The Impact of Bestseller Rank on Demand: Evidence From a Software Market

Octavian Carare

Department of Economics
University of Maryland, College Park, MD.
E-mail: carare@econ.umd.edu

September 1, 2010

Abstract

This paper uses daily data on the sales ranking of the top 100 applications sold through Apple's *App Store* to provide evidence of the impact of today's bestseller rank information on tomorrow's demand. The estimates indicate that the willingness to pay of consumers is about $4.50 greater for a top ranked product than for the same product when it is unranked. The results also indicate that the effects of bestseller status on demand decline steeply with rank at the top ranks, but remain economically significant for apps in the first half of the bestseller list.

> When men are brought together they no longer decide at random and independently of one another; they influence one another. Multiplex causes come into action. They worry men, dragging them to right or left, but one thing there is they can not destroy, this is their Panurge flock-of-sheep habits. And this is an invariant.
>
> Henri Poincaré, *The Foundations of Science*, 1913

1 Introduction

Consider a consumer who has decided to purchase an entertainment good like a music album, a computer game, or a book. Which such good would one choose among the myriad of possible choices? Since these goods are often experience goods, the qualities of these goods are, to varying extents, uncertain to the consumer prior to consumption.

Consumers of experience goods may base quality predictions on their own experience. For example, many consumers purchase creative works of authors that they liked in the past. Some consumers may base their quality inferences on recommendations received from friends and relatives. Other consumers have great confidence in the recommendations of professional reviewers.

More recently, it has become commonplace for consumers to learn about new products by evaluating information concerning the past purchases of consumers they do not know personally. In today's marketplace, information concerning past purchases is increasingly often summarized and readily available in the form of bestseller lists. Bestseller lists are compiled for goods as diverse as books, music, DVDs, movies, computer games, electronic gadgets and household appliances. [1]

This paper aims to measure the effect on the consumers' willingness to pay of changes in the public information concerning the bestseller rank of products. The primary contribution of the paper is to provide empirical evidence that the past

[1] The first bestseller list of books was published in the U.S. by *The Bookman* in 1895 (see Bassett and Walter 2001). *Publishers' Weekly* compiled its first bestseller list in 1913. The influential *New York Times* bestseller list was first published in 1942. The first music bestseller list (then called a *hit parade*) was first published by *Billboard Magazine* in 1940. Currently, many retailers publish bestseller lists for several categories of goods. Amazon.com, for instance, displays on its website hourly-updated bestseller lists for about 30 product categories including books, electronics, automotive, home improvement, and grocery and gourmet food.

purchases of other consumers, as summarized by bestseller lists, affect consumer demands. This paper provides empirical evidence that the public sales rank information available in a large marketplace has a direct and causal effect on demand.

Recent evidence about the effect of information regarding the choices of other consumers on demand comes primarily from controlled experiments; see Salganik, Dodd and Watts (2006) and Cai, Chen and Fang (2009). These papers indicate that popularity begets popularity by showing that public information about the past popularity of products is a an important determinant of future popularity. A question that is not addressed by these studies concerns the differential impact of popularity rank information on demand. It is important to know from a policy perspective if the demand for the most popular item is different from from the demand for the other popular items. Intuitively this is so, but the question of how bestseller rank information affects future product demand appears to have eluded researchers. This paper provides a rank-by-rank measure of the impact on demand of information about the popularity of the other consumers' past choices.

The empirical application in this paper uses data from Apple's *App Store*, an on-line store that sells applications, or *apps*, developed for Apple's *iphone* and *ipod touch* devices. I collected daily data concerning the sales ranking of the top 100 free and top 100 paid apps between January 1 and June 16, 2009. I use these ranks to estimate the relationship between the consumers' willingness to pay and an app's position in the best selling list.

The novel method of estimation developed in this paper overcomes several important obstacles imposed by the data available for this study. The most important obstacle is the absence of information concerning unit sales from which the rankings are constructed. When the sales volumes of products can be compared over time, one can employ existing estimation methods designed for models with ranked dependent variables. Given the data available, product unit sales cannot be compared over time. In particular, a product with a sales rank of 30 today may have a higher or a lower volume of sales than a product with a sales rank of 40 tomorrow. The key insight that allows the formulation of an empirical model is that if sales follow a power law, log-sales are exponentially distributed. In the development of the estimation procedure I use a property of exponential order statistics according to which rank-weighted differences of consecutive order statistics from an exponential distribution are independently and identically exponentially distributed. Since the

expected value of these differences is invariant, estimates can be computed using simulation methods.

Prior to purchasing an app, consumers may become aware of the download ranks of the top apps during the previous day. Since download ranks change significantly over time, prior rankings can be used as a regressor to measure the effect of the public rank information on the consumers' willingness to pay in the present period. An important concern in the evaluation of estimation results is the possible endogeneity of previous sales ranks. While yesterday's ranks cannot be affected by the unobservables that affect app demands today, it is possible that the shocks that affect the demand for a product are correlated over time. If so, yesterday's ranks would be correlated with today's demand shocks and the least-squares estimates of the ranks' effect on demand need not reflect a causal relationship between rank information and willingness to pay.

The institutional details of the app market suggest good proxies for the past rank information that are uncorrelated with the demand shocks. The list of top 100 apps is displayed on several screens. The number of apps per screen varies from device to device, but the ranks 25 and 50 provide natural breakpoints for the visibility of apps. For instance, on both the iphone and the ipod touch devices the top five downloaded apps appear on the first screen, then the apps with ranks 6 to 25 can be browsed by scrolling down continuously. The ranks 26-50 are then available by following a link that appears on the first screen just below the app with rank 25. As instruments for the lagged ranks I use the movements of past sales ranks from a particular rank of ranks, as determined by these breakpoints. These movements are correlated with the past ranks, but uncorrelated with the current demand shocks.

The estimation results indicate that the position of an app on the list of previous-period bestsellers is an important determinant of present-period demands. The results show that, all else equal, the willingness to pay of consumers is decreasing steeply with the sales rank of an app. For instance, the value attributable to the previous-period rank of 1 is roughly twice as large as the corresponding value for rank 2.[2] In turn, the value attributable to rank 2 is about 30% larger than the corresponding value for an app with rank 3.

Online retailers commonly compile and prominently display bestseller lists. Given the cost of doing so, these retailers must gain some economic advantage. Re-

[2] As for most bestseller lists, increasing ranks correspond to less popular products.

cently, retailers have started offering prices that are dependent on the market share ranks of items.[3] The advent of such pricing rules may prove to be a significant development that is brought about by the gradual shift from traditional, brick-and-mortar retailing to online retailing. The main finding of this paper is that publicly available information concerning download ranks influences the willingness to pay of consumers in a significant way. This finding may at least in part explain these developments.

The remainder of the paper proceeds as follows. Section 2 provides a brief overview of the related literature. The market and data are presented in Section 3. A model is developed in Section 4, and estimation results are presented in Section 5. Conclusions are in Section 6.

2 Related Literature

In his criticism of the state of demand theory, Morgenstern (1948) was among the first authors in the modern economics literature to state that certain non-market interactions between consumers may give rise to what he called "non-additivity" of demand curves.[4] An example of non-additivity are "fashions, where one person buys because another is buying the same thing." By incorporating various aspects of external consumption into the theory of consumer demand, Liebenstein (1950) further developed Morgenstern's observation. Liebenstein's theoretical model is static, but he suggested that a dynamic model would better capture some aspects of the problem. The papers on observational learning by Banerjee (1992), Welch (1992) and Bikhchandani, Hirschleifer and Welch (1992) develop such dynamic models.[5] These models show that inefficient outcomes may arise because information is updated sequentially on the basis of the actions of other consumers, and not on the basis of the information available to other consumers.

When choosing a product a consumer may follow other consumers because either she believes that the other consumers have superior information, or because

[3] For instance, *Amie Street,* an independent on-line music retailer, offers for download songs that are priced between 0 and 98 cents depending on their popularity.

[4] Before Morgenstern, Pigou [1913] observed that individual demand curves may be interrelated. Mason [1995] provides a comprehensive historical account of the way in which interpersonal effects have been dealt with in consumer demand theory.

[5] The website www.info-cascades.info maintained by Ivo Welch provides an excellent annotated guide to the vast literature spawned by these models.

the first consumer prefers to act like the others. The former is called expectation interaction and the latter is called preference interaction. Manski (2000) emphasized that it is important for analysis and for policy to distinguish between expectation interactions and preference interactions. In our market, when choosing a product, a consumer may follow other consumers because she believes that the other consumers have better information, or because she has a preference for imitation.

Consumers endowed with bestseller information may also follow other consumers because they believe that any idiosyncratic noise that affects their product quality expectations does not affect the aggregate consumption decisions. In reality, we know very little about the reasons why consumers follow other consumers. It is likely that in the app market expectation interactions are at work in shaping the purchase decisions of the consumers. The anonimity of bestseller lists may dampen the role of preference interactions in markets. In any case, the present analysis maintains an agnostic view about the particular mechanism through which the aggregate consumption decisions of the crowd affect consumers' preferences.

This paper is closely related to two papers that investigate the impact on consumers of learning about the past consumption choices of other consumers. The first paper by Cai, Chen and Fang (2009) reports results of a randomized field experiment conducted in a restaurant setting. Their experiment is designed to distinguish between the informational effect of expanding the consumers' knowledge of the choice set (called the *saliency effect*) and the informational effect of observational learning. The authors provide evidence that the demands for the most popular five dishes increase significantly as a result of revealing information about the ranking of the top five dishes. They find little support for the hypothesis that sales are driven by a saliency effect. The results of their controlled field experiments indicate that most of the increase in the sales of the top dishes occur as a result of observational learning. The authors suggest that a partial explanation for the common practice of displaying popularity information on retailers' websites is observational learning. The present paper brings a new contribution to this valuable insight by providing direct market evidence that popularity information affects demands. Cai, Chen and Fang (2009) consider "lumpy" informational effects in the sense that the effect of revealing information about the most popular dish is not distinguished from the effect of revealing information about the subsequent four popular dishes. I complement their findings by providing estimates of how demands are affected by each of

the publicly available popularity ranks.

A second paper, by Salganik, Dodd and Watts (2006) suggests that the inability of experts to predict with accuracy which products will succeed in the marketplace is due in part to social influences that increase the unpredictability of success. To study the effect of social influences, the authors created a large artificial music market and instructed the experiment participants to listen to songs, rank them, and also offered participants the option of downloading (for free) the songs they liked. The authors' experimental treatments expose the participants to different information structures. In particular, in one treatment the participants received information about the number of times a song was downloaded previously. The authors' experimental data support the hypothesis that the information concerning the choices of others shapes the participants' preferences. Importantly, the authors also argue that the effect of social influences in real markets may be stronger than in their experiment. I seek to complement their results by measuring the strength of these influences in a real market and by quantifying these influences on a rank-by-rank basis.[6]

The analysis in this paper relies on the assumption that daily product sales are Pareto distributed. This assumption is consistent with the limited sales volume data available. The assumption is also in the tradition of empirical analyses of the distribution of firm size pioneered by Simon and Bonini (1958). A recent application of the properties of the Pareto distribution is in the analysis of Chevalier and Goolsbee (2003). The authors measure price competition in the online book market by proposing a method for turning sales ranks into sales volumes. They assume that sales follow a power law and devise simple and inexpensive experiments that permit the estimation of the power law parameters. Brynjolfsson, Hu and Smith (2003) also assume that sales follow a power law, and compute the parameters of the power law using a property of the competitive market equilibrium. While I

[6] A few other papers provide evidence that peer effects have an important role in shaping individual economic decisions. Conley and Udry (2009) study the role of social learning in the diffusion of a new agricultural technology. Duflo and Saez (2002) investigate the role of peer effects in retirement savings decisions. Foster and Rosenzweig (1995) use household panel data to test the implications of a model that incorporates learning-by-doing and informational spillovers. Moretti (2009) uses a rich data set on movies to quantify the effect on consumption of information received from peers. Munshi (2004) provides a test of social learning in a heterogeneous population of farmers in India. Zhang (2009) discusses observational learning in the kidney transplant market. Tucker and Zhang (2009) analyze the effect of popularity on consumers' choice using data from a field experiment involving wedding vendors.

retain the assumption that sales follow a power law, my estimation procedure is different from the procedures developed by Chevalier and Goolsbee and Brynjolfsson, Hu and Smith, and permits estimation of the demand parameters of interest without knowledge of the parameters of the power law.

People have a perhaps innate tendency to join the crowd. As the experiments of Migram, Bickman and Berkowitz (1969) suggest, this may happen even when the people in the crowd do little else than stare at the empty sky. The tendency to join the crowd could imply that individuals perceive a distinct economic advantage from following the actions of others. If so, one may expect to see that popularity begets popularity in markets where popularity is public. The experimental findings of Cai, Chen and Fang (2009) and Salganik, Dodd and Watts (2006) show that popularity begets popularity in two important markets. Popularity also begets popularity in the news media, as shown by Thorson (2008). Her analysis indicates that most-emailed lists displayed in the on-line editions of newspapers like the *New York Times* and *Los Angeles Times,* as well as the larger family of news recommendation engines, affect consumers behavior by providing consumers with new ways to navigate information. Citations of academic research seem to follow a similar popularity pattern.[7]

There is very little written in the economics literature on the subject of bestseller lists. Sorensen (2007) analyzes the impact of the *New York Times* bestseller list on sales and on product variety. He finds that the listing of a book on the bestseller list causes a modest increase in sales. The objective of this research, like one of Sorensen's (2007) research objectives, is to measure the effect of bestseller information on demand. The availability of prices is the most tangible advantage of the data available for this study. Because of limited information on sales prices, Sorensen was able to only focus on how bestseller status affects the autocorrelation of sales. Since the app market data contain detailed price information, this study investigates how both prices and bestseller status affect product demands.

I turn next to a short presentation of the institutional details of the software market and of the data.

[7]See Merton's (1968) discussion of the "Matthew effect" in science. For an opposing view, see Simkin and Roychowdhury (2005), who argue that citations generated according to a process by which a scientist picks three papers at random, cites them, and also randomly copies a quarter of these papers' references fits the empirical citation distribution quite well.

3 The App Market and Data

The App Store is a platform for apps that run on the Apple's *iphone* and *ipod touch* devices.[8] As of the June 2010, the store, created by Apple Inc. in July 2008, made available more than 225,000 such apps for download. According to Apple, there have been more than 5 billion app downloads from the store between July 2008 and May 2010. The apps are classified in 20 categories that include books, business, education, games and social networking. During the data collection period the category with most apps was games. As of July 2010, according to mobile advertiser mobclix.com, the numbers of books and game titles available for download were tied at about 41,000.[9]

Very few apps are developed by Apple itself. Thus, an overwhelming majority of the apps are sold by third-party developers. For third-party apps, Apple acts as an intermediary. In particular, it facilitates sales and retains a fraction of revenue. At the beginning of September 2009, as reported by Apple, there were 125,000 registered app developers. The potential market for these apps comprise more than 50 million users of iphone and ipod touch devices.[10] As the very large number of developers indicates, the licensing and capital requirements for becoming an app developer are modest. The typical developer is an independent programmer. However, a few large gaming software firms like Ubisoft and Electronic Arts released several apps. The high number of developers, the low licensing costs, and the high number of apps indicate that the market for apps is very competitive.

There are two classes of apps: free and paid. Free apps are either versions of paid apps with reduced functionality, or apps that display advertisements. Some developers prefer to release their apps free of charge and to retain all the advertising proceeds. If a developer releases a paid app, the developer releasing a paid app retains only 70% of the price.[11]

Apps can be downloaded wirelessly from the iphone or the ipod touch devices

[8] These apps also run on the *ipad*, released by Apple in April 2010, but the ipad released after the data collection period.

[9] See *http://www.mobclix.com/appstore/1*, accessed July 8, 2010.

[10] See *http://www.apple.com/pr/library/2009/09/28appstore.html* and *http://news.cnet.com/8301-13506_3-10362544-17.html*; accessed July 2010. Clearly, since there were only 85,000 apps at the time, some developers have had yet to release apps.

[11] There were no advertisements in paid apps until the recent release in August 2009 of an app by CNN that displays advertisements. The release was followed by a wave of acrimony on the part of consumers who bought the app.

through a mobile app store interface, or through a wired connection to a personal computer. The interface for the wired download is a multi-platform program called *itunes* that is developed and offered as a free download by Apple. Downloads require a user ID and a password. Consumers are uniquely identified by their user ID. In order to register with Apple, a consumer must provide a valid credit card number and the card's billing address.

The very large number of apps in the store makes it impractical for consumers to sample of the entire set of apps. Apple is well aware of this fact and facilitates the process of product discovery by consumers. It provides prominently displayed links to the most downloaded apps both on the mobile app interface and on the itunes program. Importantly, the mobile app store interface displays on its main page links to the top 25 most downloaded free and paid products, and ranks 26-50 are available by following a subsequent link. See Figure 3 in the Appendix. The wired itunes interface also prominently displays the top 10 free and paid apps on its first page. Lists of the top 100 most downloaded free and paid apps are available in the itunes interface one click away from the main store page. In addition to these lists, Apple facilitates product discovery by consumers through lists of featured products and, on the itunes interface, staff favorite products. Also, Apple provides lists of most downloaded apps by category and allows consumers to search products by name.[12]

Developers and industry experts perceive the top 100 free and paid lists as the most important ways for consumers to learn about apps. Importantly, since apps are sold in many countries, the rankings are computed separately for each of the more than 60 country stores. The data employed in the present paper concern the download ranks of apps sold in the United States store, which is the largest store in terms of number of apps sold.

Aside from releasing quarterly aggregate download figures, Apple does not release information about the number of apps app downloaded from its store. Importantly, Apple has not released any information about the details of computing the top 100 most downloaded apps.

The information available concerning the computation of download ranks is

[12] In September 2009, with the release of itunes 9.0, Apple has added a list of the top 100 grossing apps to the previously published lists of top 100 free and top 100 paid apps by the number of downloads.

indirect and comes from mobile advertisers who are able to track app downloads and usage. Mobile advertisers agree that Apple's rankings are based on unit sales and that Apple uses a 24-hour window to compute download ranks. Unit sales are the only criterion used for computing the top 100 lists. Product characteristics like the number and quality of reviews, user ratings, and prices, do not affect the way the top 100 lists are compiled. Re-downloads of an app by the same user do not count. As such, downloads of updated versions of previously downloaded apps do not count in the calculation of the download rankings. Given the institutional features of the app market, it is not practical for a developer to affect the download rank of an app by repeatedly downloading it.

I collected daily data on the most downloaded 100 paid and 100 free apps for a period of 166 days, between January 1, 2009 and June 16, 2009. The data were collected using a program that accessed daily the sales rank information made available by Apple.

On two occasions my data collection program malfunctioned because of a power outage and a network maintenance issue. On the first occasion the ranking data were not collected for one day. On the second occasion the data were collected, but the collection program failed to produce the desired output for six consecutive days.[13] On the first occasion almost all of the rank data were recovered from the information available, but on the second occasion the missing information could only be recovered for two days. The missing data only impacts the result through a reduced number of observations. The fact that there are missing rankings that correspond to a few consecutive days does not cause any other issues with estimation.

The data collected from Apple include download ranks, the names and the developers of the top 100 paid and free apps, previous day download ranks, the date when the current version of the app was released, the app category and, for paid apps, their price. In addition to this information, I added information concerning the date when the apps in the data were first released. This date is generally different from the version release date that is available in the initial ranks data. The dates of first release for the apps in the data were collected from two app tracking sources, *apptism.com* and *appshopper.com*. I also collected from Apple information about the size of the app files.

The raw data contain ranking information about 912 apps, 509 of which are

[13] The two malfunctions were observed at the end of the collection period.

paid apps. Since the empirical part of this paper uses only data on paid apps, I now discuss some of the characteristics of the paid app data. The average price of a paid app during the observation period was $2.72, with a standard deviation of $2.25. The minimum price of an app was $0.99, and the maximum price was a sizable $29.99. The median price of an app during the observation period was $1.99. On average, a paid app appears in the top 100 most downloaded list on 31 days. However, the apps that have at least once reached the top 50 appear on the top 100 list for an average of 45 days, and the apps that have reached the top 10 appear on the top 100 list for an average of 63 days. The distribution of the number of days on the 100 most downloaded list is highly skewed, as indicated by the median number of days on the list equal to 18. This skewness indicates that a few apps survive on the list for a long time. Some 15 apps remained on the top 100 list for the entire observation period. However, most apps remain on the top 100 list for three weeks or less. Some apps appear on the top 100 list, then leave the list and return to the list after some time. This process may repeat a few times for a few apps in the data. Figure 1 displays a histogram of the number of days on the top 100 most downloaded list app.[14]

[14] Clearly, the number of days on the list for apps that entered the list at the beginning and toward the end of the observation period is not accurate. However, the same highly skewed distribution of number of days on the list obtains when I drop the apps that entered the list during the first and last 20 days of the observation period. Notable is that 15 apps have survived on the list during the entire observation period.

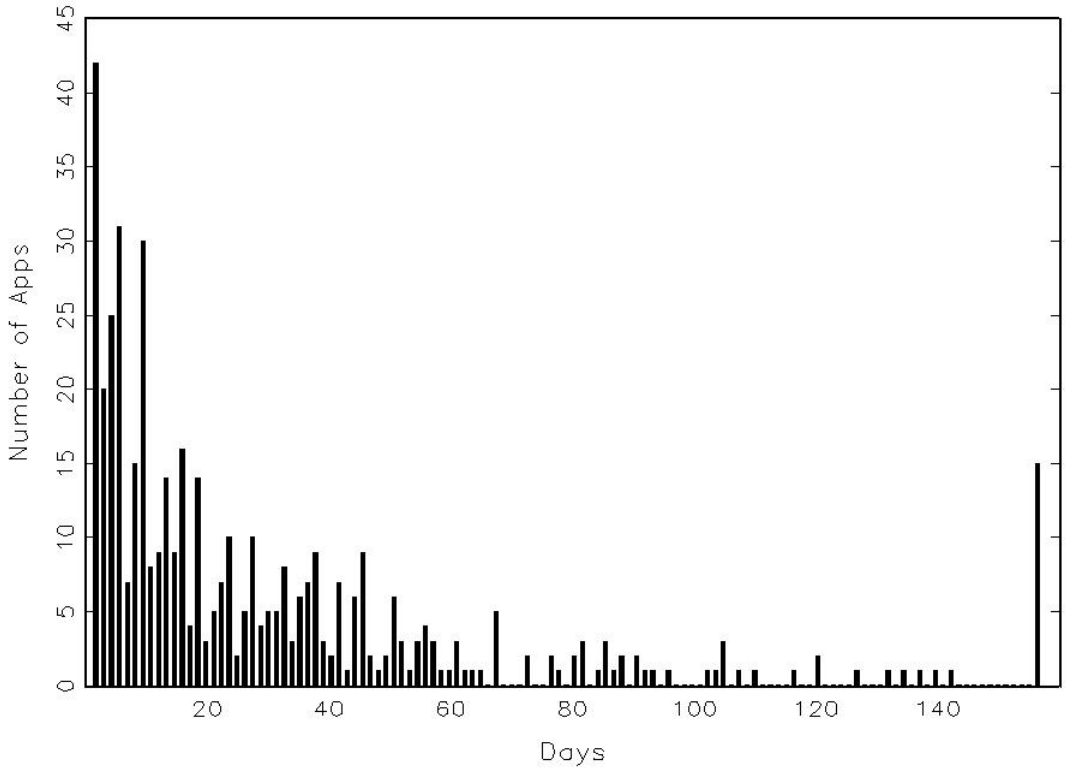

Figure 1: Histogram – Number of Days on the Top 100 List

The data collected on paid apps contain 16,387 unique rank-day observations. Some apps were removed by Apple or by their developers during the data collection period. As such, some of their characteristics were not available in the data provided by Apple and these observations were dropped from the final sample. Since lags up to order three of app ranks were used in estimation, some more observations had to be deleted. The final sample contains a number of 13,996 observations.

Summary statistics of the data are provided in Table 1 in the Appendix.

4 The Model

I consider a discrete choice setting where consumers choose at most one product among the N_t products available at time t.[15] Prior to making their consump-

[15] With daily data, this is a reasonable approximation. A recent survey of 1,117 App Store users by AdMob, the largest mobile advertiser, shows that about half of the consumers buy less than an app per month. On average, the consumers in the survey have purchased 5 apps on which they spent

tion choices, consumers may become aware of the past sales ranks of the top 100 100 products by market share. Sales ranks are compute using unit sales. Denote by $r(j,t)$ the sales rank of product j at time t. It is also useful to denote by $R(\cdot,t) : \{1,...,100\} \to \{1,...,N_t\}$ the mapping from sales ranks into products at time t. Denote by s_{jt} the market share of product j at time t, defined as:

$$s_{jt} = \frac{S_{jt}}{\sum_{i \leq N_t} S_{it}},$$

where S_{jt} denotes the number of units of product j sold at time t. It is useful to notice that the ranking of sales and market shares are identical. For any two products k and l, a lower sales rank implies a higher market share, so at any time t, if $r(k,t) < r(l,t)$, then $s_{kt} \geq s_{lt}$. In the data, the product with sales rank 1 is the product with the highest market share and the product with sales rank 100 is the product with the lowest market share.

At each time t the sales ranks of the top 100 products, but not the actual market shares of these products, are known. The main goal of this section is to construct an estimable equation that relates market shares to sales ranks. Throughout, I maintain the following three assumptions.

Assumption 1 *The sales S_{jt} at time $t \geq 1$ of products $j \in \{1,...,100\}$ are distributed according to a Pareto distribution with parameters $S_t^m > 0$ and $\theta > 0$. The probability that the sales S of a product are greater than some value $s \geq S^m$ is given by $Pr[S > s] = \left(\frac{S_t^m}{s}\right)^{\theta}$.*

The parameter S^m of the Pareto distribution is the lower boundary of its support. The scale parameter θ is often referred to as the Pareto index. Note that the Pareto distribution is a power law.

Assumption 2 *The sales S_{100t} at time $t \geq 1$ of the products with sales rank 100 are strictly greater than zero.*

Assumption 3 *Before choosing a product today, consumers may observe yesterday's sales ranks of the top 100 products. Conditional on the market share rank*

about $9. See *http://arstechnica.com/apple/news/2009/08/app-store-buyers-spend-an-average-of-9-per-five-downloads.ars*.

$r(j,t-1)$ of product j at time $t-1$, the mean utility provided to consumers is given by

$$\delta_{jt} = X_j\beta - \alpha\left(P_{jt} - M_{r(j,t-1)}\right), \tag{1}$$

where X_j are characteristics of product j, P_{jt} is the price of product j at time t, and M_i ($i = 1,...,100$) are the rank-dependent parameters of principal interest in estimation. Consumer i derives random utility u_{ijt} from choosing product j at time t, given by:

$$u_{ijt} = \delta_{jt} + \varepsilon_j + \varepsilon_{jt} + \xi_{ijt}, \tag{2}$$

where ε_{jt} is an i.i.d., across products and across time, disturbance representing j's unobserved component of utility at time t, ε_j represents a product fixed effect, and the shocks ξ_{ijt} are distributed independently across consumers, products and time according to the Type-I extreme value distribution.[16]

Assumption 1 is consistent with evidence from app developers and industry analysts that product sales in the upper tail of the sales distribution are governed by a power law. For the developers that made public their app ranks and sales, a power law appears to fit their data quite well.[17] Importantly, I assume a power law distribution of unit sales only in the upper tail of the distribution of sales, not for the entire distribution.

Assumption 2 is most likely innocuous in large markets. It makes possible the calculation of log market shares and ensures that the support of sales in the upper tail is bounded away from zero.

Assumption 3 is standard in the empirical literature on discrete consumer choice (see Berry 1994). The difference between the mean-utility specification in (1) and the utility specification in the standard discrete choice model is the presence of the parameters M. These parameters represent changes in willingness to pay that are due to ranking information. Positive values of the parameters M represent increases in the willingness to pay of consumers, while negative values represent decreases in the willingness to pay.

The following lemma provides a relationship between the the sales rankings of products and differences in market shares under the assumption that sales are Pareto

[16] While the theoretical model assumes that the disturbance ε_{jt} is i.i.d., in Section 5 I discuss the possible implications of autocorrelation and provide some robustness checks.

[17] One such account is at *www.joelcomm.com/app_store_ranktosales_revealed.html*.

distributed.

Lemma 1 *Suppose Assumptions 1 and 2 hold. Denote by* $s_{1t} \geq s_{2t} ... \geq s_{100t}$ *the market shares (by number of units sold) of the top 100 products by market share at time* t*. For* $k \in \{1, ..., 99\}$*, the random variables* $y_{kt} = k\theta \left(\ln s_{kt} - \ln s_{k+1t} \right)$ *are independently distributed according to an exponential distribution with unit mean.*

The proof of Lemma 1 is in the Appendix. Notice that since I allow one of the parameters of the Pareto distribution to vary over time, product sales need not be identically distributed over time.

Given Assumption 3, the following lemma aids considerably with the formulation of a feasible estimation procedure.

Lemma 2 *(Berry inversion, 1994). Given the utility specification in (2), market shares of products* $k = 1, ..., N_t$ *can be expressed as*

$$\ln s_{kt} = \delta_{kt} + \ln s_{0t} + \varepsilon_k + \varepsilon_{kt} \,, \tag{3}$$

where s_{0t} *is the market share of the outside good at time* t*.*

Proof. See Berry (1994). ■

It is well known (see e.g., Berry 1994) that the utility formulation in (2) gives rise to *a priori* unreasonable substitution patterns among products. Other models, like the nested logit (McFadden 1978; Cardell 1991) and the random coefficients logit (Nevo 2000) avoid this issue, but impose data requirements that are considerably more demanding than the rank data available for this analysis.

If data on market shares were available, the equation of interest for estimation would be

$$\ln s_{kt} - \ln s_{0t} = X_k \beta - \alpha \left(P_{kt} - M_{r(k,t-1)} \right) + \varepsilon_k + \varepsilon_{kt} \,, \tag{4}$$

where the dependent variable is the difference between the the market share s_{kt} of product k at time t and the market share s_{0t} of the outside option at time t. Aside from the parameters M, this equation is the standard equation employed in the estimation of models of discrete consumer choice. This equation permits the estimation of demand elasticities and of the effect of the observable product characteristics X on consumer utility. The parameters M have a straightforward interpretation. These

parameters express the differences in consumer utility that result from the public information concerning the market share ranks of a product at time $t-1$. To the extent that the utility of consumers who choose a product today is affected by the public information concerning the market share ranks of products yesterday, one expects the estimates of rank-dependent parameters M to be significantly different from zero. Estimating equation (4) is complicated by the possible endogeneity of price and lagged ranks. This possible endogeneity is discussed in the next section. The fixed product effects help to capture any unobservable components of product quality that are time-invariant and may be correlated with price or with previous ranks.

The available data do not allow the direct estimation of (4). Note, however, that the market share of the product with download rank $k+1$ at time t can be expressed as

$$\ln s_{k+1,t} - \ln s_{0t} = X_{k+1}\beta - \alpha\left(P_{k+1t} - M_{r(k+1,t-1)}\right) + \varepsilon_{k+1} + \varepsilon_{k+1t}. \qquad (5)$$

Subtracting (5) from (4), we have

$$\ln s_{kt} - \ln s_{k+1,t} = \hat{X}_{kt}\beta - \alpha\left(\hat{P}_{kt} - M_{r(k,t-1)} + M_{r(k+1,t-1)}\right) + \hat{\varepsilon}_k + (\varepsilon_{kt} - \varepsilon_{k+1t}), \quad (6)$$

where $\hat{X}_{kt} = X_{kt} - X_{k+1,t}$, $\hat{P}_{kt} = P_{kt} - P_{k+1,t}$ and $\hat{\varepsilon}_k$ is the difference between the fixed effects that correspond to the products with download ranks k and $k+1$. Notice that (6) does not contain the share of the outside good. However, the differences in market shares on the left hand side are not observed. Lemma 1 shows that when sales are Pareto distributed, the left hand side of (6) is exponentially distributed. If rescaled appropriately, the parameter of the distribution that governs the left hand side differences is not dependent on the parameters to be estimated. Accordingly, these unobserved market shares can be simulated and the equation of interest can be estimated as if market shares were known. Replicating the process many times with different draws of the left hand side variable permits estimation of the coefficients M. A relationship between the rank-specific values M and the sales ranks of the top products can be formulated based on the following theorem.

Theorem 1 *Given assumptions 1-3, for $k \in 1,...,99$ and for $t \geq 1$, the following hold:*

i. $y_{kt} = k\theta\left(\ln s_{kt} - \ln s_{k+1t}\right)$ *are i.i.d. exponentially distributed with mean*

1; and

ii.

$$\frac{y_{kt}}{k} = \hat{X}_{R(k,t)}\beta\theta - \alpha\theta\left[\hat{P}_{R(k,t)t} - \hat{M}_{r(R(k,t),t-1)}\right] + \theta\hat{\varepsilon}_{R(k,t)} + \hat{\varepsilon}_{kt} \quad (7)$$

where the hat variables are defined as follows:

$$\hat{X}_{R(k,t)} = X_{R(k,t)} - X_{R(k+1,t)} \quad (8)$$

$$\hat{P}_{R(k,t)t} = P_{R(k,t)t} - P_{R(k+1,t)t} \quad (9)$$

$$\hat{M}_{r(R(k,t),t-1)} = M_{r(R(k,t),t-1)} - M_{r(R(k+1,t),t-1)} \quad (10)$$

$$\hat{\varepsilon}_{k,t} = \varepsilon_{R(k,t)} - \varepsilon_{R(k+1,t)} \quad (11)$$

and where the term $\hat{\varepsilon}_{kt}$ is a zero-mean error.

Proof. Part *i* follows directly from Lemma 1. To prove *ii*, use Lemma 2 to express the differences in consecutive order statistics of market shares as:

$$\ln s_{kt} - \ln s_{k+1t} = \delta_{R(k,t)t} + \ln s_{0t} + \varepsilon_{R(k,t)} + \varepsilon_{R(k,t)t} - \delta_{R(k+1,t)t} - \ln s_{0t} - \varepsilon_{R(k,t)} + \varepsilon_{R(k,t)t}.$$

Multiplying by $k\theta$ and collecting terms yields (7). ■

Theorem 1 constitutes a critical step in the development of an estimation procedure. Since the distribution of the rescaled differences $k\theta(\ln s_{kt} - \ln s_{k+1t})$ is independent of the parameters to be estimated, these unobserved differences can be replaced by simulated, statistically independent unit-mean exponential random variables.

Note that the coefficient α on price and the βs are only identified up to a positive multiple – the parameter of the power law. Also note that the the estimating equation does not permit identification of coefficients on variables like daily dummies that do not change over the course of a day. Rather then a shortcoming of the model, this is one of its strengths. It implies that any daily effects like those due to operating system updates and advertising campaigns are differenced away. As such, variables of this kind do not affect the estimation results. A final note concerns the unidentifiable parameters of the assumed power law distribution. I assume that the power law exponent θ is the same over time, but the assumed upper tail distribution

of sales need not be the same. The reason is that the location parameter S_t^m of the power law is allowed to vary over time. The variation of this parameter does not affect the results because in the estimation procedure this parameter is differenced away.

I turn next to a discussion of the estimation procedure and of the results.

5 Estimation and Results

5.1 Estimation Procedure

This section provides details about the estimation procedure and discusses the potential biases that may arise due to the endogeneity of some variables.

In the data, an observation is a date-rank tuple (t,k,j) that corresponds to app j. There are $13,996$ date-rank-app tuples in the data. Apps are uniquely identified in the data by their ID code.

Denoted by $M = (M_1, M_2, ..., M_{100})$ is the vector of rank-specific valuations to be estimated. While the app fixed effects help to capture any unobserved components of quality, a possible concern is that prices are endogenous in the following dynamic sense. There are more than 400 app price changes in the data. Many of these price changes appear to be caused by changes in download ranks. If, for instance, developers who see declining download ranks for their apps tend to reduce the price of their apps, as the folklore on the online developer forums suggests, then price would be endogenous in (7) and the estimated price coefficient would be biased. I instrument the price variable using the logarithm of lagged sales ranks. Since changes in download ranks do not immediately trigger price changes, I use lags of order two and three of the sales ranks as price instruments. Both price instruments are strongly significant in the first stage.

Unit mean pseudo-random exponential variables were generated using a multiplicative congruential random number generator (see Kennedy and Gentle 1980, pp. 136-147). These variables are the $\hat{Y}$s on the left-hand-side of the estimating equation:

$$\hat{Y} / K = \hat{X} \times (\alpha^* M, \alpha^*, \beta^*, \varepsilon^*) + \hat{\varepsilon}, \tag{12}$$

where K denotes the vector of current app ranks and an asterisk denotes parameters that are multiples (by θ) of the parameters of interest. The data matrix $\hat{X}$ contains

previous-day rank dummies, the app price, as well as app specific information that including age, version age and size, and app fixed effects.

I turn next to a discussion of the empirical findings.

5.2 Estimation Results

Figure 2 contains a plot of the estimates M against the corresponding ranks (values are expressed in dollars). The mean estimates are depicted with a red continuous. Mean estimates are bracketed by lower and upper 95% confidence intervals drawn using dashes. Means and standard deviations are computed using the results of 10,000 simulation runs. These estimates are not required to satisfy any particular functional form. The estimated rank-specific value that corresponds to the first ranked apps is substantially greater than the average price ($1.75) of the products that occupy the top rank.

The estimates suggest that there are two important changes in the way previous sales ranks affect demands at ranks around 25 and 50. These ranks correspond to the highest ranks displayed by the mobile app store interface on its two rank pages (see Figure 3 in the Appendix). This finding suggests that the demand for apps that have just reached the top 25 and top 50 list exhibits a discrete increase.

The estimated relationship between ranks and values depicted in Figure 2 is not monotone. Mean estimates decrease from rank 1 to rank 15, but become somewhat noisy at higher ranks. The estimation results indicate that, all other things equal, consumers attribute a slightly lower value to apps previously ranked toward the bottom of the top 100 list than to apps that are previously unranked. Because unranked apps are typically newer apps, this finding indicates that consumers believe that the apps at the bottom of the top 100 list are worse that apps that have not had a chance to reach the top 100 list.

More detailed results of estimation are contained in Tables 2 and 3 in the Appendix. Table 2 contains estimates of the starred parameters that correspond to the estimating equation (12). Table 3 reports means and standard errors of the rank coefficients M, determined as the ratio between the estimates of M^* and the estimate of the price coefficient α^*.

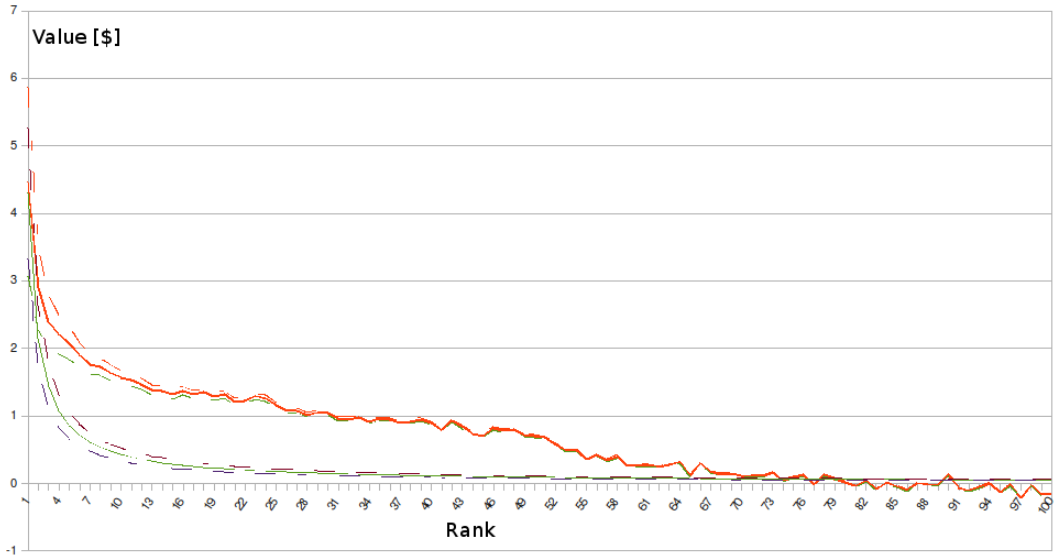

Figure 2: Rank-Specific Values

5.3 Auto-correlated Demand Shocks

An important concern for estimation is the possible endogeneity of the previous-day rank dummies in the estimating equation 12. While yesterday's download ranks are not affected by the shocks that affect app demands today, demand shocks for a product may be serially correlated. This would cause yesterday's ranks to be correlated with today's demand shocks. In effect, the measured effect of ranks on demand could reflect this autocorrelation and not a direct effect of rank information on demand.

To see whether endogeneity could affect the results of estimation, the effect of yesterday's ranks may be estimated using the method of instrumental variables. However, given the relative dearth of variables in the data, obtaining credible instruments for all the past rank variables is impractical.[18] Instead, a specification is used that restricts the coefficients M to be proportional with the reciprocal of the previous sales rank.

The reciprocal of previous ranks may be instrumented using movements of

[18] More granular movements to and from particular rank ranges (e.g., to and from ranks 1-5, 6-10, etc.) could provide good instruments for the previous-day rank variables. It turns out that while many of these instruments are significant, the goodness of fit for most of the first-stage regressions is poor (most first-stage F statistics are lower than 4). Not surprisingly, given the weak instruments, the second-stage estimates of the effect of public rank information on demand were very noisy.

second-lagged ranks to and from particular rank ranges that are natural breakpoints in the way the top list is displayed. Clearly, movements of past download ranks from one rank range to another are correlated with current ranks. Because the apps in the bestseller list are displayed sequentially, apps displayed at the top of the best-seller list may receive more attention from consumers than apps at the bottom of the list. An app that has just made the top 50 (or 25) is more likely to maintain its top ranking than another app that has not just crossed such a threshold. In effect, the increased visibility of an app may affect its subsequent sales rank. Importantly, while movements of ranks two days ago across various thresholds are correlated with yesterday's ranks, these movements are unlikely to be correlated with today's demand shocks. In principle, these rank movements are good instruments for yesterday's ranks.

The estimated parameters M are close to the estimate of the impact of previous sales rank on demand obtained by restricting the parameters M to be proportional with the reciprocal of the sales rank. The estimated coefficient on the reciprocal of sales rank is equal to 4.29, with a standard error of 0.48. The estimated impact of sales ranks on demand computed using this estimate, bracketed by a 95% confidence interval, is depicted in Figure 2 with the thinner continuous green line.

To address the potential endogeneity of prices and previous sales ranks, these variables were instrumented using the lags of order two and three of sales ranks, as well as two binary variables that measure whether the twice-lagged sales rank of an app has reached the list of top 50 or the top 25 apps from below. All instruments are strongly significant in the first stage regressions. First stage results are summarized in Table 6 in the Appendix. As indicated by the results of over-identification tests, the null hypothesis that the instruments and the error term in (12) are orthogonal was not rejected.[19]

The instrumental variables estimate of the reciprocal of past rank coefficient was equal to 4.37, with a standard error of 0.49; notice that this is only slightly higher than the ordinary least squares coefficient of 4.29.[20] This finding is a strong indication that serial correlation is unlikely to be a cause of concern when the parameters of interest M are estimated by ordinary least squares.

[19] The maximum value of the Sargan statistic was 1.6, lower than the 1% $\chi^2(2)$ critical value of 9.21. The mean value of the Sargan statistic for the 10,000 simulation runs was 0.73.

[20] The IV coefficients and their standard errors are given in Table 5 in the Appendix.

5.4 Discussion of Results

The most striking result of this analysis is that the estimated rank-dependent values M for the top two ranks are quite large relative to the price of the apps that occupy these ranks. The results also indicate that the value attributable to the highest rank is almost double the value attributable to the second highest rank. Another striking result is the apparent precipitous drop in the estimated parameters M in the neighborhood of ranks 25 and 50. These ranks are the last ranks displayed on the two pages of the top app list on the iphone and ipod touch devices. The results also show that the rank-dependent values decline steeply with rank for the top 10 products. The effect of product rank on demand for apps ranked 50 or higher become very small.

One possible concern is that the estimation results are affected of unobserved quality differences between products. These unobserved quality differences are unlikely to affect the estimation results because they are absorbed by the app fixed effects.[21] Another possible concern is that the findings are not indicative of a causal effect of sales rank information on demand, but a mere reflection of the serial correlation between the unobserved factors that affect the demand for a product. Movements of past ranks across ranges determined by the manner the top apps are displayed were used to control for the possible endogeneity of ranks. These movements were shown to provide good instruments for past ranks. The instrumental variables estimates are very close to the least-squares estimates, suggesting that the potential endogeneity of past ranks does not significantly affect the estimation results. Furthermore, the apparent steep jumps that occur at ranks 25 and 50 indicate that changes in the sales ranks cause significant changes in demand. These discrete demand jumps are unlikely to be due to auto-correlated errors.

The results indicate that, all things equal, releasing a new version of an app does not increase demand. This finding runs counter to the folklore on developers' forums, according to which releasing a new version increases the visibility of a

[21] A different kind of unobserved quality arises if the quality of an app changes significantly from one version to the other. App developers prefer to release apps with a significantly altered functionality under a slightly different name (e.g., Scramble 2, Tap-tap-revenge 2, etc.). As such, it is unlikely that the quality of the same product is improved significantly over time. Even if it were, the instrumental variables method makes it possible to estimate the relationship between demands and ranks using the component of rank variability that is uncorrelated with the unobserved differences in product quality (see Angrist and Krueger [2001]).

product; it also indicates that the lists of newly released apps made public by Apple may have little effect on consumer choice. However, as implied by the negative and significant coefficient on the version age variable ($VAge$), demand is lower for apps that are not regularly updated.

The estimate of the price coefficient is negative, as expected, and strongly significant. Since the estimated coefficient is the the unknown scale parameter of the power law multiplied by the true price coefficient, own-price elasticities are not identified.

The interpretation of the empirical results is subject to a caveat. Since apps could reach a large, but finite number of potential consumers, it is possible that the demands for some products become saturated over time. I attempt to control for this saturation effect by using the age of an app as a determinant of demand. The estimated coefficients on both app age and its square were positive and significant. In addition, survey evidence (see footnote 15) indicates that more than half of iphone and ipod touch users download less than one paid app per month. Additionally, the data indicate that the mean survival time for an app on the most downloaded list is close to a month. These suggest that saturation is not likely to significantly affect the results.

Given the relatively large estimated value attributed by consumers to the apps ranked on top of the best selling list, it is puzzling why top apps do not maintain their ranks for longer periods than observed. A candidate answer is that there are many factors other than rankings that affect consumer choice. For instance, advertising by Apple and conversion of freely downloadable apps into paid ones play an important role in shaping consumer preferences. With the help of additional data on advertising and conversion of free apps, the effect of factors other than past app rankings that affect consumer choice may be identified.

6 Conclusions

The models of herding and information cascades of Banerjee (1992), Bikhchandani, Hirshleifer and Welch (1992) and Welch (1992) have analyzed interactions between consumers that may give rise to inefficient outcomes. These models of behavior assume the existence of certain forms of learning from the actions of others. Testing this assumption may turn out to be critical for our understanding of economic

phenomena. If learning is shown not to permeate relationships among individuals, then assuming away information externalities may be appropriate. Conversely, if learning from the actions of others is shown to play an important role in shaping consumer behavior, then its role in economic theory and policy may need to be re-assessed.

A few recent papers (Salganik, Dodd and Watts 2006; Cai, Chen and Fang 2009; Conley and Udry 2009) provide important and convincing evidence of the effects of observational learning on demand. In line with the findings of these studies, this paper provides evidence that public past demand information significantly affects the purchase decisions of consumers. This paper strengthens and complements the findings of the recent experimental literature on observational learning by providing rank-by-rank estimates of the effect of popularity information on demand.

A central challenge in identifying the causal effect of public rank information on demand is the possible endogeneity of past ranks. Differences in the manner by which the products are displayed on the bestseller list suggest credible instruments for the past ranks. The instrumental variable estimates are very close to the ordinary least squares ones, suggesting that the estimation results are not significantly affected by the potential endogeneity of previous-day ranks.

One of the main findings of this paper is that the public bestseller status of top ranked apps is a very important determinant of demand. However, the results of this paper indicate that in the app market the component of willingness to pay that is attributable to bestseller status information declines very steeply for the top 10 products, but levels off somewhat at economically meaningful values for products ranked between 11 and 50. At higher ranks, a change in bestseller status does not cause a significant change in demand.

The data available for this study do not contain information on the market shares of products, but contain information on their sales rankings. This lack of data calls for an estimation procedure that avoids the market share data requirements of existing empirical models of discrete consumer choice. In this paper I developed a procedure that permits the estimation of market demands using rank data. One of the costs in terms of loss of generality of using this procedure is likely modest and comes from the assumption that sales in the upper tail of the distribution follow a power law. This assumption is consistent with evidence from several markets of digital goods, including the market for apps studied in this paper. A second, more

important cost of using the procedure is a drastic reduction in the range of empirical models of discrete consumer choice that can be taken to the data.

Appendix

Proof of Lemma 1.

Proof. Observe first that $\ln(S_t/S_t^m)$ is exponentially distributed with parameter θ. To see this, note that $\Pr[\ln(S_t/S_t^m) > s] = \Pr[S_t > S_t^m e^s] = e^{-s\theta}$. Observe next that, for $k \in 1,...,99$, $\ln(S_{kt}/S_t^m) - \ln(S_{k+1t}/S_t^m) = \ln S_{kt} - \ln S_{k+1t} = \ln s_{kt} - \ln s_{k+1t}$, the last equality because $s_{kt} = S_{kt}/\sum_{i=0}^{N_t} S_{it}$ (S_{0t} denotes the sales of the outside good at time t). I prove next, following Rényi (1953), that the differences $y_{kt} = k(\ln s_{kt} - \ln s_{k+1t})$ are independently and identically exponentially distributed with parameter θ. The first step is to show that these differences are exponentially distributed. Since $\ln s_{kt}$ are exponentially distributed, for $a, b > 0$,

$$\Pr[\ln s_{kt} - \ln s_{k+1t} > a | \ln s_{k+1t} = b] = \Pr[\ln s_{kt} > a + b | \ln s_{k+1t} = b].$$

Since the exponential process is memoryless,

$$\Pr[\ln s_{kt} > a + b | \ln s_{k+1t} = b] = \Pr[\ln s_{kt} > a].$$

The right hand side of the last equality requires k of the i.i.d. exponential draws to be greater than a, so

$$\Pr[\ln s_{kt}\text{-}\ln s_{k+1t} < a | \ln s_{k+1t} = b] = 1 - [\exp(-\theta a)]^k = 1 - \exp(-k\theta a). \quad (13)$$

Note that the conditional probability in 13 is not a function of b, so that, for $k \in 1,...,99$, the variables $k(\ln s_{kt} - \ln s_{k+1t})$ are exponentially distributed with parameter θ. I prove next that the differences $k(\ln s_{kt} - \ln s_{k+1t})$ are statistically independent. Clearly, the property of the exponential distribution used above implies that

$$\Pr[\ln s_{kt} - \ln s_{k+1t} < a | \ln s_{k+1t} - \ln s_{k+2t} = b_{k+1}, ..., \ln s_{100t} = b_{100}] \quad (14)$$

is equal to $1 - \exp(k\theta a)$. Note that the probability in 14 is independent of the parameters b upon which it is conditioned. It follows that the differences $\ln s_{kt} - \ln s_{k+1t}$ and the variables y_{kt} (which are their multiples) are statistically independent. Since the mean of the differences $\ln s_{kt} - \ln s_{k+1t}$ is equal to $1/(k\theta)$, the variables y_{kt} have unit mean and the proof of the Lemma is concluded. ■

Variable	Description	Count	Mean	Stdev	Median	Min	Max
ID	Unique app ID number	452					
Rank	App rank on top 100 list	13,996	47.86	28.35	47	1	100
L(Rank)	Lagged rank	13,996	46.76	27.50	46	1	101
L_2(Rank)	Second lag	13,996	46.07	26.96	45	1	100
L_3(Rank)	Third lag	13,996	46.09	27.06	45	1	100
Price	Price in US Dollars	13,996	2.72	2.25	1.99	0.99	29.99
Age	Days since first release	13,996	109.40	83.15	85	0	340
VAge	Age of current version	13,996	92.63	75.85	66	0	340
Size	App size (mega bytes)	13,996	17.20	29.11	7.8	0	412
NewVer	New Version Dummy	13,996	0.0162	0.1263	0	0	1

Table 1: Summary Statistics

Notes: Lagged rank 101 corresponds to a previously unranked app. The size of apps is rounded to the nearest tenth of a megabyte, so one app (for which there are five total observations) has a size of zero. The squares of age, version age and size used in the regressions below are normalized.

Variable	Estimate	Stderr	t-value	Variable	Estimate	Stderr	t-value
M_1^*	2.0451	0.1833	11.1564	M_{31}^*	0.4498	0.0887	5.0704
M_2^*	1.3401	0.1712	7.8266	M_{32}^*	0.4455	0.0879	5.0683
M_3^*	1.1095	0.1637	6.7764	M_{33}^*	0.4592	0.0909	5.0525
M_4^*	1.0294	0.1625	6.3356	M_{34}^*	0.4284	0.0849	5.0493
M_5^*	0.967	0.158	6.1206	M_{35}^*	0.4504	0.0892	5.051
M_6^*	0.8883	0.1531	5.8025	M_{36}^*	0.444	0.0883	5.0308
M_7^*	0.8254	0.1457	5.6641	M_{37}^*	0.4213	0.0835	5.0471
M_8^*	0.808	0.1463	5.5245	M_{38}^*	0.4291	0.0852	5.0361
M_9^*	0.7679	0.1399	5.4887	M_{39}^*	0.4465	0.0887	5.033
M_{10}^*	0.7363	0.1358	5.4229	M_{40}^*	0.4223	0.0838	5.037
M_{11}^*	0.72	0.1341	5.3707	M_{41}^*	0.371	0.0735	5.0507
M_{12}^*	0.6896	0.1306	5.2812	M_{42}^*	0.4367	0.0869	5.0263
M_{13}^*	0.6497	0.1237	5.2526	M_{43}^*	0.3936	0.0781	5.0417
M_{14}^*	0.6402	0.122	5.2495	M_{44}^*	0.3479	0.0689	5.0482
M_{15}^*	0.6178	0.1184	5.2158	M_{45}^*	0.3302	0.0655	5.039
M_{16}^*	0.6441	0.1242	5.188	M_{46}^*	0.3859	0.0765	5.0441
M_{17}^*	0.6174	0.1197	5.1559	M_{47}^*	0.3679	0.0729	5.0484
M_{18}^*	0.636	0.1235	5.1491	M_{48}^*	0.3776	0.0749	5.0427
M_{19}^*	0.603	0.1171	5.1502	M_{49}^*	0.3368	0.0669	5.0365
M_{20}^*	0.6191	0.1207	5.1289	M_{50}^*	0.3274	0.065	5.0374
M_{21}^*	0.5671	0.1103	5.1428	M_{51}^*	0.3183	0.0631	5.0432
M_{22}^*	0.5716	0.1116	5.1232	M_{52}^*	0.2762	0.0547	5.0452
M_{23}^*	0.6092	0.1192	5.1129	M_{53}^*	0.227	0.0453	5.0113
M_{24}^*	0.5926	0.1161	5.1047	M_{54}^*	0.2313	0.0458	5.055
M_{25}^*	0.545	0.1066	5.1147	M_{55}^*	0.1701	0.0337	5.0523
M_{26}^*	0.5047	0.0991	5.0931	M_{56}^*	0.1972	0.0392	5.026
M_{27}^*	0.5069	0.1	5.069	M_{57}^*	0.1602	0.0317	5.0594
M_{28}^*	0.4771	0.0937	5.0925	M_{58}^*	0.1851	0.035	5.2824
M_{29}^*	0.4939	0.097	5.09	M_{59}^*	0.1235	0.0247	4.9991
M_{30}^*	0.4928	0.0971	5.0725	M_{60}^*	0.1212	0.0246	4.9193

Variable	Estimate	Stderr	t-value	Variable	Estimate	Stderr	t-value
M^*_{61}	0.1254	0.0252	4.9787	M^*_{85}	-0.0189	0.005	-3.7812
M^*_{62}	0.1142	0.0234	4.8734	M^*_{86}	-0.0481	0.0098	-4.9196
M^*_{63}	0.1298	0.026	5.0001	M^*_{87}	0.0041	0.0036	1.1293
M^*_{64}	0.1497	0.0301	4.9815	M^*_{88}	-0.0061	0.0037	-1.6292
M^*_{65}	0.0523	0.011	4.7531	M^*_{89}	-0.0114	0.0039	-2.9459
M^*_{66}	0.141	0.029	4.8551	M^*_{90}	0.0555	0.0121	4.6004
M^*_{67}	0.0792	0.0164	4.8328	M^*_{91}	-0.0286	0.0065	-4.414
M^*_{68}	0.062	0.0132	4.6888	M^*_{92}	-0.05	0.0102	-4.8778
M^*_{69}	0.0719	0.0151	4.7558	M^*_{93}	-0.0287	0.0066	-4.3501
M^*_{70}	0.055	0.0119	4.628	M^*_{94}	0.0017	0.0039	0.4408
M^*_{71}	0.0475	0.0103	4.6112	M^*_{95}	-0.0613	0.0124	-4.9495
M^*_{72}	0.0478	0.0103	4.66	M^*_{96}	-0.0096	0.0037	-2.6214
M^*_{73}	0.0782	0.0161	4.8609	M^*_{97}	-0.1022	0.0201	-5.0721
M^*_{74}	0.0204	0.0056	3.6263	M^*_{98}	-0.013	0.0038	-3.3996
M^*_{75}	0.0391	0.009	4.3517	M^*_{99}	-0.0768	0.0153	-5.0348
M^*_{76}	0.0553	0.012	4.6115	M^*_{100}	-0.0697	0.0141	-4.9563
M^*_{77}	-0.0052	0.0036	-1.4208	Price (α^*)	-0.4707	0.095	-4.9572
M^*_{78}	0.0603	0.0129	4.6603	Age	0.0069	0.0014	4.8745
M^*_{79}	0.0327	0.0078	4.1673	Age^2	0.2235	0.0427	5.2299
M^*_{80}	0.0093	0.0042	2.193	Vage	-0.0089	0.0018	-4.9586
M^*_{81}	-0.0182	0.0048	-3.753	$VAge^2$	-0.0147	0.0047	-3.1109
M^*_{82}	0.0202	0.0057	3.5561	NewVer	-0.0333	0.012	-2.7833
M^*_{83}	-0.0417	0.0087	-4.8152	Size	0.0004	0.0006	0.6839
M^*_{84}	0.0053	0.0037	1.4373	$Size^2$	6.0596	1.2425	4.8769

Table 2: Estimation Results

Notes: Not reported are the coefficients on app dummies (417 coefficients were significant at the 10% confidence level). The estimating equation is (12). Price is instrumented using the logarithms of twice- and thrice-lagged ranks; the first-stage F statistic is equal to 437.19 (R^2 is 0.95). Means and standard errors are computed using 10,000 replications.

Variable	Estimate	Stderr	t-value	Variable	Estimate	Stderr	t-value
M_1	4.4691	0.6982	6.4005	M_{31}	0.9567	0.0161	59.4172
M_2	2.9001	0.3087	9.3958	M_{32}	0.9475	0.0158	60.1395
M_3	2.3892	0.1941	12.3068	M_{33}	0.9766	0.0152	64.1258
M_4	2.2109	0.1466	15.0846	M_{34}	0.9111	0.0148	61.4237
M_5	2.0741	0.1217	17.0437	M_{35}	0.9579	0.0145	65.8779
M_6	1.9012	0.0917	20.7281	M_{36}	0.944	0.0136	69.2522
M_7	1.7646	0.0761	23.1902	M_{37}	0.896	0.0137	65.1802
M_8	1.7256	0.0638	27.0385	M_{38}	0.9124	0.0136	66.9403
M_9	1.6396	0.0604	27.1627	M_{39}	0.9494	0.0132	71.8266
M_{10}	1.5712	0.0525	29.9495	M_{40}	0.898	0.0129	69.6503
M_{11}	1.5356	0.0475	32.3545	M_{41}	0.7891	0.0131	60.3625
M_{12}	1.4698	0.0417	35.2203	M_{42}	0.9285	0.0129	71.9879
M_{13}	1.3844	0.0365	37.9804	M_{43}	0.837	0.0123	68.1065
M_{14}	1.3642	0.0363	37.5848	M_{44}	0.7398	0.0123	60.2381
M_{15}	1.316	0.0332	39.6454	M_{45}	0.7021	0.0118	59.7517
M_{16}	1.3717	0.0315	43.5146	M_{46}	0.8205	0.0122	67.1208
M_{17}	1.3144	0.0293	44.8485	M_{47}	0.7823	0.0121	64.7306
M_{18}	1.3539	0.0282	48.0626	M_{48}	0.8031	0.0115	69.924
M_{19}	1.2837	0.0268	47.8388	M_{49}	0.7161	0.011	65.177
M_{20}	1.3176	0.026	50.6229	M_{50}	0.6961	0.011	63.2299
M_{21}	1.2071	0.0256	47.2066	M_{51}	0.6769	0.0106	63.6641
M_{22}	1.2166	0.024	50.618	M_{52}	0.5873	0.0114	51.3203
M_{23}	1.2964	0.0233	55.5654	M_{53}	0.4825	0.0101	47.7304
M_{24}	1.261	0.0226	55.7326	M_{54}	0.492	0.0097	50.9021
M_{25}	1.1598	0.0219	52.9999	M_{55}	0.3618	0.0091	39.6516
M_{26}	1.0739	0.0199	53.8683	M_{56}	0.4192	0.0096	43.5468
M_{27}	1.0783	0.0181	59.7022	M_{57}	0.3408	0.0096	35.5607
M_{28}	1.0152	0.0178	56.875	M_{58}	0.3946	0.0149	26.4781
M_{29}	1.0508	0.0186	56.5053	M_{59}	0.2627	0.0092	28.7073
M_{30}	1.0482	0.0168	62.3163	M_{60}	0.2576	0.009	28.545

Variable	Estimate	Stderr	t-value	Variable	Estimate	Stderr	t-value
M_{61}	0.2666	0.0086	30.9159	M_{81}	-0.0389	0.0085	-4.5635
M_{62}	0.2426	0.0091	26.556	M_{82}	0.0429	0.0086	4.9976
M_{63}	0.2761	0.009	30.8259	M_{83}	-0.0888	0.0079	-11.1993
M_{64}	0.3183	0.0086	36.9556	M_{84}	0.0111	0.0079	1.4143
M_{65}	0.1113	0.0088	12.7127	M_{85}	-0.0401	0.0078	-5.1524
M_{66}	0.2993	0.0087	34.5094	M_{86}	-0.1026	0.0081	-12.7142
M_{67}	0.1684	0.009	18.693	M_{87}	0.0084	0.0078	1.0831
M_{68}	0.1316	0.0088	15.0381	M_{88}	-0.0131	0.0084	-1.5608
M_{69}	0.1527	0.0088	17.2784	M_{89}	-0.0244	0.0076	-3.2073
M_{70}	0.1167	0.0088	13.3079	M_{90}	0.1178	0.0081	14.5665
M_{71}	0.1009	0.0081	12.5227	M_{91}	-0.0608	0.0078	-7.8078
M_{72}	0.1016	0.0083	12.3041	M_{92}	-0.1064	0.008	-13.3759
M_{73}	0.1662	0.008	20.7469	M_{93}	-0.0611	0.0078	-7.7782
M_{74}	0.0432	0.0082	5.2984	M_{94}	0.0033	0.0088	0.3712
M_{75}	0.083	0.0087	9.4914	M_{95}	-0.1305	0.0083	-15.6426
M_{76}	0.1173	0.0081	14.5286	M_{96}	-0.0207	0.0078	-2.6512
M_{77}	-0.0111	0.0081	-1.3722	M_{97}	-0.2174	0.0072	-30.2328
M_{78}	0.1279	0.0078	16.3914	M_{98}	-0.0283	0.0088	-3.1959
M_{79}	0.0692	0.008	8.6572	M_{99}	-0.1636	0.0081	-20.1672
M_{80}	0.0195	0.0083	2.3649	M_{100}	-0.1484	0.0081	-18.2835

Table 3: Rank-Dependent Values

Notes: Reported values are the means and standard deviations of the ratios between the estimates of M and the estimate of α for 10,000 replications. The estimating equation is (12). The value of M_{101} was normalized to zero.

Dep. variable: $\hat{P}$			Dep. variable: $L_1(\hat{Rank})^{-1}$		
Variable	Estimate	Stderr	Variable	Estimate	Stderr
$ln(L_2(\hat{Rank}))$	-0.2430	0.0442	$\hat{FB}_{25}$	0.0253	0.0045
$ln(L_3(\hat{Rank}))$	-0.0246	0.0318	$\hat{FB}_{50}$	0.0222	0.0043
F(443,12040)=524.892; $R^2 = 0.951$			F(443,12040)=15.169; $R^2 = 0.358$		

Table 4: First Stage Results

Notes: First stage regressors also include the variables in the main estimating equation (12): age, age squared, version age, version age squared, new version dummy, size, size squared, and fixed effects dummies. The variable FB_{25} takes the value of 1 if the third lagged download rank L_3(Rank)>25 and the twice lagged rank L_2(Rank)<26, and zero otherwise. FB_{50} takes the value of 1 if L_3(Rank)>50 and L_2(Rank)<51, and zero otherwise. All hat variables represent differences between the variables that correspond to successive daily ranks (see Theorem 1). The dependent variable in the right panel is reciprocal of lagged ranks of observations with successive daily ranks that correspond to the restricted model in which $M_k = 1/k$.

Variable	Estimate	Stderr
$L_1(\text{Rank})^{-1}$	2.5635	0.7362
Price	-0.5997	0.2762
Age	0.0083	0.0038
Age^2	0.2565	0.1112
VAge	-0.0103	0.0047
VAge^2	-0.0222	0.0125
NewVer	-0.0463	0.0280
Size	0.0029	0.0017
Size^2	7.3006	3.4145

Table 5: IV Estimates

Notes: The dependent variable is $\hat{Y}$ in (12). The first two variables are instrumented using the instruments in Table 4. Fixed-effects coefficients are not reported (421 out of the 423 fixed-effects coefficients are significant at 10% confidence level). Means and standard errors computed for 10,000 replications.

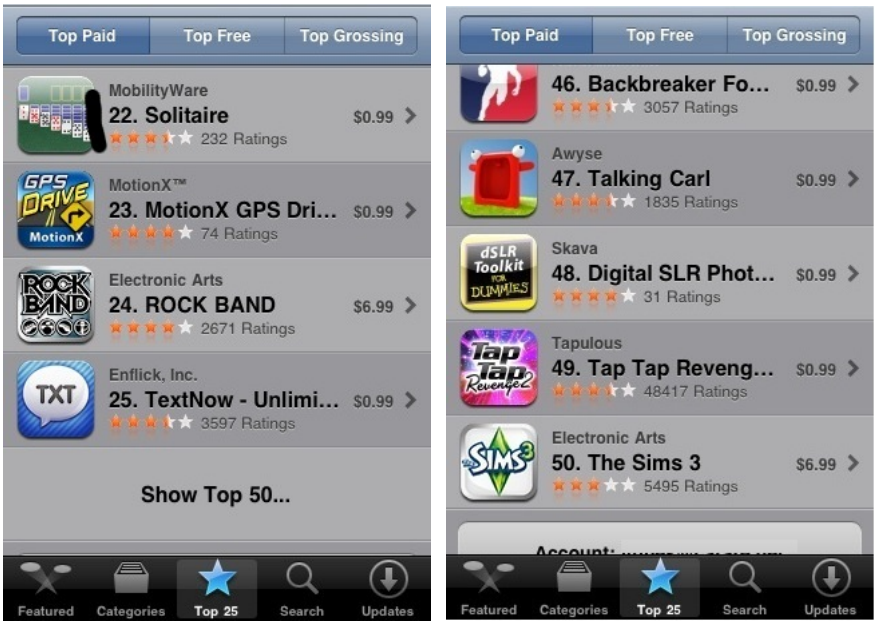

Figure 3: Top Apps Displayed on the Mobile Interface

References

[2001] Angrist, Joshua D. and Alan B. Krueger (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives* 15 (4), pp. 69-85

[1992] Banerjee, Abhijit V. (1992), "A Simple Model of Herd Behavior," *Quarterly Journal of Economics* 107 (3), pp. 797-818

[2001] Bassett, Troy J. and Christina M. Walter (2001) "Booksellers and Best-sellers: British Book Sales as Documented by 'The Bookman,' 1891-1906," *Book History* 4, pp. 205-236

[1994] Berry, Steven T. (1994), "Estimating Discrete-Choice Models of Product Differentiation," RAND Journal of Economics 25 (2), pp. 242-262

[1992] Bikhchandani, Sushil, David Hirshleifer and Ivo Welch (1992), "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades," *Journal of Political Economy* 100 (5), pp. 992-1027

[2003] Brynjolfsson, Erik, Yu Hu and Michael D. Smith (2003), "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science* 49 (11), pp. 1580-1596

[2009] Cai, Hongbin, Yuyu Chen and Hanming Fang (2009), "Observational Learning: Evidence from a Randomized Natural Field Experiment," *American Economic Review* 99 (3), pp. 864-882

[1991] Cardell, N. Scott (1991), "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory* 13 (2), pp. 185-213

[2003] Chevalier, Judith and Austan Goolsbee (2003), "Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com," *Quantitative Marketing and Economics* 1 (2), pp. 203-222

[2009] Conley, Timothy G. and Christopher R. Udry (2009), "Learning About a New Technology: Pineapple in Ghana," forthcoming, *American Economic Review*

[2002] Duflo, Esther and Emmanuel Saez (2002), "Participation and Investment Decisions in a Retirement Plan: The Influence of Colleagues' Choices," *Journal of Public Economics* 85 (1), pp. 121-148

[1995] Foster, Andrew D. and Mark R. Rosenzweig (1995), "Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture," *Journal of Political Economy* 103 (6), pp. 1176-1209

[1980] Kennedy, William J., Jr., and James E. Gentle (1980), *Statistical Computing*, Marcel Dekker, Inc., NY

[1950] Leibenstein, Harvey (1950), "Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand," *Quarterly Journal of Economics*, 64 (2), pp. 183-207

[2000] Manski, Charles (2000), "Economic Analysis of Social Interactions," *Journal of Economic Perspectives* 14 (3), pp. 115-136

[1995] Mason, Roger (1995), "Interpersonal Effects on Consumer Demand in Economic Theory and Marketing Thought, 1890-1950," *Journal of Economic Issues* 29 (3), pp. 871-881

[1978] McFadden, Daniel (1978), "Modelling the Choice of Residential Location," in *Spatial Interaction Theory and Planning Models*, A. Karlqvist *et al.* (eds.), Amsterdam: North Holland

[1968] Merton, Robert K. (1968), "The Matthew Effect in Science," *Science* 159, pp. 56-63

[1969] Milgram, Stanley, Leonard Bickman and Lawrence Berkowitz (1969), "Note on the Drawing Power of Crowds of Different Size," *Journal of Personality and Social Psychology* 13(2), pp. 79-82.

[1948] Morgenstern, Oskar (1948), "Demand Theory Reconsidered," *Quarterly Journal of Economics* 62, pp. 165-201

[2009] Moretti, Enrico (2009), "Social Learning and Peer Effects in Consumption," working paper, University of California Berkeley

[2004] Munshi, Kaivan (2004), "Social Learning in a Heterogeneous Population," *Journal of Development Economics* 73, pp. 185-213

[2000] Nevo, Aviv (2000), "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand," *Journal of Economics and Management Strategy* 9 (4), pp. 513-548

[1913] Pigou, Arthur C. (1913), "The Interdependence of Different Sources of Demand and Supply in a Market The Interdependence of Different Sources of Demand and Supply in a Market," *The Economic Journal*, 23 (89), pp. 19-24

[1953] Renyi, Alfréd (1953), "On the Theory of Order Statistics," *Acta Mathematica Hungarica* 15, pp. 191-227

[2006] Salganik, Matthew J., Peter S. Dodds and Duncan J. Watts (2006), "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," *Science* 311, pp. 854-856

[2005] Simkin, Mikhail V. and Vwani P. Roychowdhury (2005), "Copied Citations Create Renowned Papers," *Annals of Improbable Research* 11 (1), pp. 24-27

[1958] Simon, Herbert A. and Charles P. Bonini (1958), "The Size Distribution of Business Firms," *American Economic Review* 48 (4), pp. 607-617

[2007] Sorensen, Alan T. (2007), "Bestseller Lists and Product Variety," *Journal of Industrial Economics* 55 (4), pp. 715-738

[2008] Thorson, Emily (2008), "Changing Patterns of News Consumption and Participation," *Information, Communication and Society* 11 (4), pp. 473-489

[2009] Tucker, Catherine and Juanjuan Zhang (2009), "How Does Popularity Information Affect Choices? A Field Experiment," mimeo, MIT

[1992] Welch, Ivo (1992), "Sequential Sales, Learning, and Cascades," *Journal of Finance* 47 (2), pp. 695-732

[2009] Zhang, Juanjuan (2009), "The Sound of Silence: Observational Learning in the U.S. Kidney Market," forthcoming, *Marketing Science*