Using P2P sharing activity to improve business decision making: proof of concept for estimating product life-cycle

Sudip Bhattacharjee , Ram Gopal , Kaveepan Lertwachara ,
James R. Marsden

Abstract

Estimating the life-cycle or duration of a product can be an important input into a firm's decision-making related to production and marketing. In the music industry, online Peer-to-Peer (P2P) networks have attracted millions of potential music consumers and have had substantial impact on the music business. In this paper, we investigate the possible use of P2P information in estimating product "shelf-life," in particular the duration of a music album on the Billboard 100 chart. We identify and track the music albums that appear on the Top 100 of the Billboard Charts, spanning a period of six months. We show that P2P sharing activity can be used to help predict the subsequent market performance of a music album.

1. Introduction

Estimating the life-cycle or duration of a product can be an important input into a firm's decision-making related to production and marketing. This is of special significance for entertainment firms that deal with multiple products with short life-cycles, such as music and movies. In standard analysis of such issues, researchers have had little difficulty structuring the objective functions – businesses are profit maximizing and can manage their interaction with consumers. And then along came free peer-to-peer (P2P) networks that provided distinctly "non-commercial" means of exchange. In fact, recent studies have observed high level of free-riding among users in these networks [10,14]. A recent article in *Fortune* highlighted the non-business approach of KaZaA and its originators, Janus Friis and Niklas Zennstrom, that has so greatly impacted entertainment companies. The KaZaA developers apparently proceeded without a business plan, taking a "just go and do it approach" with subsequent failed attempts to work out licensing deals with major entertainment companies. KaZaA, despite being a virtual non-business, has become the "top search term on Yahoo" [Fortune, 2004] and, together with its P2P counterparts such as WinMx and Grokster, continues to vex the entertainment industry. But we suggest that businesses might actually leverage P2P networks as information sources to make better production and marketing decisions. In the material that follows, we present our initial investigations into the gathering and potential industry use of P2P activity information.

P2P application software such as KaZaA and WinMX are extremely popular and commonplace among potential music purchasers. As explained later, we developed a custom software application that directly observes and takes "snapshots" of P2P music file-sharing activity. Using this data, we are able to directly address whether measures of such sharing activity might be useful in business decision making. In the work presented here, we focus on estimating product "shelf-life," in particular the duration of a music album that appears on the Billboard 100 chart. Could entertainment firms utilize information on P2P file sharing to better determine the success (measured in chart duration) of music albums?

Before discussing data collection or data analysis, we think it prudent to consider characteristics of the P2P "data venue." In the P2P music sharing setting, incentives are rather different than the normal business/consumer market. Without a profit-maximizing (or similar) objective, what incentive is there for support or customer service tools such as enhanced search tools? In fact, as one would expect, the search options allowed on P2P networks are quite limited. In such environments, searches for digital goods such as music are primarily directed searches. That is, a consumer gathers information from offline as well as online sources about music items (album and artist names) and then searches for a particular item to sample. Certainly, some searches may be broader (e.g., based on music genres), but information overload is likely to quickly occur since there is no support for search aids such as "ranked list of relevant results" in these networks since there are no incentive to provide such services. This deters a "browsing based" search behavior on a particular genre or artist, for example, since a random presentation of search results quickly increase the search cost for a consumer. Hence, the directed nature of the search implies that consumers are more likely to sample those music items for which "information availability" is high (from offline and online sources). This indicates a generally higher sampling and sharing activity for well-known albums that also likely have higher sales based on consumers' information awareness.

With these two factors (most searches are likely to be directed searches and heavily influenced by information availability), we begin our discussion of data collection followed by presentation of initial P2P data analysis. It has been observed that while radio airplay measures the advertising effort for given music albums [11], airplay does not closely predict consumer interest in such albums [5,12]. In fact, anecdotal evidence points to misjudgment of consumer interest and related promotional activities of new artists and albums by record companies (see, for example [16,18]. Given the increasing interest in research on products such as music [2,3,6–9,13], we posit that sharing information on P2P networks may be used to predict consumer interest and subsequent sales for music albums. Our proof of concept approach to investigating the possible value of P2P information in business decision begins with consideration of the following three research questions related to albums that appear on the Billboard charts:

(i) Can sharing information on online networks during initial weeks on chart be a valid predictor for survival duration on the charts?

(ii) Does such early sharing information offer predictive ability beyond such factors as debut rank on the charts?

(iii) Finally, is there any relationship between the predictive ability of early sharing information and album visibility or "album information availability"?

We also provide a cursory "proof of concept" relating to whether continuing P2P sharing might be helpful to business in determining the life-cycle for albums that have already survived for some period of time on the Billboard chart. The issue is posed in the following question:

iv) For albums continuing on the chart, can subsequent sharing activity levels help predict how much longer an album will remain on the charts?

2. The data

2.1. Weekly billboard charts

Billboard magazine, an authority in the music business, publishes various charts of music albums and singles in different genres. In the music business, traditionally sales of certain singles preceded its album sales, creating a promotional effect. However, singles sales has been a loss-making proposition for many albums [1], and the proportion of singles shipped (to albums shipped) has decreased dramatically – 7% (1999), 4% (2000), to 1% (2003). Hence we focus on album sales in this research, since 99% of albums debut without a corresponding authorized early pubic sales of some of their singles.

To study the market impact of P2P activity on albums, we use ordinal sales information from the Billboard Top 200 charts (www.billboard.com/bb/charts/bb200.jsp). Ranking on this chart is based on sales estimates as reported by Nielsen SoundScan using results from their national sampling of retail stores. The official Billboard web site (mentioned above) provides a list of the weekly Top 100 albums at no charge. Since this information was freely available to consumers on the web, we decided to focus on this set of albums rather than the Billboard 200 list for which consumers would have to pay. That is, we focus on the set of ranked albums P2P users might easily and freely access on the web.

This ranking information from the chart is also widely used in the music industry to identify the "winners" in the previous week's retail sales. Specifically, the Top 20 from the Billboard 200 is distributed to subscribers through an alert system before public release. Inclusion of an album in the Top 20 list has been traditionally viewed in the industry as a major marketing success for that album.

We developed a Windows-based application to directly extract the Billboard 100 information each week and to create a list of keywords based on artist and album names. As new albums are released and begin to appear on the charts, the keyword list is updated. As explained next, this keyword list plays a vital role in obtaining the requisite data on sharing activity for each item appearing on the Billboard chart.

2.2. Sharing activity data

In order to complete our analysis, we need data on file sharing behavior. To gather this data, we developed a Windows-based application program to automatically search for the relevant audio files available via WinMX. We chose WinMX over KaZaA because search results on KaZaA have a hard upper limit, while that is not the case for WinMX. Thus, using KaZaA might result in significant understatement of the level of sharing activity. Though appearing on the scene after KaZaA, WinMX is a fast-closing second in current popularity [17].

Using the keywords obtained from our Billboard data collection outlined above, our WinMX search program conducts a Boolean search for copies of songs on each music album available on the network. A data snapshot is recorded and then automatically stored for processing and entry into a relational database. The data gathering process is fully automated. Using a set of dedicated PC's, search start times/key word pairs are randomly generated for each new search. The data utilized here was collected from searches conducted daily over the period October 25, 2002–April 12, 2003 for each of the Top 100 albums for the applicable weekly Billboard chart. Each new Billboard chart signifies a new week of data collection. Our activities involved observing and recording activities that individuals placed into

Table 1
Correlation between dependent and independent variables

	Chart life	Debut rank	Sharing activity during debut week
Chart life	1		
Debut rank	−0.5680	1	
Sharing activity during debut week	0.3980	−0.3605	1

the public domain. We captured only the file information described above and did not perform any downloading of any copyrighted content from any computer on a WinMX network.

3. The analysis

Our data collection provided information on 210 albums that debuted on the Billboard Top 100 chart during the October 25, 2002–April 12, 2003 period. While the data collection snapshots included a variety of technical information (file size, bit rate, etc.), our data requirements focused on measures of chart rankings, sharing activity, and "information availability." The first two are provided directly by the data gathering processes detailed in the previous section. "Sharing activity" is measured as the number of copies of an album that are available on the network in any given period. If more copies of songs from album A is shared in a given period than from album B, we assume that album A has higher sharing activity in that period. For the third, "information availability," we use categories or groupings based on the level at which an album debuts on the chart. For example, albums debuting at a ranking in the Top 20 are rated as having higher information availability than albums debuting in say rankings 21–30 or 31–40. Our measure of information availability can also be construed as product visibility similar to many other rankings such as business schools or sports teams. While we all might be familiar with the top ones, our recollections become much less firm as we move down the ranking. The higher information availability of Top 20 albums is also borne out by the Top 20 album alert system and related marketing efforts mentioned earlier (Section 2.1).

Our "proof of concept" investigation is directed at asking whether P2P sharing activity might prove helpful for firms trying to estimate an album's life-cycle measured in chart duration. [1] Firms already know an album's debut ranking [4]. Our analysis deals with whether P2P sharing activity provides significant additional information. Table 1 provides a simple correlation table for the dependent (chart life) and two non-categorical independent variables (debut rank and P2P sharing activity). We use the third variable discussed above, "information availability," only to group observations for comparison. The values in Table 1 suggest the expected direction of relationship (remember that high initial rankings are indicated by or low initial numerical ranking) between each of the two independent variables and the dependent variable. That is, low numerical values for an album debut and greater sharing activity tend to be associated with longer life-cycle. However, the correlation between the two independent variables is at a level suggesting multi-collinearity should not pose a problem.

In our analysis, we estimate the following regression model:

$$\text{Chart life} = f(\text{rank in week1, sharing activity in week1})$$

We begin by considering groupings representing "high information availability" and "low information availability". Separate regressions were run for albums grouped by debut ranking. Table 2 provides the regression results for albums that debut in spots 1–20 and 21–100, respectively.

[1] While it is possible for an album to drop off the chart and then reappear, this is actually quite rare. Of the total of 210 albums appearing on the Top 100 chart during our sampling period, only nine returned to the chart after falling off.

Table 2
Rank and sharing impact in debut week on chart

		Constant	Rank in week1	Sharing in week1
Model: chart life = f (rank in week1, sharing activity in week1)				
Debut rank 1–20 (Adj. $R^2 = 0.217$, $N = 86$)	Coefficient	12.68	−0.38	0.0012
	t value	7.16^{****}	-2.72^{**}	2.44^{*}
Debut rank 21–100 Adj. $R^2 = 0.178$, $N = 124$	Coefficient	7.80	−0.007	0.0003
	t value	8.73^{****}	-5.097^{****}	0.80
Coefficient difference (z value)		23.53^{****}	20.37^{****}	8.10^{****}

$*$ $p < 0.05$, $**$ $p < 0.01$, $****$ $p < 0.0001$.

In the Top 20 case (Table 2), the P2P sharing activity variable appears to add to any explanatory power provided by the debut rank. In terms of statistical significance, sharing activity was slightly less significant than debut ranking (p-value of 0.017 versus 0.217).

There were a total of 124 albums that debuted on the Billboard chart at rankings from 21 to 100 (Table 2). In this case, sharing activity in week 1 was not significant for any common level of significance (p-value of 0.425) while the debut week rank was still very significant for this group. Thus, for albums with low information availability, the P2P sharing activity does not appear to provide explanatory power for determining album chart life. We also note that an analysis of the coefficient estimates between rank 1–20 and 21–100 show that both "Rank in week 1" and "Sharing in week 1" have significantly different values, along with the difference in the intercept term. This shows that although rank and sharing significantly affect the chart life, the effects are starkly different for the two groups.

Taken together, these results provide at least initial support for positive response to our first three proof-of-concept questions:(i) Can sharing information on online networks during initial weeks on chart be a valid predictor for survival duration on the charts? (ii) Does such early sharing information offer predictive ability beyond such factors as debut rank on the charts? (iii) Finally, is there any relationship between the predictive ability of early sharing information and album visibility or "album information availability"?

Our initial analysis suggests that the answer to the first two questions depends on where an album debuts on the chart. For highly rated albums (Top 20), our preliminary results indicate a "yes" response for the first two questions. For lower rated album debuts (i.e., positions 21–100), the preliminary results indicate a "no" response to the same two questions. Taken together, these outcomes suggest a "yes" response to the third question. Assuming that high debut ranking provides a measure of album visibility or "information availability," the preliminary results support a "yes" response to our third research question.

Our fourth research question focused on whether subsequent sharing activity for albums surviving on the charts might help predict how much longer an album will remain on the chart. We estimated the following model:

$$\begin{aligned}\text{Chart life} = f(&\text{rank}_t, \text{rank}_{t+1}, \text{rank}_{t+2}, \text{rank}_{t+3},\\ &\text{rank}_{t+4}, \text{sharing}_t, \text{sharing}_{t+1}, \text{sharing}_{t+2},\\ &\text{sharing}_{t+3}, \text{sharing}_{t+4}), \quad \text{where}\\ t &= \text{week1(debut week)},\\ t+1 &= \text{week2, etc.}\end{aligned}$$

Our analysis of this question employed a stepwise regression analysis for albums that have appeared on the chart for at least five weeks. Tables 3 and 4 provide the results for albums debuting in the Top 20 and 21–100, respectively.

We note the following, all relating to albums that survived on the chart for at least five weeks:

(i) for albums debuting in the Top 20, Sharing Activity in week 5 enters first, but rank in week 5 also adds explanatory power; and,

(ii) for albums debuting at ranks between 21 and 100, Rank in week 5 enters first with Sharing Activity in week 3 also providing explanatory power.

Table 3
Albums that debuted at ranks 1–20 on Billboard chart

	Constant	Rank in week 5	Sharing in week 5
Dependent variable: chart life or number of weeks on chart			
Coefficient	14.241	−0.0552	0.0013
t value	8.979****	−2.186*	2.894**
(Adj. $R^2 = 0.243$, $N = 63$)			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Table 4
Albums that debuted at ranks 21–100 on Billboard chart

	Constant	Rank in week 5	Sharing in week 3
Dependent variable: chart life or number of weeks on chart			
Coefficient	11.472	−0.05	0.0017
t value	5.417****	−2.701*	2.726*
(Adj. $R^2 = 0.2$, $N = 34$)			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Comparing result (i) with the results detailed earlier (Table 2) provides the suggestion that, in explaining chart life, sharing activity appears to be an important indicator as an album demonstrates survival on the chart (at least our example of five week survivors). Result (ii) suggests a lesser, but still helpful, explanatory role for sharing activity (this time Sharing Activity in week 3 enters the stepwise estimation model) for albums debuting in the Bottom 80 category. We must note, however, that albums debuting in the Bottom 80 drop off the chart relatively quickly so the number of observations here is only 34 compared to 124 in the Table 3 above.

4. Summary and concluding remarks

We began by suggesting that the ability to predict product life-cycle has significant value to the firm. The earlier a firm can estimate product life-cycle, the better since the firm can either avoid continuing costs for short-cycle products or make strategic decisions in support of longer cycle products. The difficulty, however, is that there often is scant early information on which firms can build reasonable estimates.

While the digital good industry does not typically have the same lead time issues as traditional manufactured products, early knowledge on the likelihood of product success and product life-cycle remain important elements in the profit equation. Here, we have provided an initial analysis of new information – P2P sharing activity – which we suggest may be helpful to estimation of product life-cycle of digital goods such as music. We demonstrated the importance of this new information relative to the previously existing information – rankings on the Billboard Top 100 chart.

It is important to stress the preliminary nature of our analysis and results. We did develop a detailed data set on music sharing activity for music appearing on the Billboard Top 100 charts. We did find indications of the relevance of P2P sharing activity. We did provide tentative positive responses to our research questions. But we would argue that our analysis should be viewed as a "proof of concept" demonstration. The results are suggestive, not definitive. The current research is an aid in helping us shape a rigorous and comprehensive research study. This would include a detailed investigation through rigorous estimation of sophisticated models (e.g., hazard models) of the impact of sharing activity on the lifecycle of albums. This also leads to the development of advanced forecasting tools to predict lifecycle from actual consumer activity. These stochastic predictive models can be designed and focused for other digital goods such as digitized movies, books and video games.

Our results also provide tantalizing insights on how the design of P2P networks can be enhanced, to the benefit of both consumers and music companies. The fact that sharing activity is a good predictor only for albums with high information availability speaks directly to the poor design of current P2P systems. As such, these applications function well only as repositories of music; repositories that are useful to consumers only if they know what they are looking for. Finally, we note that new and unknown artists and record companies can actively enter the P2P arena and develop enhanced "search and find" functionalities (e.g. ranked lists based on sharing characteristics) that help consumers explore and experience new music products – possibly through emerging fee-based P2P services. Design of such mechanisms remain an active and fruitful area of research in P2P systems.

Acknowledgements

The authors gratefully acknowledge the use of the Gladstein Endowed MIS Research Lab and the financial support of Treibick Electronic Commerce Initiative, the GE Endowed Professor Fund and the Shenkman Family Endowed Chair fund which helped make this research possible.

References

[1] Bernstein Research, The music industry and the Internet: learning to live the single life, December 8, 2000.

[2] S. Bhattacharjee, R.D. Gopal, K. Lertwachara, J.R. Marsden, Whatever happened to payola? An empirical analysis of online music sharing, Decis. Support Syst. (forthcoming, 2004).

[3] S. Bhattacharjee, R.D. Gopal, G.L. Sanders, Digital music and online sharing: software piracy 2.0?, Commun. ACM 46 (7) (2003) 107–111.

[4] Eric T. Bradlow, Peter S. Fader, A bayesian lifetime model for the hot 100 billboard songs, J. Am. Stat. Assoc. 96 (454) (2001).

[5] BusinessWeek Online. File trading as CD sales predictor?, BusinessWeek online, February 20, 2003. Available from: <interrefhttp://www.businessweek.com/technology/content/feb2003/tc20030220_6958_tc121.htmurlhttp://www.businessweek.com/technology/content/feb2003/tc20030220_6958_tc121.htm>.

[6] M. Givon, V. Mahajan, E. Muller, Software piracy: estimation of lost sales and the impact on software diffusion, J. Marketing 59 (1995) 29–37.

[7] R.D. Gopal, G.L. Sanders, S. Bhattacharjee, M. Agrawal, S. Wagner, A behavioral model of digital music piracy, J. Org. Comp. Elect. Com. (forthcoming, 2004).

[8] Kamel Jedidi, Robert E. Krider, Charles B. Weinberg, Clustering at the movies, Marketing Lett. 9 (4) (1998) 393–405.

[9] Robert E. Krider, Charles B. Weinberg, Competitive dynamics and the introduction of new products: the motion picture timing game, J. Marketing Res. 35 (February) (1998) 1–15.

[10] R. Krishnan, M. Smith, Z. Tang, R. Telang. The virtual commons: why free-riding can be tolerated in file sharing networks? in: International Conference on Information Systems, 2002, Barcelona, Spain.

[11] Wendy W. Moe, Peter S. Fader, Modeling hedonic portfolio products: a joint segmentation analysis of music compact disc sales, J. Marketing Res. 38 (August) (2001) 376–385.

[12] Alan L. Montgomery, Wendy W. Moe, Should Record Companies Pay for Radio Airplay? Investigating the Relationship Between Album Sales and Radio Airplay, Working paper, Carnegie Mellon University, June 2000.

[13] Sonja Radas, Steven M. Shugan, Seasonal marketing and the timing of new product introductions, J. Marketing Res. 35 (August) (1998) 296–315.

[14] K. Ranganathan, M. Ripeanu, A. Sarin, I. Foster. To Share or Not To Share: An Analysis of Incentives to Contribute in Collaborative File-Sharing Environments", Workshop on Economics of Peer-to-Peer Systems, 2003.

[15] Daniel Roth, Pirates of the NET, Fortune 9 (February) (2004) 64–67.

[16] Salon.com. Courtney Love does the math, Technology and Business, June 14, 2000. Available from: <http://dir.salon.com/tech/feature/2000/06/14/love/index.html>.

[17] A. Schatz, Web's Most Wanted 2003, Lycos.com. Available from: <http://50.lycos.com/2003review.asp>.

[18] Brian Steinberg, A CD spins full circle at AOL – a hard-to-peg band named wilco was out – then back in, Wall Street J. 8 (May) (2002) B9–H18.