&= \frac{1 - P(XY = 0)}{P(X > 0)}\\
&= \frac{1 - P(X = 0) - P(Y = 0) + P(X + Y = 0)}{P(X > 0)}.
\end{aligned}
$$

The only difficult part in the above expression is $P(X + Y = 0)$. This can be done by creating a composite (and fictitious) $X + Y$ brand. Because the Dirichlet world assumes independence between X and Y, the parameters a of this $(X + Y)$ brand are $a_{X+Y} = a_X + a_Y$ and $b_{X+Y} = \sum_{i=1}^{5} a_i - a_{X+Y}$. Hence, from the equation of marginal distribution $P(X = x)$, we have

$$
P(X + Y = 0) = \left(\frac{\alpha}{\alpha+1}\right)^r \frac{b_{X+Y}}{S} {}_2F_1\left(r, b_{X+Y} + 1, S + 1, \frac{1}{\alpha+1}\right).
$$

Let $s = \sum_{i=1}^{5} a_i$ for the market of five sites. From expressions 6.1 and 6.2, we derive duplication as follows:

$$
\begin{aligned}
&Duplication(X, Y)\\
&= \frac{1 - P(X = 0) - P(Y = 0) + P(X + Y = 0)}{1 - P(X = 0)}\\
&= 1 - \frac{P(Y = 0) - P(X + Y = 0)}{1 - P(X = 0)}\\
&= 1 - \left(\frac{\alpha}{\alpha+1}\right)^r ((b_Y/S) {}_2F_1(r, b_Y + 1, S + 1, 1/(\alpha+1))\\
&\quad - (b_{X+Y}/S) {}_2F_1(r, b_{X+Y} + 1, S + 1, 1/(\alpha+1)))\\
&\quad \cdot (1 - (\alpha/(\alpha+1))^r (b_X/S) {}_2F_1(r, b_X + 1, S + 1, 1/(\alpha+1)))^{-1}.
\end{aligned}
$$

B.7. The VBA Code for Customizing the ${}_2F_1$ Function in Excel

```
Public Function GHF(a As Double, b As Double, c As
Double, z As Double) As Double
Dim i As Integer
Dim j As Integer
Dim temp As Double
  GHF = 1
  temp = 1
  For j = 1 To 500
      temp = temp * (a + j - 1) * (b + j - 1) * z /
      ((c + j - 1) * j)
      GHF = GHF + temp
  Next j
  End Function
```

B.8. Algorithm LIND

We present algorithm LIND below to compute the market share and SoW for a focal firm j.

```
  Input: Firm j's customer purchase data, other K − 1
firms' penetration.
  Output: Market share and SoW estimation for all K
firms.
  1 LL=0 /*initial log-likelihood is set to 0 */
  2. Forall customers (i ∈ (1…N) {
  3 Compute BB/sNBD probability p(X_i=x_i) according to
Equation (A5)
  4 LL= LL + Ln(p)
  5 }
  6 Maximize likelihood according to the formulation
(7)
  7 Obtain the K+2 parameters as the solution to (7)
  8 Compute Market Share according to Equation (8)
  9 Compute SoW according to Equation (9)
  10 Output: MS and SoW
```

References

Bhattacharya, B. 1997. Is your brand loyalty too much, too little, or just right? Explaining deviations in loyalty from the Dirichlet norm. *Internat. J. Res. Marketing* **14**(5) 421–435.

Bickart, B., D. Schmittlein. 1999. The distribution of survey contact and participation in America: Constructing a survey-based estimate. *J. Marketing Res.* **36** (2) 286–294.

Boulding, W., R. Staelin, M. Ehret, W. Johnston. 2005. A customer relationship management roadmap: What is known, potential pitfalls, and where to go. *J. Marketing* **69**(10) 155–166.

Chen, Y., S. Yang. 2007. Estimating disaggregate models using aggregate data through augmentation of individual choice. *J. Marketing Res.* **44**(November) 613–621.

Danaher, P., I. Wilson, R. Davis. 2003. Comparison of online and offline consumer brand loyalty. *Marketing Sci.* **22**(4) 461–476.

Davis, J. 2007. *Measuring Marketing: 103 Key Metrics Every Marketer Needs*. John Wiley and Sons (Asia), Singapore.

Du, R., W. Kamakura, C. Mena. 2007. Size and share of customer wallet. *J. Marketing* **71**(4) 94–113.

Dunn, R., S. Reader, N. Wrigley. 1983. An investigation of the assumptions of the NBD model as applied to purchasing at individual stores. *J. Roy. Statist. Soc. Ser. C* (*Applied Statistics*) **32**(3) 249–259.

Ehrenberg, A. 1959. The pattern of consumer purchases. *Appl. Statist.* **8**(1) 26–41.

Ehrenberg, A. S. C. 2000. Repeat buying. *J. Empirical Generalisations Marketing Sci.* **5**(2) p1–p375.

Ehrenberg, A. S. C. 1988. *Repeat Buying: Theory and Applications*. Charles-Griffin, London.

Ehrenberg, A. S. C. 1995. Empirical generalizations, theory and methods. *Marketing Sci.* **14**(3) 20–28.

Ehrenberg, A., M. Uncles, G. Goodhardt. 2004. Understanding brand performance measures: Using Dirichlet benchmarks. *J. Bus. Res.* **57** 1307–1325.

Fader, P., B. Hardie. 2000. A note on modeling underreported poisson counts. *J. Appl. Statist.* **27**(8) 953–964.

Fader, P., J. Lattin. 1993. Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Sci.* **12**(2) 304–317.

Fader, P., L. Lodish. 1990. A cross-category analysis of category structure and promotional activity of grocery products. *J. Marketing* **54**(3) 52–65.

Fader, P., D. C. Schmittlein. 1993. Excess behavioral loyalty for high-share brands: Deviations from the Dirichlet model for repeat purchasing. *J. Marketing Res.* **30** 478–493.

Fader, P., B. Hardie, K. Jerath. 2007. Estimating CLV using aggregated data: The Tuscan lifestyles case revisited. *J. Interactive Marketing* **21**(3) 55–71.

Farris, P., N. Bendle, P. Pfeifer, D. Reibstein. 2006. *Marketing Metrics;* 50+ *Metrics Every Executive Should Master*. Wharton School Publishing, Philadelphia.

Fox, E., J. Thomas. 2006. Predicting retail customers' share-of-wallet using shopper loyalty card data. Working paper, Marketing Department, Southern Methodist University, Dallas.

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

Jeuland, A. B., F. M. Bass, P. Wright. 1980. A multi-brand stochastic model compounding heterogeneous Erland timig and multinomial choice processes. *Oper. Res.* **28** 255–277.

Johnson, L., S. Kotz, A. Kemp. 1992. *Univariate Discrete Distributions*, 2nd ed. John Wiley & Sons, New York.

Kahaner, L. 1998. Competitive intelligence: How to gather, analyze, and use information to move your business to the top. *Touchstone*. Touchstone, New York.

Kahn, B., M. Kalwani, D. Morrison. 1988. Niching versus change-of-pace brands: Using purchase frequencies and penetration rates to infer brand positioning. *J. Marketing Res.* **25**(11) 384–390.

Kallberg, J., G. Udell. 2003. The value of private sector credit information sharing: The U.S. case. *J. Banking Finance* **27**(3) 449–469.

Lawrence, R. J. 1980. The lognormal distribution of buying frequency rates. *J. Marketing Res.* **17** 212–220.

Liu, Q., K. Serfes. 2006. Customer information sharing among rival firms. *Eur. Econom. Rev.* **50** 1571–1600.

Menon S., S. Sarkar. 2007. Minimizing information loss and preserving privacy. *Management Sci.* **53**(1) 102–116.

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) 145–159.

Musalem, A., E. Bradlow, J. Raju. 2008. Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *J. Marketing Res.* **45**(December) 715–730.

Musalem, A., E. Bradlow, J. Raju. 2009. Bayesian estimation of random-coefficients choice models using aggregate data. *J. Appl. Econometrics* **24**(3) 490–516.

Padmanabhan, B., Z. Zheng, S. Kimbrough. 2001. Personalization from incomplete data: What you don't know can hurt. *Proc. 7th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (*KDD01*), San Francisco, 154–163.

Padmanabhan, B., Z. Zheng, S. Kimbrough. 2006. An empirical analysis of complete information for eCRM models. *MIS Quart.* **30**(2) 247–267.

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

Roberts, J., J. Lattin. 1997. Consideration: Review of research and prospects for future insights. *J. Marketing Res.* **34**(3) 406–410.

Schmittlein, D. C., A. Bemmaor, D. Morrison. 1985. Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Sci.* **4** 255–266.

Sharp, B. 2010. *How Brands Grow: What Marketers Don't Know*. Oxford University Press, Oxford, UK.

Sichel, H. S. 1982. Repeat buying and the generalized inverse Gaussian Poisson distribution. *Appl. Statist.* **31** 193–204.

Umyarov, A., A. Tuzhilin. 2011. Using external aggregate ratings for improving individual recommendations. *ACM Trans. Web* (*TWEB*) **5**(1) 22–26.

Uncles, M., A. Ehrenberg, K. Hammond 1995. Patterns of buyer behavior: Regularities, models, and extensions. *Marketing Sci.* **14**(3) 71–78.

Winkelmann, R. 2008. *Econometrics Analysis of Count Data*, 5th ed. Springer, Berlin.

Yang, Z., S. Zhing, R. Wright. 2005. Privacy-preserving classification of customer data without loss of accuracy. *Proc. 5th SIAM Conf. Data Mining* 92–102.

Zheng, Z., B. Padmanabhan. 2006. Selectively acquiring customer information: A new data acquisition problem and an active learning based solution. *Management Sci.* **52**(5) 697–712.