Using Virtual Stock Exchanges to Forecast Box-Office Revenue via Functional Shape Analysis

Wolfgang Jank
Department of Decision and Information Technologies
Robert H. Smith School of Business
University of Maryland
wjank@rhsmith.umd.edu

Natasha Foutz
Department of Marketing
Robert H. Smith School of Business
University of Maryland
nfoutz@rhsmith.umd.edu

ABSTRACT

In this paper we propose a novel model for forecasting innovation success based on online virtual stock markets. In recent years, online virtual stock markets have been increasingly used as an economic and efficient information gathering tool for the online community. It has been used to forecast events ranging from presidential elections to sporting events and applied by major corporations such as HP and Google for internal forecasting. In this study, we demonstrate the predictive power of online virtual stock markets, as compared to several conventional methods, in forecasting demand for innovations in the context of the motion picture industry. In particular, we forecast the release weekend box office performance of movies which serves as an important planning tool for allocating marketing resources, determining optimal release timing and advertising strategies, and coordinating production and distributions for different movies. We accomplish this forecasting task using novel statistical methodology from the area of functional data analysis. Specifically, we develop a forecasting model that uses the entire trading path rater than only its final value. We also employ trading dynamics and we tease out differences between different trading paths using functional shape analysis. Our results show that the model has strong predictive power and improves tremendously over competing approaches.

1. INTRODUCTION

Pre-release forecasting of demand for innovations is critical for allocating limited marketing resources, determining optimal pricing and advertising strategies, and coordinating production and distributions. Such forecasting is also difficult particularly for experiential new products and innovations with short lifecycles. Marketing researchers have used a variety of information sources in pre-release forecasting, such as historical demand for similar innovations and cross-sectional product features, demand for the same innovation in other distributional channels or geographical markets, advance purchase orders, as well as consumer surveys and consumer clinics.

In recent years, online virtual stock markets (VSMs) have been increasingly used as an information gathering tool for online communities and as an effective forecasting tool. They belong to a more general class of information markets, also known as prediction market, idea market, event futures, or betting exchange. In online VSMs, assets are created whose final cash value is tied to a particular event (e.g., will the next U.S. president be a Republican) or parameter (e.g., opening weekend box office revenues of a movie). Participants trade the assets over time and the market price at each time point can then be interpreted as predictions of the probability of the event occurring or the expected value of the parameter. Those who buy low and sell high are rewarded for improving the market prediction, while those who do the opposite are punished for degrading the market prediction.

Online VSMs have been used in a variety of domains, such as politics, sports and entertainment, economic and business events (e.g. HedgeStreet), natural disasters (Hurricane Futures Market at University of Miami), and other events of interests to the public. Political stock markets have a long history in the U.S. The most well-known are organized information markets on Wall Street (1880 - 1944), the Iowa Electronic Market (1988 - present) that has predicted the U.S. presidential elections more accurately than traditional polls 75% of the time, and TradeSports (2001 - present). Related to the entertainment industry, the Hollywood Stock Exchange (HSX hereafter) was established in 1996 in which traders buy and sell shares of movies, actors, directors, and film-related options. HSX has correctly predicted 35 of 2005's 40 big-category Oscar nominees and 7 out of 8 top category winners.

Online VSMs can efficiently aggregate public and privately held information and have been demonstrated in various empirical studies to provide reliable and accurate forecasts for future events [Spann and Skiera, 2003, Forsythe et al., 1999, Pennock et al., 2001a, Pennock et al., 2001b, Leigh and Wolfers, 2006, Berg and Rietz, 2003]. The success of online VSMs in forecasting has inspired a growing stream of research from the fields of political science, policy analysis, economics, finance, computer science, and information technologies. However, its potential applications in marketing decision making, particularly in forecasting demand for innovations, have been extremely limited.

EC'07, June 11–15, 2007, San Diego, CA.
Copyright 2007 ACM – ...$5.00.

In this study, we demonstrate the predictive power of online VSMs, as compared to several conventional methods, in forecasting demand for innovations in the context of the motion picture industry. Specifically, in contrast to existing methods that focus on using only the very last trading price in pre-release forecasting [Pennock et al., 2001a, Pennock et al., 2001b, Spann and Skiera, 2003], we develop an innovative model that uses the entire trading path. By trading path we mean the entire history of VSM trading values. For instance, our model considers the shape of the trading history, whether the path is convex or concave, whether it is increasing or decreasing, the rate of increase (or decrease), as well as the volatility of the trading path. Our results show that there is significant value in the trading path. In particular, the trading path of the online VSMs, specifically that of HSX, has tremendous predictive power on the release weekend box office performance of movies, which is considered by the industry as the best indictor of any film's overall success in theaters as well as in subsequent markets such as videos and international markets.

Our modeling approach is housed within the modern statistical framework of *functional data analysis* (FDA). In contrast to classical statistics where the focus is on a sample of data-vectors, the interest in FDA centers on a sample of functional objects such as the trading paths of VSMs. Using a smooth representation of the trading path, FDA also allows us to gauge trading dynamics such as the trading velocity and its acceleration. Our model teases out the difference in shape between different trading paths and trading dynamics. To that end, we make use of *functional shape analysis*. Using the resulting trading shapes and the corresponding shape dynamics, we achieve accurate, early, and dynamic pre-release forecast of box office revenue.

In addition, we also show how our modeling approach can be used to extract knowledge from *partial trading paths*. By partial trading paths we mean trading histories that are available only for part of the entire trading time. On HSX, movies are traded for several months, sometimes even years, before the release of the movie, and in the motion picture industry, the most important marketing decisions e.g. release timing, advertising or distribution planning occur prior to the actual release of the movie. Therefore, pre-release and early forecasting of a movie's demand is essential. We show how to use partial trading shapes, i.e. shapes that are available only until several weeks or months before the release of the movie. We also determine the optimal time for decision-making by quantifying the incremental amount of information that is gained for each additional week of trading towards the movie release.

Our results show that online VSMs provide superior pre-release forecasts, as compared to several conventional methods, such as historical demand and cross-sectional information. The efficiency of online VSMs in aggregating pre-release information from the online community is further demonstrated as the inclusion of other pre-release information such as film characteristics and pre-release advertising adversely impacts the predictive performance of online VSM.

2. DATA

We use functional data analysis to develop a novel forecasting model for box office performance. Our model operates on the trading history of movies from the Hollywood Stock Exchange (HSX).

HSX is a virtual stock market established in 1996 where more than half a million active members trade virtual stocks of upcoming films based on their expectations of the films' theatrical revenues. A registered user is given a free membership and two HSX dollars (H$2), equivalent to real world $2 million, to start with and can increase the value by strategically managing his/her portfolio. Each film is IPO-ed months or sometimes years prior to its theatrical release and is traded up to four weeks after its wide release in theaters. We use every Friday's daily average trading prices and trading volumes (or shares traded)[1].

In the following we describe in detail our data sources. The primary source is data from HSX. We supplement this data with box office data from Nielsen EDI and advertising data from TNS Media Intelligence

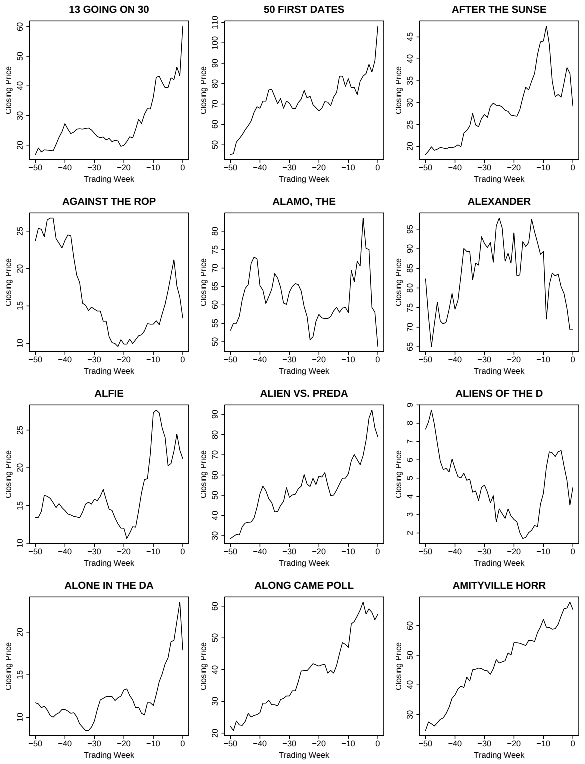

Figure 1: Trading paths of the first 12 movies in our data. The x-axis denotes time *from* the release of the movie; the y-axis denotes weekly HSX closing prices. We can see that the shapes of the trading paths vary quite considerably across movies.

2.1 Description of HSX Data

Our movie database consists of information on 262 movies. Figure 1 shows trading paths for the first 12 movies in our data. We can see that the shapes of the paths vary quite considerable across movies. Specifically, while the HSX price increases towards the movie release date for some movies, it

[1] We use weekly rather than daily data since we are mainly interested in longer-term trends rather than short-term fluctuations.

decreases for others. Moreover, the way prices increase (or decrease) varies quite a lot. While in some movies (e.g. *13 going on 30*) it increases sharply only at the end, other movies (e.g. *Amittyville Horror*) show a rather gradual increase. Moreover, some movies (e.g. *Alexander*) show both increase and decease in price, and there exist considerable differences in the amount of price variation (e.g. comparing *13 going on 30* and *Alexander*). Our goal is to capture these differences in price paths and to use them to better forecast box office revenue.

2.2 Description of Box-Office Revenue Data

We supplement the data from HSX with box office data from Nielsen EDI and advertising data from TNS Media Intelligence. After merging all available databases, we end up with 262 films released to the U.S. theaters between December 2003 and July 2005. The Nielsen EDI data contains weekly theatrical box office revenues for each film's first ten weeks' screening. Films have short life cycles. The theatrical release week often represents forty percent and the first four weeks eighty of a film's total theatrical revenues as box office revenues often exhibit a exponential decay pattern.

Table 1: Descriptive Statistics of our data sample

Sequel: number (%)	30 (11.45%)
Average production budget ($m)	45.84
Genre: number (%)	
ACTION	22 (8.40%)
ANIMATED	12 (4.58%)
COMEDY	92 (35.11%)
DRAMA	73 (27.86%)
HORROR	13 (4.96%)
SCI-FI	14 (5.34%)
SUSPENSE	22 (8.40%)
OTHERS	14 (5.34%)
MPAA Rating: number (%)	
G, NC-17, NR	10 (3.82%)
PG, PG-13	162 (61.83%)
R	90 (34.35%)
Average run time (minutes)	106.94
Studio: number (%)	
BUENA-VISTA	25 (9.54%)
FOX	25 (9.54%)
MGM-UA	11 (4.20%)
MIRAMAX	17 (6.49%)
PARAMOUNT	21 (8.02%)
SONY	33 (12.60%)
UNIVERSAL	20 (7.63%)
WARNER	25 (9.54%)
OTHERS	85 (32.44%)

The data from Nielsen EDI also include various film characteristics: sequel, production budget, genre, MPAA rating such as R or PG-13, run time or duration of the film, studio such as Twentieth Century Fox, and cumulative box office revenues. Table 1 displays descriptive statistics of the 262 films used in our analysis.

We obtain weekly pre-release advertising spending across all media (e.g. newspaper and TV), from TNS Media Intelligence (formally known as TNS Media Intelligence/CMR). In the motion picture industry, more than half of the theatrical advertising budget is spent prior to the release of the film. Pre-release advertising typically starts 15 weeks prior to the release of a film and increases over the weeks leading to the release of the film. It usually peaks at the release weekend in an effort to maximize awareness of the film.

3. FUNCTIONAL FORECASTING MODEL

In this section we describe how to use functional data analysis (FDA) to derive a powerful forecasting model for box office revenue. The use of FDA for forecasting is rather novel and has been explored only recently. Recent successful examples include dynamic, real-time forecasting models for the price-trajectories of online auction prices [Jank et al., 2006, Wang et al., 2007]. Here, the setting is different. In this paper, we use the trajectories of HSX trading prices to forecast a movie's box office success. Our approach is also different from classical time series models. In time series models, the objective is to forecast a single sequence of (univariate of multivariate) observations over time. In contrast, our goal is to learn from patterns common across many sequences. Moreover, we also characterize differences in shape and dynamics which lends itself to a very powerful forecasting approach.

Derivation of our forecasting model consists of three steps. First, we derive *smooth* trading paths from the observed trading patterns. Smooth trading paths eliminate extraneous noise from the observed patterns. Moreover, the smoothness also allows for a derivation of the corresponding trading dynamics. As we will see, trading dynamics will play a crucial role in our forecasting model. After derivation of the smooth trading paths, we apply *shape analysis* to extract movie-specific trading features that distinguish the trading shape of one movie from another. These trading shapes play a central role in our forecasting model and we incorporate shapes not only of the trading paths but also of their dynamics. And lastly, after extracting trading shapes, we perform variable selection to identify only those shapes that have a significant impact on forecasting box office revenue. We explain these three steps in detail next.

3.1 Smooth Trading Paths

In this section we derive smooth trading paths from the observed HSX weekly closing prices. The idea is to represent the weekly closing prices as a smooth continuous curve in order to eliminate extraneous noise and to extract trading dynamics. This representation is accomplished using ideas from a rather novel set of tools from statistics, referred to as *functional data analysis* (FDA) [Ramsay and Silverman, 2005].

One very flexible and computationally efficient smoothing technique is the penalized smoothing spline [Ruppert et al., 2003]. Let $\tau_1, \ldots, \tau_L$ be a set of knots. Then, a polynomial spline of order p is given by

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p + \sum_{l=1}^{L} \beta_{pl}(t - \tau_l)_+^p, \quad (1)$$

where $u_+ = uI_{[u \geq 0]}$ denotes the positive part of the function u. Define the roughness penalty

$$\mathrm{PEN}_m(t) = \int \{D^m f(t)\}^2 dt, \quad (2)$$

where $D^m f$, $m = 1, 2, 3, \ldots$, denotes the mth derivative of the function f. The penalized smoothing spline f minimizes the penalized squared error

$$\mathrm{PENSS}_{\lambda,m} = \int \{y(t) - f(t)\}^2 dt + \lambda\, \mathrm{PEN}_m(t), \quad (3)$$

where $y(t)$ denotes the observed data at time t and the smoothing parameter λ controls the trade-off between data-fit and smoothness of the function f. Using $m = 2$ in (3) leads to the commonly encountered cubic smoothing spline. Other possible smoothers include the use of B-splines or radial basis functions [Ruppert et al., 2003].

In this study, we use smoothing splines of order $p = 4$, a smoothing parameter of $\lambda = 50$, and we place a knot at every week of the trading period. Since we consider 52 trading weeks for each movie, this results in a total of 52 knots per trading shape. While the choice of smoothing parameters can appear arbitrary, our specific selection is guided by the goal of obtaining smooth functional objects that (visually) represent the original data well (see Figure 2). Moreover, we also conducted a robustness study and found that the results do not vary much for different choices of the smoothing parameters.

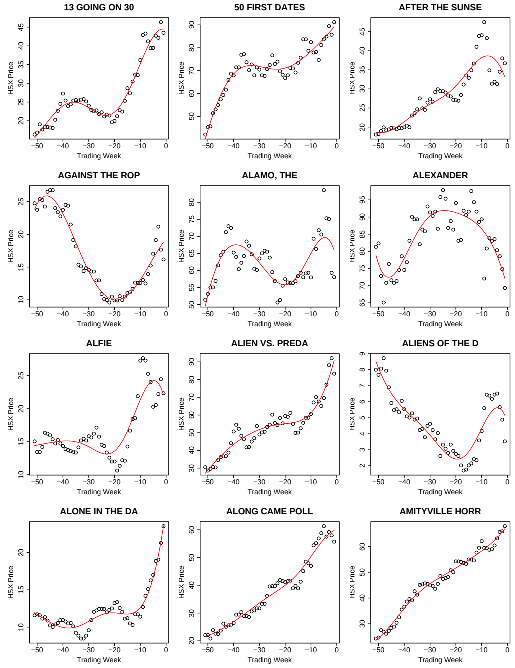

Figure 2: Smooth Trading Paths of the first 12 movies in our data. The circles show the actual weekly prices; the line shows their smooth representation.

Figure 2 shows smooth trading paths for the first 12 movies in our data. We can see that, compared to the raw trading paths in Figure 1, the smooth paths eliminate noise (e.g. *Alexander*) and extract the underlying pattern. In fact, there are movies with distinctively different patterns: Some movies (e.g. *13 Going on 30*) start low and end high, which is quite the opposite for other movies (e.g. *Aliens of the Dark*). Moreover, the *way* in which many movies behave between the start and the end is very different: While both *13 Going on 30* and *Amityville Horror* go from low at the beginning to high at the end, trading prices for the latter movie grow at a constant (almost linear) rate during the entire trading period, while the growth rate changes drastically for the former movie. In fact, prices for *13 Going on 30* are almost constant for most of the earlier trading periods, while they shoot up towards the end. Trading patters differentiate further: Some movies show rather convex shapes (*Against the Ropes*) while the shape is concave for others (*Alexander*). We believe that these different patterns carry significant information about the movie and its success and therefore we set out to incorporate them into our forecasting model.

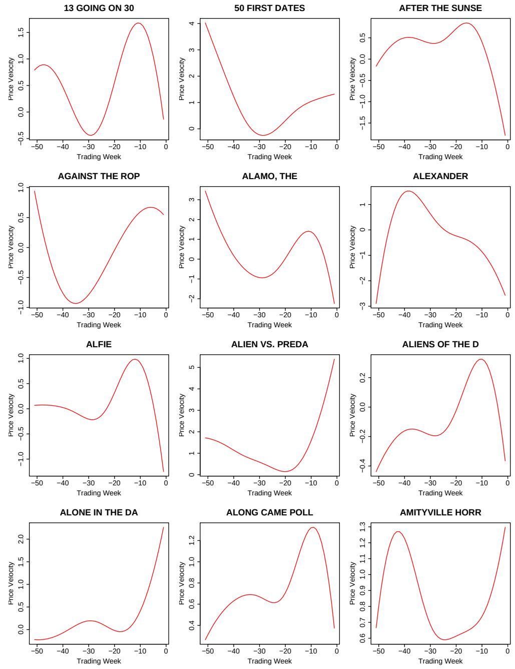

Figure 3: Trading velocity for the 12 movies.

3.2 Trading Dynamics

The smooth curves $f(t)$ in Figure 2 represent trading prices at any time t for a particular movie. We refer to $f(t)$ as the price path or the trading path. While $f(t)$ describes the exact *position* of price for any t, it does not reveal how fast the price is *moving*. Attributes that we typically associate with a moving object are its *velocity* (or its *speed*) and its *acceleration*. Because we use smoothing splines to obtain $f(t)$, velocity and acceleration can be readily computed for each movie via the first and second derivatives of $f(t)$, respectively.

Figure 3 shows the velocity of trading price for the 12 movies from Figure 2. (We also compute the trading acceleration. But for lack of space, the results are not shown here.) We can see that price velocity is quite heterogeneous both across all 12 movies but also within each movie individually. Take for instance *30 Going On 30*. For that movie, price velocity increases between week -50 and -40 implying that the speed at which price increased grew faster and faster during that time period. After week -40 or so, price velocity decreases. In fact, at about week -30, price velocity turns negative ("deceleration") indicating that prices actually go down (this can also be verified from Figure 2).

There are several reasons for considering the trading dynamics in addition to the trading paths. First, dynamics (velocity, acceleration) are measures of instantaneous change. While a change in the trading path (e.g. up/down) is associated with a change in dynamics (e.g. velocity positive/negative), dynamics make changes more pronounced and can also filter-out small and seemingly unimportant changes. Second, dynamics *anticipate* change. What we mean by that is that, e.g., a change in velocity is preceded by a change in acceleration; similarly, a change in the trading velocity occurs *before* a change in the trading path. In that sense, dynamics are *forward-looking* which makes them especially powerful in a forecasting setting. And lastly, dynamics can capture many of the unobservable factors influencing the trading path. For instance, while we can observe that the price increases sharply between two time periods, we cannot observe *why* it increases. Reasons for the increase may be the release of new movie information, or the start of a new advertising campaign. But reasons may also be more intricate such as a new advertising campaign for one movie and the simultaneous advertising of another movie. Changes may also be caused by an alteration of a movie's perception in a changing social or political environment (e.g. movies with a theme on climate, war, elections, etc). While we generally cannot observe the *cause* of price changes from one time point to another, dynamics capture their immediate *impact*.

3.3 Shape Analysis of Trading Paths

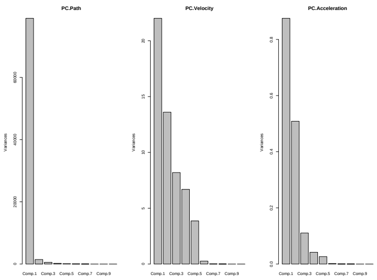

Figure 4: Scree plot of the first 9 principal components. The left panel shows the scree plot for the trading path, the middle panel shows the trading velocity, the right panel shows the trading acceleration.

The key idea of our forecasting model is to take advantage of different shapes in the HSX trading pattern. The rationale is that a trading pattern that increases very sharply towards the movie release time carries different information than one that grows at a much smaller rate. The sharp increase could be indicative of a last moment "hype" associated with the movie and such a hype will have a much different effect than movies without this hype. Similarly, a concave shape (sharp increases early, then flattening out or even decreasing) carries different information than a convex shape (little to moderate increases early, followed by a period of no activity, then sharp increases towards the end). Also, movies with a large amount of inter-trading volatility (ups and downs in the trading path, or equivalently, changes from negative to positive in the trading dynamics) could be indicative of an increasing amount of uncertainty associated with the movie; it could also point to simultaneous movies that directly compete for the same viewership. All-in-all, while there may be many reasons that cause heterogeneity in trading patterns, our goal is to capture their shape and to use shape-information to our advantage in the forecasting model.

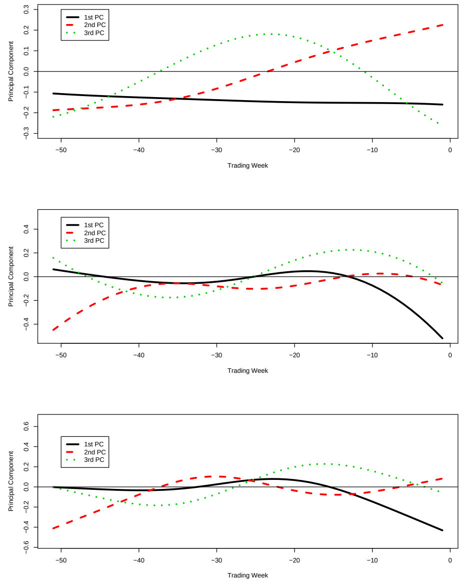

Figure 5: Principal component based trading shapes of the trading path and the trading dynamics. The top panel shows the first 3 principal components for the price path. Similarly, the middle and bottom panel show the corresponding principal components for the price velocity and acceleration, respectively.

In order to capture differences (and similarities) across trading patterns, we employ shape analysis. Shape analysis is done via functional principal component analysis [Ramsay and Silverman, 2005]. Let $\mathbf{Y}^s = (\mathbf{y}_1^s, \ldots, \mathbf{y}_n^s)$ denote the matrix of all smooth trading paths in our data, where n denotes the number of movies ($n = 262$ in this case). Let $\mathbf{R} := \mathrm{Corr}(\mathbf{Y}^s)$ be the correlation matrix obtained from $\mathbf{Y}^s$. Functional principal analysis decomposes $\mathbf{R}$ into $\mathbf{P}^T \mathbf{\Lambda} \mathbf{P}$ where $\mathbf{\Lambda}$ is a matrix of eigenvalues and $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_t]$ is a matrix of eigenvectors. Then, $\mathbf{e}_1$ contains $\lambda_1 \times 100\%$ of the variability in $\mathbf{Y}^s$, where λ_1 is the first diagonal component of $\mathbf{\Lambda}$. We can think of each of the eigenvectors $\mathbf{e}_i, i = 1, \ldots, t$, as shape-defining characteristics. For instance, an eigenvector may bring out the difference in price-magnitude between the start and the end of the trading period. Alternatively, an eigenvector may bring out the change in price (gradual/steep or convex/concave) between adjacent trading periods. Each eigenvector is associated with a corresponding eigenvalue. The common practice is to choose only those eigenvectors that correspond to the largest eigenvalues (i.e. those that explain most of the variation in $\mathbf{Y}^s$). By discarding those eigenvectors that explain no or only little of the variation, we get rid of extraneous noise. The eigenvector corresponding to the largest eigenvalue is typically denoted the first principal component. (Similarly, the second principal component is the eigenvector with the second highest eigenvalue, and so on.)

We select the significant eigenvectors according to the proportion of the total variation that they carry. Figure 4 shows the corresponding scree plot. We can see that for the trading path, the first principal component explains almost all of the variation. While the second and third component still carry some information, higher order components are negligible. The picture is somewhat more diffuse for the trading velocity and -acceleration. While the first principal component still carries most of the velocity information, components 2-5 also explain a significant portion of the total variation. Similarly, the first 5 principal components of the trading acceleration carry most of the information. For that reason, we initially retain the first 5 principal components for the trading dynamics. We reduce this number later and let the data decide which components have a significant impact on forecasting box office revenue.

The resulting principal components are shown in Figure 5. (Notice that we only show the first 3 principal components of the trading dynamics to avoid cluttering.) The principal components rotate the original data into a space where all the data-dimensions are orthogonal to one another. In that sense, each principal component can be interpreted as weighting factor. If the component value is high, then the corresponding data point receives a large weight; conversely, a data point receives only a small weight if the value is low. In similar fashion, principal component values with opposite signs (positive/negative), exhibit weighting in opposite directions. In fact, the weight that the principal component exhibits on the original data is referred to as "principal component score" (PC score). For instance, the first PC score for the first observation is defined as $S_1 = y^s_{11}e_{11} + \cdots + y^s_{1t}e_{1t}$, where $\mathbf{y}^s_1 = (y^s_{11}, \ldots, y^s_{1t})$ denotes the smooth trading path (evaluated on a finite grid) and $\mathbf{e}_1 = (e_{11}, \ldots, e_{1t})$ denotes the components for the first principal component. In other words, the first PC score for the first trading path can be thought of as the inner product between $\mathbf{y}^s_1$ and $\mathbf{e}_1$. (Actually, the principal components operate on standardized data, so it is more precise to think of the PC score as the inner product of the first eigenvector and the first *standardized* trading path.)

With this in mind, we can make the following observations from Figure 5. Notice that the first PC for the trading path (solid black line, top panel) is negative over the entire period and almost constant. Like a function that puts almost equal weight on each trading period, this first PC captures the difference in average trading levels between two different movies. For example, consider Figure 6 which shows the first 3 PC scores for three select movies. The first PC score for *13 Going On 30* is denoted by the circle in the bottom panel (similar for the other two movies). Recall that this PC score is essentially equivalent to the inner product of the first principal component (black line in top panel of Figure 5) and the movie's trading path (red line in top left panel of Figure 6). We notice that the first PC scores for the three movies in Figure 6 are very different: while the score is positive for *13 Going On 30*, it is negative for the other two movies and lowest in value for *Alexander*. The reason for this is that, as pointed out above, the first principal component merely measures differences in average trading levels: The average trading price across all movies and all trading periods is H$32. Notice that the average price for *13 Going on 30* across all trading periods is only H$27; on the other hand, the average prices for *Alexander* and *Amityville* are H$84 and H$47, respectively. The first PC score captures this difference and characterizes *13 Going on 30* as significantly below the average, whereas the other two movies are significantly above the average[2].

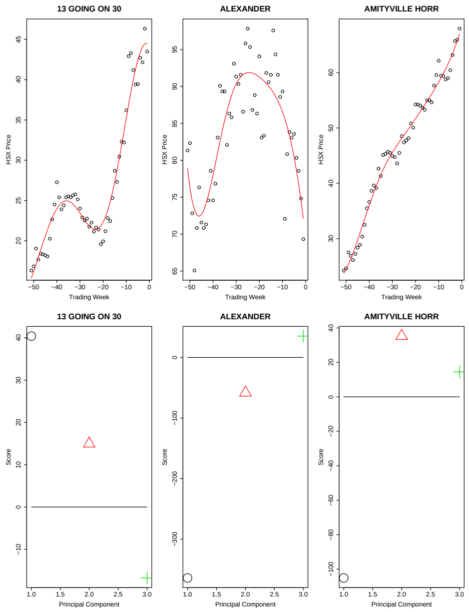

Figure 6: Smooth trading paths for three select movies (top panel) and their corresponding PC scores (bottom panel). The circle denotes the first PC score, the triangle denotes the second PC score and the cross denotes the third PC score.

We can draw a similar conclusion for the second principal component of the trading path (dashed red line in Figure 5). The second principal component is negative in the first half of the trading period, but positive in the second half. This implies that it contrasts differences between early and late trading. Take again the three examples in Figure 6. The second PC scores are given by the red triangles in the bottom panel. We can see they are positive for *13 Going On 30* and for *Amityville*, but negative for *Alexander*. This implies that differences between early and late trading are strong for *13 Going On 30* and for *Amityville*, but these differences are weak for *Alexander*. Differently put, the price for the former two movies grows strongly between the start and the end of trading, while it does not grow much for the latter movie (in fact, the price for *Alexander* at the time of release is almost the same as 50 weeks prior to release). Put differently yet another time, the second PC scores contrast the *trading shape* between the start and end of trading (which is linear or super-linear for the two former movies, but concave for the latter movie).

Using similar rationale, we can conclude that the third principal component (dotted green line in Figure 5) compares mid-term trading with early and late trading. In-

[2] In other words, the first principal component essentially captures differences in the magnitude of predicted box office revenues, say, differences between a $50 million movie and a $500 million movie.

deed, the corresponding third PC scores in Figure 6 (green crosses) indicate that mid-term trading for *13 Going On 30* and for *Alexander* are very different compared to early and later trading. However, differences are not as pronounced for *Amityville*.

We can make similar observations for the shapes of the trading dynamics. For instance, the first PC of the trading velocity emphasizes velocity spurts during the last trading moments (the principal component values are almost zero except for at the end, with increasing magnitude towards the movie release time). In contrast, the second PC emphasizes velocity spurts at the beginning of the trading period. Similarly, the third PC measures volatility in trading velocity between the first and second half of the trading period. Table 2 summaries our findings.

Table 2: HSX Trading Shapes and Dynamics

	PC	Shape Name
Path	PC1	Average trading level
	PC2	Early vs. late trading
	PC3	Mid-term trading
Velocity	PC1	Last moment velocity spurts
	PC2	Early velocity spurts
	PC3	Velocity volatility
Acceleration	PC1	Last moment acceleration spurts
	PC2	Early acceleration spurts
	PC3	Acceleration volatility

3.4 Variable Selection of Trading Shapes

Recall that based on Figure 4, we retained the first 3 principal components for the trading path and the first 5 principal components for the trading dynamics (velocity & acceleration) as potentially useful predictors of box office revenue. In the following we will investigate whether this rather large number of predictors can be reduced to a smaller, more parsimonious set. To that end, we will investigate the usefulness of the predictors for estimating box office revenue and we will employ classical measures of model fit.

For ease of notation, let P.PC1 denote the first principal component corresponding to the trading path and let P.PC2 denote its second principal component. In similar fashion, let V.PCi and A.PCi denote the principal components for the trading velocity and acceleration, respectively. Note that $i = 1, \ldots, 5$ since we retained the first 5 principal components.

In order to investigate the usefulness of the above predictors, we run a linear regression model using all (3+5+5)=13 predictors on box office revenue. (We run this model on the training sample; see also next section.) The results are shown in Table 3. We can see that except for V.PC1, all predictors are highly significant. In fact, the R-Squared value equals 0.90 (Adjusted R-Squared = 0.90). Interestingly, when eliminating the insignificant predictor and re-running the model, the significance levels of all remaining predictors reduce dramatically (while the R-Squared value remains almost identical). Since this kind of phenomenon frequently occurs when multicollinearity is present among the predictors, we investigated the pairwise correlations among all of the 15 predictors. Notice that e.g. P.PC1 and P.PC2 are orthogonal by construction of the principal components and thus pairwise correlations are no problem for predictors from within the same class (e.g. predictors from within the trading path or within the trading velocity). However, there exist correlations between predictors across different classes (e.g. between, say, the trading velocity and the trading acceleration). This is not surprising since it is plausible that some of the effects captured by the velocity are also captured by the acceleration. All-in-all, we eliminated 8 predictors with extremely high pairwise correlations. Using the remaining 5 predictors, we re-run the model. The results are shown in Table 4. Notice that we now have reduced the forecasting model to set of 5 parsimonious predictors The R-Squared/Adjusted R-Squared values now both equal 0.88, only slightly smaller compared to the previous, much larger model. We also investigate the effect of the dynamic shape variables. That is, we eliminate the two variables V.PC1 and V.PC2 from the model in Table 4 and re-run the regression. This results in R-Squared/Adjusted R-Squared values of 0.82 and 0.81, a drop compared to the previous model. We thus take this as evidence that the model in Table 4 is the most parsimonious description of the effect of trading shapes and their dynamics on a movie's success. Moreover, in order to give the variable selection approach described in this section even more credibility, we will also contrast our approach with two alternative approaches: classification & regression trees (CART) and generalized additive models (GAM) (see next Section).

Table 3: Estimating box-office revenue using all trading shape data.

	Estimate	StdErr	P-Val
(Intercept)	1.481e+07	4.672e+05	0.0000
P.PC1	-3.426e+10	1.220e+10	0.0055
P.PC2	1.744e+11	6.434e+10	0.0074
P.PC3	5.025e+11	1.737e+11	0.0043
V.PC1	6.223e+10	3.854e+10	0.1082
V.PC2	3.006e+12	1.063e+12	0.0052
V.PC3	-3.369e+12	1.160e+12	0.0041
V.PC4	-2.988e+12	1.095e+12	0.0070
V.PC5	9.326e+11	3.269e+11	0.0048
A.PC1	-2.457e+11	7.990e+10	0.0024
A.PC2	-2.773e+11	1.007e+11	0.0065
A.PC3	-8.544e+11	3.982e+11	0.0333
A.PC4	4.814e+11	1.583e+11	0.0027
A.PC5	7.962e+11	2.707e+11	0.0037

Table 4: Estimating box-office revenue after eliminating multicollinear shapes.

	Name	Estimate	StdErr	P-Val
(Intercept)	NA	15088081	495128	0.0000
P.PC1	Avg.trade.level	-50890	2556	0.0000
P.PC2	Early.late.trade	84355	14418	0.0000
P.PC3	Mid.term.trade	-72382	24900	0.0041
V.PC1	Late.vel.spurts	-1118574	113200	0.0000
V.PC2	Early.vel.spurts	-709968	211212	0.0009

4. RESULTS

4.1 Insight from the Forecasting Model

We use the model in Table 4 to forecast a movie's box office revenue. Interpreting its coefficients requires extra care because the model's input variables are the principal component scores derived in the previous section. Each principal component can be regarded as a linear combination of the information from the entire trading path and therefore has no easy interpretation. Moreover, the score corresponding to a particular principal component compares an individual movie to the average behavior of all movies. For instance, a movie's score on the first principal component of the trading path (top panel in Figure 5) will be large (in absolute value), if the movie consistently trades very high (or very low) relative to the average. In particular, its score will be a *large negative number* if it is traded very high above the average price and it will be a *large positive number* in the other case. In other words, since the principal component curve in Figure 5 is negative, the sign of the coefficients in Table 4 have to be interpreted very carefully.

With that in mind, we can make the following observations. The coefficient of the average trading level is negative; hence movies that trade high above the average (i.e. have a large negative score) will result in a higher box office revenue. This effect is strongly correlated with the final trading value: if the final trading value is high, its trading path is also likely to be high.

Table 4 also shows that the coefficient for the shape that compares early trading with late trading is positive. This implies that larger differences in price between the start and end of trading will have larger effects on box office revenue. Specifically, if the difference is positive (e.g. *13 Going On 30* or *Amittyville* in Figure 6), it will add a positive premium to box office revenue. On the other hand, if the difference is small, so will be the corresponding premium. A small (or even negative) difference between early and late trading can occur in several ways. One possible way is a concave shape of the trading path. For instance, the price for the movie *Alexander* initially increases only to decrease towards the movie's release. The resulting principal component score (Figure 6) is negative, indicating that this movie's trading path results in a negative premium.

A similar conclusion can be derived from the shape coefficient that compares mid-term trading. In particular, we notice that the coefficient is negative and hence "penalizes" movies with high mid-term trading. In other words, this shape implies a premium for movies that are traded predominantly towards the movie release. Take again the example of *13 Going On 30* (Figure 6). We notice that this movie experiences most of its trading activity in the last weeks before the movie release. (Notice the almost exponential increase in price in the last 20 weeks.) Consequently, its score on the mid-term principal component is low (in fact it is negative), implying that this movie's box office revenue will experience a premium. Now compare the other two movies in Table 4. Both of these movies have high mid-term trading activity and as a result experience negative box office premiums.

The conclusion from the two remaining coefficients (Late velocity spurts and early velocity spurts) are in line with the previous findings. Both coefficients are negative and thus reward movies with faster price increases either in the beginning or at the end. It is also noteworthy that the magnitude of the coefficient for late velocity spurts is one order higher than that of early velocity spurts: while early increases in trading speed add a premium to a movie's success, the effect almost doubles if that same increase happens late.

In summary, our forecasting models shows that different trading shapes have different effects on a movie's success: In general, the higher the trading price, the higher will be a movie's box office success. However, for two movies with the same trading price, the one that shows a stronger increase between early and late trading and particularly a faster increase towards the end will have higher box office revenue. It is particularly noteworthy that mid-trading activity results in a penalty.

4.2 Performance of the Forecasting Model

Our model focuses on the pre-release forecasting of the initial demand or release weekend demand for motion pictures. Forecasting demand accurately is important for movie marketing managers which is evidenced by the rich stream of literature on the topic [Shugan, 1998, Eliashberg et al., 2000, Krider and Weinberg, 1998, Foutz and Kadiyali, 2004]. In the following we measure the quality of our forecasting model and compare it to a set of competitor models. To that end, we perform a training sample/ validation sample approach. That is, we partition our data randomly into a 70% training sample which we use to estimate the parameters of our model and its competitors. We then use the remaining 30% to measure the predictive accuracy. All model comparisons are made on this validation sample.

We contrast a set of 7 competitor models. The baseline model is a model that consists of only information "traditionally" used to predict a movie's success. By traditional information we mean typical movie characteristics such as a movie's genre, budget, rating and runtime, whether or not the movie is a sequel, and also the producing studio. We refer to this baseline model as Model A. Besides these typical movie characteristics, an additional factor that has a strong impact on a movie's success is the amount of pre-release advertising. Consequently, our second competitor model (Model B) includes all typical movie characteristics plus pre-release advertising information for 10 weeks prior to the movie release. Our next competitor model (Model C) contains the same information as in Model B, plus the trading shapes from Table 4. In some sense, Model C examines how much can be gained by using the prediction market information *in addition* to movie characteristics and advertising information. Model D uses only the very last pre-release HSX trading price and ignores all the shape information in the trading paths [3]. And finally, our last model (Model E) uses *only* the trading shapes from Table 4. This model measures the pure impact of the shape of trading information fond on HSX's virtual stock exchange[4].

Models A-E are based on linear regression models. For comparison, we also run 2 additional models that serve as a contrast to the variable selection approach described in Section 4.4: A classification and regression tree (CART)

[3] Model D is similar but not identical to the model in [Spann and Skiera, 2003]. In contrast to the Spann & Skiera model, Model D only uses the final HSX trading price.

[4] Combining Models D and E does not lead to improved results because the final HSX trading price (Model D) is highly correlated with the first principal component of the trading path (first shape in Model E).

Table 5: Competitor Models

A	Genre, Budget, Rating, Runtime, Sequel, Studio
B	Same as in A, plus pre-release advertising
C	Same as in B, plus trading shapes from Table 4
D	Only very last pre-release HSX trading price (no trading shapes)
E	Only trading shapes from Table 4
CART	Using CART on all the 15 trading shapes from Table 3
GAM	Using GAM on all the 15 trading shapes from Table 3

and a generalized additive model (GAM). We run these two models on all of the 15 trading shapes in Table 3. This results in a total of 7 models. All models are summarized in Table 5.

We train all models on the training data. After estimating the parameters, we use the estimated models to predict box office revenue on the validation sample. The results are shown in Table 6. We can see that Model A has the worst performance. In fact, using only movie characteristics, the prediction error is almost 60%. One reason for this poor performance is that movie characteristics capture only information about the movie, but not how the movie is perceived among the public. Some of this perception is captured in Model B: It includes pre-release advertising information and one can argue that advertising is a successful tool for manipulating the public's perception. Consequently, Model B improves upon Model A by about 20%. Yet, true perception of a movie is unlikely proportional to advertising expenditures; rather, it is captured in reviews and opinions about the movie. One of the exciting features of virtual stock exchanges is that they are able to extract and aggregate individual opinions and convert them into a simple number: the movie's trading value. Indeed, we see that Model C (including the trading shapes, pre-release advertising and movie characteristics) significantly improves upon Model B, by over 25%. Most intriguing, however, is the performance of the model using only the trading shapes (Model E). We can see that its prediction error is only 8.31%, which is a 5% improvement over Model C. This result is very interesting and it appears to suggest that much of the information contained in movie characteristics and pre-release advertising is already captured in the movie's virtual stock value and its trading path. The performance of Model D shows how much can be gained by using the shape information on top of the information contained in the final HSX trading value: the predictive accuracy of Model D is about 3.5% worse compared to the model using all the shape information[5].

In order to evaluate and benchmark the performance of our model further, we compare it with two alternate approaches, CART and GAM. We notice that the performance of both CART and GAM is worse than that of Model E. CART is a flexible method with built-in mechanisms to determine the most useful predictors. In that sense, it is often regarded as an alternative approach to formal variable selection. The worse performance of CART suggests that the variable selection procedure outlined in Section 3.4 is quite effective in identifying important predictors and rendering a parsimonious yet accurate forecasting model. On the other hand, GAM is a flexible modeling approach which uses nonparametric methods to determine the form of the relationship between predictor and response. The worse performance of GAM suggests that the linear form of our model cannot be improved upon using non-linear functional relationships.

Table 6: Comparison of predictive performance on the validation sample. ***MAPE*** **denotes the Mean Absolute Percentage Error.**

Model	MAPE
A	59.50%
B	41.22%
C	13.56%
D	11.68%
E	8.31%
CART	12.99%
GAM	11.46%

4.3 Forecasting with partial trading shapes

The results discussed in the previous section take advantage of the *full* trading shape for each movie. By full shape we mean that we used all the information available until the release of the movie. In practice, however, it is often necessary to make decision well in advance of the movie release. For instance, marketing managers need to decide about the amount of pre-release advertising or the number and location of screens during the opening weekend. These decisions are often done weeks or months before the movie release. What this means for our analysis is that, at the point when the forecast is done, HSX trading information is available only until several weeks prior to the release of the movie. We refer to this as a *partial* trading shape.

In the following we investigate the ability of our model for making advance decisions. To that end, we assume that we have observed HSX trading shapes only partially until time t, $t = -40, -39, -38, \ldots, -1, 0$, where e.g. $t = -40$ indicates that we have HSX trading information only until 40 weeks *before* the release of the movie (similarly, $t = 0$ indicates that we have the information until the week of the movie release). In other words, the larger (closer to zero) t, the more complete information the manager has available; conversely, small values of t indicate smaller portions of available HSX information. Consequently, we refer to a trading shape derived from only partial information as a *partial* trading shape.

We conduct the same analysis as earlier, only this time we use partial trading shapes of length $-52, -51, \ldots, t$, where $t = -40, -39, -38, \ldots, -1, 0$. In that notation, using $t = 0$ recovers our analysis from the previous section. Figure 7 shows the results. The left panel of Figure 7 shows the reduction in MAPE as the length of the trading shape increases. We can see that for $t = -40$, the MAPE is about 36%. In other words, using only information available up to 40 weeks prior to the movie release, we can predict box

[5] Further improvements of Model E could be achieved by e.g. also controlling for the total time between the introduction of the stock and the release of the movie. Also, in contrast to [Spann and Skiera, 2003], we have no comparisons of our predictions with expert judgements.

office revenue with a relative error of 36%. This improves to about 8% if we use all the information available until the release of the movies ($t = 0$).

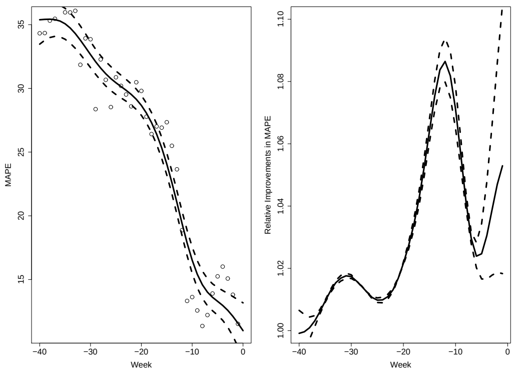

Figure 7: MAPE and length of the trading shape. The circles in the left panel show the individual MAPE values for different values of t ($t = -40, -39, -38, \ldots, -1, 0$). The solid line corresponds to a smoothing spline through the data. The dashed lines correspond to 95% confidence bounds. The right panel shows the relative improvement in MAPE, i.e. $MAPE_t/MAPE_{t-1}$.

The right panel of Figure 7 shows the *relative improvement* in MAPE as we go from one week to the next. For instance, the value of about 1.015 at week -30 implies that as we go from week -31 to week -30 (i.e. as we include one additional week worth of trading information into our decision making process), the MAPE increases by 1.5%. We notice that the largest relative improvement occurs at week -12: an improvement of 1.086 or 8.6%. We also notice that there are three local optima: the first optimum occurs early, at week -30; the second optimum occurs at week -12; and lastly, the third (and global) optimum is at week 0, the time of the complete trading information.

This finding has several implications for the decision maker. First, if most accurate predictions about a movie's success are desired, then one should wait as long as possible (i.e. until week 0). However, if decision making at an earlier time point is necessary, then there are two scenarios: for medium-range decisions (e.g. advertising decisions), the decision maker should wait until about 12 weeks before the movie release. This will ensure that one captures the most significant week-by-week improvements in predictive accuracy. On the other hand, for decision that need to be taken at even an earlier time point (e.g. release timing decisions), that decision should be done approximately 30 weeks before the movie release.

In summary, our model allows for early and dynamic predictions of a movie's box office success. The only required input into that prediction is trading information from a virtual stock exchange. Our model uses a movie's virtual stock value as well its trading pattern from the past to accomplish this forecasting task. In order to distinguish between different trading patters and tease out differences, it uses functional shape analysis combined with variable selection.

5. REFERENCES

Berg, J. E. and Rietz, T. A. (2003). Information markets as decision support systems. *Information Systems Frontiers*, 5(1):79–93.

Eliashberg, J., J. J.-J., Sawhney, M., and Wierenga, B. (2000). Moviemod: An implementable decision support system for pre-release market evaluation of motion pictures. *Marketing Science*, 19(3):226–243.

Forsythe, R., Rietz, T. A., and Ross, T. W. (1999). Wishes, expectations, and actions: A survey on price formation in election stock markets. *Journal of Economic Behavior & Organization*, 39:83–110.

Foutz, N. Z. and Kadiyali, V. (2004). Evolution of preannouncements and their impact on new product release timing: Evidence from the u.s. motion picture industry. *Working Paper, University of Maryland.*

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statistical Science*, 21:155–166.

Jank, W., Shmueli, G., and Wang, S. (2006). Dynamic, real-time forecasting of online auctions via functional models. *Proc. 12th ACM SIGKDD (Philadelphia, PA)*, pages 580–585.

Krider, R. E. and Weinberg, C. B. (1998). Competitive dynamics and the introduction of new products: The motion picture timing game. *Journal of Marketing Research*, 35(1):1–15.

Leigh, A. and Wolfers, J. (2006). Competing approaches to forecasting elections: Economic models, opinion polling and information markets. *Economic Record*, 82:325–337.

Pennock, D. M., Lawrence, S., Giles, C. L., and Nielsen, F. A. (2001a). The real power of artificial markets. *Science*, 291(5506):987–988.

Pennock, D. M., Nielsen, F. A., and Giles, C. L. (2001b). Extracting collective probabilistic forecasts from web games. pages 174–183.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis (Second Ed.).* Springer-Verlag, New York.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press, Cambridge.

Shugan, S. M. (1998). Forecasting failure and scucess of new films. *Working Paper, University of Florida.*

Spann, M. and Skiera, B. (2003). Internet-based virtual stock markets for business forecasting. *Management Science*, 49(10):1310–1326.

Wang, S., Jank, W., and Shmueli, G. (2007). Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business and Economic Statistics.* (in press).

Acknowledgements

The authors thank Alex Costakis and Amy Lamare at Hollywood Stock Exchange, Michele Sitkovetsky and Nick Sanzo at TNS Media Intelligence, and Michael A. Marcell at Nielsen EDI for generously providing managerial insights and the data used in this study. We also thank two anonymous referees for valuable suggestions.