

# Free vs. For a Fee: The Impact of Information Pricing Strategy on Information Diffusion in Online Social Media

*Research-in-Progress*

## **Hyelim Oh**

Desautels Faculty of Management  
McGill University  
Montreal, Canada H3A 1G5  
hyelim.oh@mail.mcgill.ca

## **Animesh Animesh**

Desautels Faculty of Management  
McGill University  
Montreal, Canada H3A 1G5  
animesh.animesh@mcgill.ca

## **Alain Pinsonneault**

Desautels Faculty of Management  
McGill University  
Montreal, Canada H3A 1G5  
alain.pinsonneault@mcgill.ca

## **Abstract**

*Waking up to the realities of the new Internet age, print newspapers are experimenting with different pricing models for their online news content. Recently, the New York Times (NYT) rolled out a new paywall strategy. As expected, the introduction of the subscription fee model has lowered the NYT's website traffic. However, it is not clear how this new pricing strategy will impact the diffusion of the NYT's content in online social media, which we argue may directly and indirectly impact the future sustainability and growth of the NYT. Therefore, we examine the impact of a paywall rollout lowers the diffusion of the NYT's content in social media. We also find that a paywall affects the patterns of the NYT's content sharing, reshaping to a longer tail in the content sharing distribution as the diffusion extent of popular content decreases. We differentiate the paywall effect between the head and tail of the content popularity distribution in the long tail.*

**Keywords:** Information pricing, willingness-to-pay for content, word-of-mouth, information diffusion, social media, long tail

# Free vs. For a Fee: The Impact of Information Pricing Strategy on Information Diffusion in Online Social Media

*Research-in-Progress*

## Introduction

Exponential growth in the Internet user base and gradual shift in consumer's news consumption preference from offline media (i.e., print) towards online media is transforming the newspaper industry. To address the growing need for online news in the market, newspapers have digitized their news content and made it accessible through a wide variety of devices connected to the Internet. However, given the intense competition in online news market and almost zero marginal cost for providing news online, the newspapers find it difficult to charge a fee for access to their online content. Therefore, most of the online newspapers have been providing content free of charge while making money from online advertising. However, as more consumers switch from print to online news consumption, the online advertising revenue (which is significantly lower than the print ad revenue) is not counterbalancing the loss of revenue from print newspaper subscribers (Peters 2011). Waking up to the realities of new Internet age, print newspapers are experimenting with different pricing models for their online content. However, very few newspapers like Wall Street Journal have been successful with subscription based pricing model (Chyi 2005). Most of the newspaper companies are either sitting at the fence or are trying to make the subscription model work for the online content. New York Times (NYT) is one such newspaper that tried to implement a paywall (i.e., only allowing paying subscribers to access their online content) in 2005 but did not succeed (Ragan 2011). Learning from their prior failed experiment, NYT rolled out a new paywall strategy in March 2011. Since both the subscription and ad revenue are a function of the newspaper's readership, it is important for the NYT to ensure that the traffic to its website does not drop significantly. Recognizing the importance of retaining the current NYT website visitors who might have low willingness to pay, NYT has implemented a generous access policy allowing non-subscribers to read 20 articles for free (Peters 2011). However, just retaining current customers is not enough and, like any other business, a newspaper publisher needs to rely on advertising and word-of mouth (WOM) to increase awareness and to acquire new customers (subscribers as well as non-subscribers).

Recently, content sharing over online social media has become a significant way for people to consume content. As such, the interplay between online social networks and the content business ecosystem has become an important consideration. In recognition of the importance of online word-of-mouth over social media to maintain website traffic, traditional media try to harness WOM to enhance their advertising and subscription revenue stream<sup>1</sup>. When implementing a paywall, the NYT also considered website traffic through social media, allowing free access through social media, such as Twitter and Facebook. Though as expected, the introduction of paid content has lowered the NYT's online readership (also referred as the website traffic) by 5 to 15% in the first 2 weeks after the paywall rollout (Dougherty 2011), it is not clear how this new pricing strategy will impact the diffusion of NYT's content in online social media. Given the importance of online WOM for the future sustainability of online readership of newspaper companies like NYT and lack of understanding of the interaction between information pricing strategy and information diffusion, we examine the impact of paywall rollout on the *extent* and *pattern* of diffusion of NYT's content in online social media such as Twitter.

This paper makes several contributions to the theory and practice. First, this study sheds light on an emerging research topic with regard to diffusion of information content shared by individuals over online social media (e.g., Susarla et al. 2011; Stephen et al. 2011). This study enhances our understanding of how a paywall affects content diffusion. To the best of our knowledge, our work is the first study that

---

<sup>1</sup> According to Alexa (as of March 31<sup>st</sup>, 2011), Twitter and Facebook account for 2.43% and 8.63% of upstream, and 2.23% and 8.58 % of downstream NYT website traffic, respectively.

incorporates the insights of willingness-to-pay (WTP) for online content and information diffusion. Second, we theorize how the impact of a paywall influences content diffusion patterns. Building on the long-tail literature, we theorize about the shift in the concentration of content sharing from the Pareto distribution towards the long tail distribution after the paywall rollout. Third, we provide empirical evidence using objective measures of diffusion outcomes yielded from a natural experiment. Finally, we build a user-content profile model using a finite mixture modeling approach to gain insight into the user-content relationships that change with the introduction of the paywall. Further, our innovative metrics of user and content popularity profiles have a methodological contribution to the literature.

## **Research background and Hypotheses**

In this study we investigate how implementation of a paywall affects the diffusion of the NYT's content on the Twitter network, which has become an incredibly popular social media tool for sharing content. Considering that direct measurement of WOM is difficult or very costly (Godes and Mayzlin 2004; Dellarocas 2003), Twitter provides an interesting research opportunity to study the consumption of online news content and information sharing over online social networks. The openness of the Twitter platform allows us to extract rich WOM information to study the impact of the information pricing policy on online content diffusion.

### ***Paywall effect on diffusion extent***

Research has suggested the difficulty of charging a fee for content because of consumers' low WTP (Picard 2000; Chyi 2005). Given that in social media information spreads through relationships among members of the social network, news content becomes popular within the social media only if there are people to link to them (Stephen et al. 2011). Therefore, fewer users on Twitter who have access to NYT content would mean a lower number of seeds for its NYT content. The loss of visitors to NYT's website will also affect the retransmission of seeds, which facilitates diffusion of information. Therefore, we hypothesize:

*H1: The introduction of a paywall will decrease the diffusion of NYT content over online social media.*

The consumers who discontinued visiting the NYT website are more likely to switch to alternative news sources due to demand elasticity (Chyi 2005). As a result of this substitution, the sharing of free content from rival content providers will increase within the social media. Hence, we hypothesize:

*H2: The introduction of a paywall will increase the diffusion of content consumed from rival media over online social networks.*

### ***Paywall effect on content diffusion patterns:***

A reduction of the extent of diffusion will have a disproportionate impact on the content diffusion patterns as a paywall implementation is likely to affect only low WTP consumers' consumption patterns. The theory of exposure (McPhee 1963) and variety-seeking literature (Simonson 1997) suggest that popular products monopolize the consumption of light consumers, whereas heavy consumers choose a mix of hit and niche products. As a result, the decrease in content sharing due to paywall implementation will occur at "heads" in the distribution of content based on popularity. Therefore, we hypothesize:

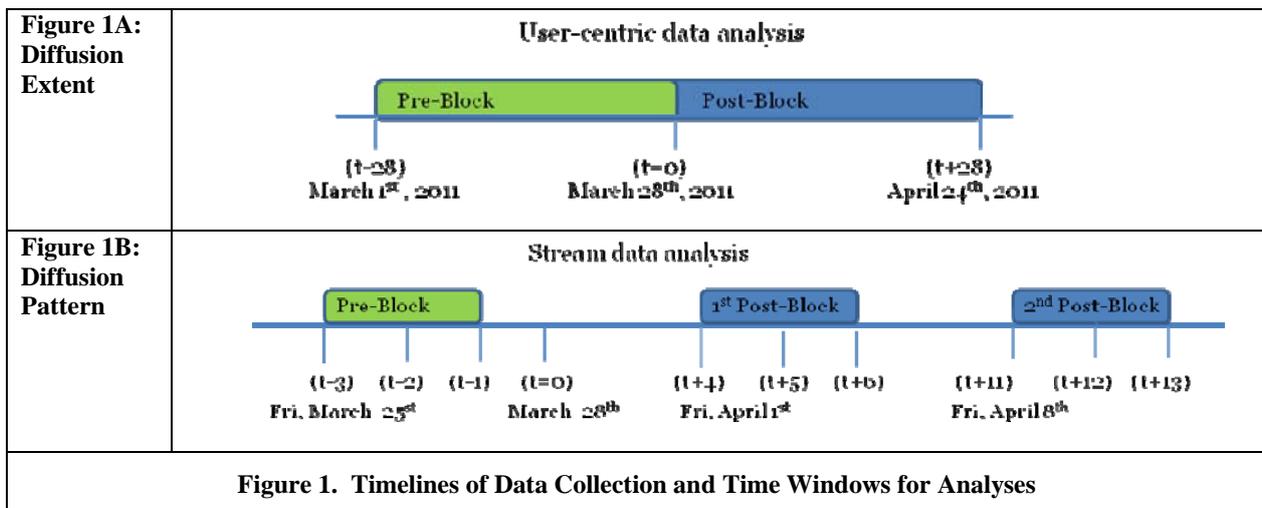
*H3: The introduction of a paywall will decrease the consumption of niche content, which in turn, leads to a longer tail (less concentrated) of content diffusion patterns over online social media.*

Based on the argument in H3, we expect that user activity and content popularity tend to exhibit an inverse relationship. We further predict that the paywall effect will have variation in the popularity of content because light users are more likely to discontinue their contribution to NYT's content diffusion. Thus, we hypothesize:

*H4: The introduction of a paywall will lower the degree of the inverse relationship between content popularity and user activity.*

## Data and Methodology

We collected data using two complementary approaches: user-centric sampling using Twitter Timeline APIs and streaming data sampling using Twitter search APIs. Both data collection approaches have advantages and disadvantages. User-centric sampling allows us to trace back through users' timeliness. However, because of Twitter's rate limiting policy, data collection is costly. On the other hand, streaming data sampling allows us real-time access using search queries. By using Twitter search APIs, we were able to collect a rich data equivalent to the approximate population-size of tweets, which confers sufficient variation to create the long-tail distribution of content sharing. However, given the perishable nature of the snapshot dataset, access to historical data is not available, and our data are limited in the current pre-paywall period. As such, the empirical part in the later sections consists of two empirical approaches: a difference-in-differences analysis using user-centric data and a long-tail analysis using streaming data (see Figure 1).



### *User-centric data*

We collected 700 Twitter users' IDs randomly drawn from our Twitter database that contains 17 million Twitter user IDs and the tweet messages posted during September 2010. This database was collected using a web crawler programs in September 2010. In the random sampling process, we first identified 5,564 Twitter users who had posted New York Times articles at least once from April 2010 to August 2010 using the NYT's custom URL shortener (<http://nyti.ms>) as a search condition, and then 700 Twitter user IDs were randomly sampled from the 5,564 Twitter users. The reason we constrained our sampling to Twitter users who had shared NYT articles previously is that without this sampling restriction, the proportion of users whose tweets contained NYT article links was very small (11 out of 5,500 randomly sampled users). Thus, it is not feasible to collect a sufficient amount of NYT link sharing through a manageable size of Twitter user IDs randomly drawn from the entire 17 million-user database. After data cleansing (e.g., removing celebrities and organization user accounts, and Canadian-location users to avoid a confounding effect due to the earlier paywall rollout date in Canada), our data set contains 626 users' tweets and the data of their profiles from March 1st, 2011 to April 24th, 2011. We further collected rival media's content sharing from the Tweet data via the users who shared NYT content in our sample (see Table 1).

**Table 1A: User-centric Data Summary**

	Before Paywall Rollout	After Paywall Rollout
Tweets which contain NYT links	881	719
Total tweets	159,239	182,022
Proportion of NYT link sharing	0.553%	0.395%
Twitter users	195	199
NYT articles	857	696
Rival media's articles	480	514

**Table 1B: Descriptive Statistics of User-centric Data**

	Pre-Block					Post-Block					Group-wise t-test
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	
NYT Content Sharing <sub>i</sub>	857	1.272	0.027	1	9	696	1.159	0.179	1	5	3.277***
Rival Content Sharing <sub>i</sub>	480	1.072	0.135	1	3	514	1.118	0.0156	1	3	-2.011**

Note: \*\*\* and \*\* represent statistically significant at the 1% and 5% levels, respectively.

### ***Difference-in-differences analysis***

As shown in Figure 1A, we compare the diffusion outcomes for 28 days before and after the paywall rollout. Because the dependent variable is a count of link sharing from either NYT or rival media on Twitter, we estimate the Poisson regression models. *Paywall<sub>i</sub>* is a dummy that equals one if the time period is after the paywall rollout period. By employing the difference-in-differences (DD) estimation strategy, we attempt to remove possible biases in comparing the treatment effect in before and after the paywall rollout that could be the result of time trends. We specify the DD model as follows:

$$\text{Content Sharing}_i \sim \text{Poisson}(\theta_i), \quad \ln(\theta_i) = \beta_0 + \beta_1 \text{Paywall}_i + \beta_2 \text{NYT Content}_i + \beta_3 \text{Paywall}_i \times \text{NYT Content}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

where the time period dummy, *Paywall<sub>i</sub>*, captures aggregate factors that would cause changes in the dependent variable in both treatment and control groups. The dummy variable, *NYT Content<sub>i</sub>*, captures the possible differences between the treatment and control groups prior to the paywall rollout. The coefficient of the interest,  $\beta_3$ , is an estimate of the interaction term, *Paywall<sub>i</sub> × NYT Content<sub>i</sub>*, which equals to one if an observation is in the treatment group in the post-paywall period.

### ***Streaming data***

We collected 541,399 tweets (posted by 137,590 Twitter users), which contain NYT links (“nyti.ms”) from March 25th, 2011 to April 11th, 2011 using the Twitter Search API in a 30-minute interval (1,500 tweets per search query). Then, we parsed the downloaded html pages that contained the tweets and user data. To make a fair comparison regarding the paywall effect on content diffusion, we counted the amount of NYT article link sharing on the same day that the content was published on the NYT website. A significant proportion of samples contained the NYT links to the past articles, which may create positive inflation of the paywall effect as the pre-paywall block simply has a longer duration in our sample. To prevent this biased treatment effect, we used a 1-day diffusion horizon in measuring the dependent variable of diffusion extent. Given the relatively short lifecycle of news content, a 1-day diffusion horizon would be reasonable to test the treatment effect. We also attempted to eliminate the possible weekday effects by creating three comparison blocks composed of the same week days: pre-paywall, 1st post-paywall (1 week after the pre-paywall block), and 2nd post-paywall (2 weeks after the pre-paywall block) (see Figure 1B and Table 2).

Table 2A: Streaming Data Summary			
	Pre-paywall	1 <sup>st</sup> Post-paywall	2 <sup>nd</sup> Post-paywall
Tweets which contain NYT links	19,434	18,918	15,895
NYT articles	1,791	1,543	1,571
Twitter users	38,158	29,506	27,708

Table 2B: Descriptive Statistics of Streaming Data												
	Pre-paywall				1 <sup>st</sup> Post-paywall				2 <sup>nd</sup> Post-paywall			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
NYT content sharing	35.76	160.09	1	3661	35.18	92.07	1	1557	32.22	90.28	1	1626
User activity	2.43	9.24	1	900	2.81	11.36	1	933	2.86	12.81	1	943

### Long-tail analysis

To test H<sub>3</sub> and H<sub>4</sub>, we employ the long-tail measures the extent to which the paywall has an effect on a shift of content sharing distribution. We create a Paywall dummy (defined as 0 if pre-paywall, otherwise 1), and interact the Paywall with  $\ln(\text{Content Popularity Rank})^2$ . Using a log-linear relationship of a power-law distribution that describes the relationship between quantity of the NYT's content sharing on Twitter and content popularity rank (Brynjolfsson et al. 2003; Brynjolfsson et al. 2011), we estimate the Pareto curve, specified as follows:

$$\ln(\text{NYT Content Sharing}_i) = \beta_0 + \beta_1 \ln(\text{Content Popularity Rank}_i) + \beta_2 \text{Paywall}_i + \beta_3 \text{Paywall}_i \times \ln(\text{Content Popularity Rank}_i) + \epsilon_i$$

## Preliminary Findings and Discussion

### Paywall Effect on Diffusion Extent

Overall, the results suggest that the paywall effect leads to a decrease of the quantity of NYT's content sharing on Twitter, implying a subsequent decrease of NYT's website traffic and awareness of the NYT as an influential newspaper (see Table 2A). Interestingly, the interaction term in the DD equation is negative and significant, indicating that rival media has substitution effects on NYT contents after the paywall rollout (see Table 3 and Figure 2).

Table 3. Poisson regressions at content-level <sup>3</sup>			
	NYT Content Sharing	Rival Content Sharing	Pooled Content Sharing
Paywall	-0.093 (.046)**	0.038 (.060)	.038 (.060)
NYT			.170 (.053)***
Paywall × NYT			-.130(.076)*
Observations	1553	994	2547
R-squared	0.0011	0.0002	0.0022
Log-likelihood	-1826.908	-1059.1063	-2886.0142

<sup>2</sup> Consistent with the prior empirical studies, the highest *NYT Content Sharing<sub>i</sub>* is ranked as the lowest value of *Content Popularity Rank<sub>i</sub>*.

<sup>3</sup> Over-dispersion tests of negative binomial regressions confirm that Poisson regression models in Table 2 are appropriate.

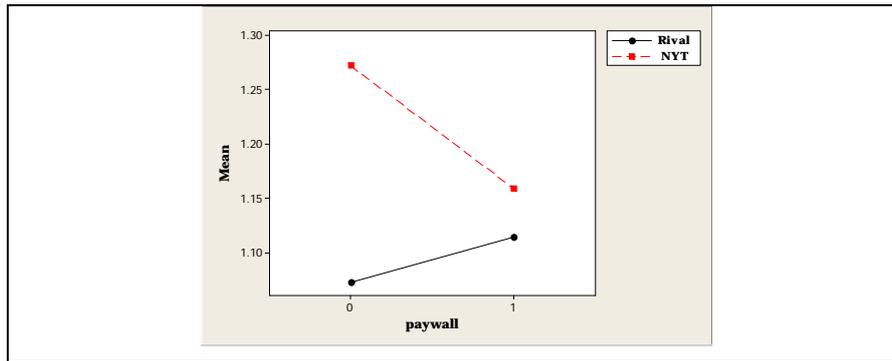


Figure 2. Interaction between Paywall and NYT's vs. Rival's Content Sharing

**Paywall Effect on Content Diffusion Patterns**

We first identify user attrition after the paywall rollout at the user-level analysis. Then, we turn to a content-level analysis using the Gini coefficient metric<sup>4</sup> for the long tail analysis. We find that the Gini coefficients in the pre-paywall periods are consistently greater than the coefficient in the post-paywall periods. The results of the Pareto curve estimation support that the long tail has become longer and flatter after a paywall rollout (see Table 4 and Figure 3).

Table 4: Pareto curve estimation at content-level		
	Pre- vs. 1st post-paywall blocks	Pre- vs. 2nd post-paywall blocks
Constant	10.332 (.069)***	10.332 (.069)***
Rank	-1.557 (.012)***	-1.557 (.012)***
Paywall	-.360 (.099)***	-.311 (.099)**
Paywall × Rank	.071 (.018)***	.039 (.018)**
Adjusted R2	0.893	0.894
Observations	3,334	3,362

Note: Absolute values of t-statistics are in parentheses. \*\*\* and \*\* represent statistically significant at the 1% and 5% levels, respectively.

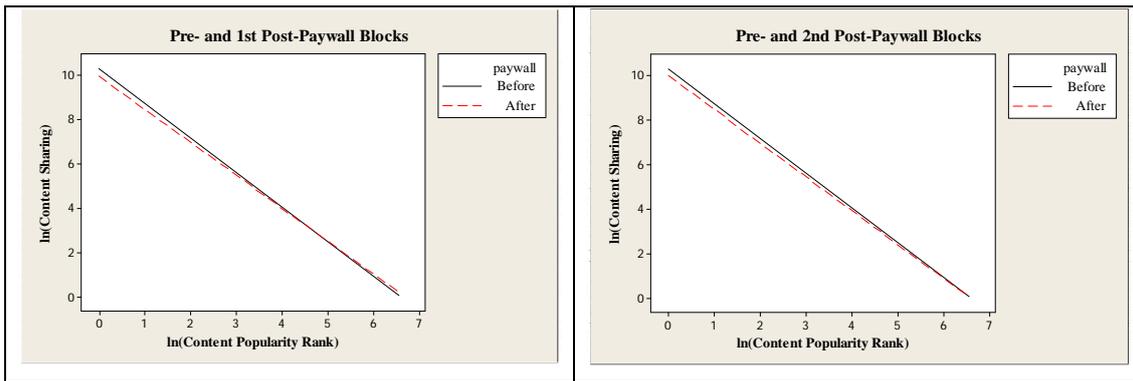


Table 3: Interaction between Paywall and Content Popularity Rank

<sup>4</sup> As the literature has emphasized (Brynjolfsson et al. 2010), the relative measure of the Gini coefficient is only useful for a comparison of long tail distributions when product assortment sizes do not change. We assume that the number of NYT news articles per day is relatively constant, and thus our use of the Gini coefficient at a content-level analysis would be appropriate.

### ***User segmentation (ongoing work)***

The primary motivation of a finite mixture model is to explore how the paywall effect varies with user segments (i.e., heavy vs. light users). Research suggests that finite mixture models provide a flexible method of modeling to capture local variations that a single parametric distribution cannot handle (Bapna et al. 2011; Gupta and Chintagunta 1994). Our preliminary results indicate that user activity and content popularity exhibit an inverse relationship ( $\rho = -.238$ ,  $p < .001$ ). We further examine the dynamics of user segmentation regarding the paywall effect on “heads” and “tails” in the content sharing distribution.

### **Ongoing Research and Conclusion**

Primarily, we find that the introduction of information pricing strategy has impacts on both the quantity and patterns of information diffusion. The NYT is currently introducing new pricing strategies (e.g., plans for iPhone and iPad apps), and thus, comparing our results with the impact of such changes would provide interesting insights. As an extension to this study we are in the process of collecting more data to examine the dynamics of WOM creation processes in the current social media research context.

### **References**

- Banpa, R., Goes, P., Wei, K. K., and Zhang, Z. 2011. “A finite mixture logit model to segment and predict electronic payments system adoption,” *Information Systems Research* (22:1), pp. 118-133.
- Bynjoľfsson, E., Hu, Y., J., and Smith, M. D. 2003. “Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers,” *Management Science* (49:11), pp. 1580-1596.
- Bynjoľfsson, E., Hu, Y., J., and Smith, M. D. 2010. “Long tail vs. superstars: The effect of information technology on product variety and sales concentration patterns,” *Information Systems Research* (21:4), pp. 736-747.
- Brynjoľfsson, E., Hu, Y. J., and Semester, D. 2011. “Goodbye Pareto principle and hello long tail,” *Management Science*, forthcoming.
- Chyi, H. 2005. “Willingness to pay for online news: An empirical study on the viability of the subscription model,” *Journal of Media Economics* (18:2), pp. 131-142.
- Dellarocas, C. 2003. “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms,” *Management Science* (49:10), pp. 1407-1424.
- Dougherty, H. 2001. “Impact of paywall on NYTimes.com,” Available at [http://weblogs.hitwise.com/heather-dougherty/2011/04/impact\\_of\\_paywall\\_on\\_nytimesco\\_1.html](http://weblogs.hitwise.com/heather-dougherty/2011/04/impact_of_paywall_on_nytimesco_1.html).
- Godes, D., and Mayzlin, D. 2004. “Using Online Conversation to Study Word of Mouth Communications,” *Marketing Science* (23:4), pp. 545-560.
- Gupta, S., and Chintagunta, P. 1994. “On using demographic variables to determine segment membership in Logit mixture models,” *Journal of Marketing Research* (31:1), pp. 128-136.
- McPhee, W. 1963. *Formal Theories of Mass Behavior*, Free Press of Glencoe. New York. NY.
- Oestreicher-Singer, G., and Sundararajan, A. “Recommendation networks,” *MIS Quarterly*, forthcoming.
- Peters, J. 2011. “The Times Announces Digital Subscription Plan,” Available at <http://www.nytimes.com/2011/03/18/business/media/18times.html>.
- Picard, R. 2000. “Changing business models of online content services – their implications for multimedia and other content producers,” *International Journal on Media Management* (2:2), pp. 60-80.
- Ragan, S. 2011. “The NYTimes.com paywall causes traffic to drop,” *The Tech Herald*, Available at <http://www.thetechherald.com/article.php/201115/7055/The-NYTimes-com-paywall-causes-traffic-to-drop>.
- Simonson, I. 1997. “The effect of purchase quantity and timing on variety-seeking behavior,” *Journal of Marketing Research* (27:2), pp. 150-162.
- Stephen, A., Dover, Y., and Goldenberg, J. 2011. “Social sharing by social pumps: The effects of transmitter activity in information diffusion over online social networks,” Working Paper, INSEAD.
- Susarla, A., Oh, J., and Tan, Y. 2011. “Social networks and the diffusion of user-generated content: Evidence from YouTube,” *Information Systems Research*, forthcoming.