

Evaluating Promotional Activities in an Online Two-Sided Market of User-Generated Content

Paulo Albuquerque, Polykarpos Pavlidis*, Udi Chatow, Kay-Yut Chen, Zainab Jamal, Kok-Wei Koh†

January 17, 2010

Abstract

We develop a modeling approach to explain demand variation for an online platform of user-generated content, and use it to measure the impact of marketing activities on decisions to visit the platform, and on decisions to create and buy content. The model explains individual-level decisions as a function of consumer characteristics, marketing activities, and behavior of other online users, allowing for the possibility of network effects and interdependence of decisions. Empirically, we apply our model to the Hewlett-Packard’s print-on-demand service of user-created magazines, named MagCloud. We use widely available aggregate-level data from Google Analytics and summary statistics from company reports, thus making our approach generally applicable. Our results show that content price and content creator marketing actions have strong effects on the number of purchases, while advertising by the firm mainly influences visits and creation of content. We provide recommendations to the level of marketing investments and quantify the benefits of “free” promotional activities from content creators. According to our findings, 8% of MagCloud’s profits are directly related to these actions. This type of “free” marketing is likely to have a substantial presence in most online services of user-generated content and must be taken into account when allocating marketing resources.

Keywords: Demand Modeling, User-Generated Content, Two-Sided Markets.

*Paulo Albuquerque is an Assistant Professor of Marketing and Polykarpos Pavlidis is a Ph.D. candidate at the Simon Graduate School of Business, University of Rochester.

†Udi Chatow, Chen Kay-Yut, are principal scientists; Zainab Jamal, and Kok-Wei Koh are researcher scientists at HP Labs.

1 Introduction

The Internet has become one of the most important marketplaces for transaction of goods and services. Recent reports (Comscore, 2007) show that non-travel online consumer spending in the US has surpassed \$100 billion and that growth rates of online demand for information goods, such as books, magazines, and software, are between 25% and 50%. In recent years, online multi-sided markets have emerged with considerable success, led by user-generated content websites such as Facebook, eBay and YouTube. According to eMarketer (2009), 82 million people in the U.S. created online content during 2008, with the majority of uploaded materials being related to social network sites, personal videos, and blogs. The number of user-generated content creators is expected to grow to 114.5 million in 2011.

In these multi-sided markets, a firm/platform usually plays the role of intermediary that maximizes its own objectives by bringing together content creators, consumers and advertisers. In some cases, the platform obtains revenue from commissions derived from transactions of products created by users, while in other cases, revenues come from advertising, such as banner ads or links placed in the users' web pages. The performance of user-generated content platforms is strongly characterized by network effects that emerge between the different participants, especially between creators of content and final users of that content. Network effects have been shown to exist and lead to the success of several multi-sided markets outside of the web (e.g., Rysman 2004, Nair et al. 2004). Given the inter-connectedness and viral community structure of Internet users, it is possible that even stronger network effects persist in the online world. If so, differences between the online and the bricks-and-mortar worlds are likely to be substantial in terms of consumer participation in the market and effectiveness of marketing mix activities on the different groups of individuals involved. Within this framework, we develop a modeling approach to explain demand variation for a platform of user-generated content, and use it to measure the impact of marketing activities on decisions to visit the platform, and on decisions to create and buy content.

The management of marketing activities for an online platform has high level of complexity due to some challenges not usually found in other products or services. First, managers face a two-sided market. A platform that is able to attract a larger number of end users is likely to be more appealing to creators of content. The effect may hold in the opposite direction as well, since increased

quantity and variety of generated content can attract more end users. Second, content users can simultaneously be content creators. It is essential for a manager interested in developing one or more sides of the market to understand potential direct and indirect effects that exist across market sides. Third, there are usually two (or more) stages in the decision to participate in a user-generated platform. Users must first opt to visit the site, and once in the site, they must decide to generate and/or consume the available content. Managers allocate their marketing budgets to influence each stage. Finally, we note that the content creators themselves frequently generate significant “free” marketing for the platform’s content, in the form of referrals and marketing campaigns in the creator websites. These activities, which are in most cases unobserved to managers, should be taken into account when predicting growth or allocating resources.

We specify an individual-level model of multiple decisions by heterogeneous users. Each user chooses either to visit the site, and conditional on visit, to purchase and create content. The decision process is thus naturally divided into two stages. First, consumers choose to visit the site, given expectations about the utility of actions once at the site. Second, consumers decide to purchase content, create content, or both.¹ We start by defining a bivariate probit model for the second stage that allows for correlation between the unobservable terms of the two decisions (creation and purchase) and a structural shift in the utility of purchasing content, if an individual creates content, based on a model proposed by Heckman (1978) on simultaneous equations in the presence of endogenous dummy variables. Each decision is a function of consumer characteristics, marketing activities, and past visitors behavior, allowing for the possibility of network effects. Once in possession of the second stage estimates, we use a binary probit model that takes into account the expectations about the utility of actions at the online platform and marketing activities to explain the visiting decision.

Empirically, we use aggregate-level data from the online service MagCloud, where content is defined as magazines. Created by Hewlett-Packard (HP), the MagCloud platform allows users to buy and sell custom and niche magazines with print-on-demand fulfillment. Each visitor to the site can create and purchase their own magazine, or purchase someone else’s magazines. Our main source of data is Google Analytics. The data set contains daily time series about the number of visits, content purchases and creations, as well as information about marketing actions from both

¹A similar approach of modeling online choice in stages is proposed in Sismeiro and Bucklin (2004).

MagCloud and creators of content. It also tracks the number of new and returning visitors to the site. The data is easy and free to obtain for most managers of online websites.

Besides our modeling contribution, we provide answers to four substantive questions:

1. Is there significant interdependence between the decisions to create and purchase content?
2. How does the effectiveness of different promotional tools vary across platform users, taking into account network externalities and interdependence of decisions?
3. How do profits of the user-generated content platform change with different levels of marketing investments in the two sides of the market?
4. How important is the impact of “free advertising” by content creators to the usage and profits of the platform?

Our results show that the dependence between the creation of content and purchases has multiple dimensions. At the individual level, visitors who have created content are more likely to buy content, both their own² and from other creators. At the aggregate platform level, and as expected, visitors are more prone to purchase content when the total amount of created content is higher.

In terms of promotional activities, we find that the content price and content creator marketing actions have strong effects on the number of purchases, while advertising by MagCloud mainly influences visits and creation of content. Feedback effects are most important within the content buyers side, with past purchasers increasing the likelihood that more buyers reach the site, while similar effects are almost insignificant within the content creators. Our model predicts that price promotions are useful to generate both sales volume and profits, partially due to the feedback effects on the side of the buyers of content. This effect leads to increased sales of content by the platform even after the promotion is over.

Finally, we find that content creators bring a significant number of potential purchasers of content to the platform, which makes the impact of their marketing actions related to MagCloud substantial. Their free marketing activities represent an increase of 8% in the profits of the platform and we suggest that HP should provide additional incentives to content creators to increase their referral activities. The impact of content creator activities is likely to be strong in the development of most

²Creators generate online content and can then purchase printed copies of that content.

user-generated content websites and it must be taken into account by managers when allocating marketing resources.

The paper continues by presenting the relevant literature in section 2. Section 3 describes our modeling approach and the data is presented in section 4. The estimation details are developed in section 5. We analyze the results in section 6 and section 7 concludes.

2 Relevant Literature

We discuss three streams of literature related to our work: research on multi-sided markets, demand models with network effects, and co-production. The study of multi-sided markets is a topic of recent popularity in the economics and empirical industrial organization literature. Early empirical research can be found in Rosse (1979), who looks at the newspaper industry, and Baxter (1983), who focuses on the role of intermediaries in matching two interrelated markets. More recently, Berry and Waldfogel (1999) analyze the market of radio broadcasting, where agents are radio stations, radio listeners, and advertisers, studying whether free entry of radio stations results in market inefficiencies and welfare loss. Gandal, Kende and Rob (2000) measure the feedback loop between the supply of compact disks (CD) and the production of CD players. Rysman (2004) analyzes the relation between advertising and consumer usage in the Yellow Pages industry, and provides welfare implications resulting from the internalization of estimated network externalities. Akerberg and Gowrisankaran (2006) estimate an equilibrium model of consumers and banks adoption of an electronic payment system (ACH), recovering structural estimates of adoption costs and network externalities. Kaiser and Wright (2006) study the multi-sided market in magazines, while Argentesi and Filistrucchi (2007) specify and estimate a two-sided model for newspapers, focusing on the examination of market power in the Italian newspaper industry. Several papers provide an overview of multi-sided markets theory, such as Rochet and Tirole (2005) and Armstrong (2006), while Evans (2003) discusses several issues in applied and anti-trust situations.

Regarding marketing specific issues, research on multi-sided markets is limited, but a few notable exceptions do exist, focusing on indirect network effects. Multi-sided markets imply the existence of indirect network effects, where valuation of a product (or service) by consumers depends on how many consumers use the product, as they attract more sellers of complementary products (Rysman,

2004). For example, Nair et al. (2004) estimate indirect network effects between hardware demand and supplied software variety in the market of PDA's, while Wilbur (2008) proposes a two-sided model to estimate the interplay between TV viewers and advertisers that purchase TV time to promote to viewers. The model estimates allow for the calculation of elasticity measures for the two demand functions and the implementation of a counterfactual experiment assessing the impact of a potential ad avoidance technology. Yao and Mela (2009) study an online two-sided market in the context of auctions, and present a structural model that measures the value of buyers and sellers and provide an empirical analysis of how the two sides should be priced. Parker and Van Alstyne (2005) show that, in two-sided markets, a monopolist could rationally decide to offer a free good in perpetuity in order to profit from a complementary market. Furthermore, they provide conditions which predict which market side is to be subsidized depending on the relative sizes of the cross-market network effects. In common with our paper, all the mentioned empirical studies quantify the existence of network effects and outline methodologies that are useful in doing so. The major differences between our work and the above papers are (1) our focus on a market where consumers are likely to simultaneously participate in both sides of the market (production and purchase of content), and (2) our objective of providing recommendations regarding marketing investments in the different sides of the market, when both the firm and the consumers (content creators) play a role in generating marketing actions.

A second relevant stream of literature includes papers that study demand of new products using individual-level primitives when only aggregate-level data is available. For example, Horsky (1990) models the adoption of durable goods by households that maximize utility with respect to consumption and leisure under their income and time budget constraint. Song and Chintagunta (2003) estimate a choice model for digital camera demand when consumers have heterogeneous preferences and are forward looking with respect to future prices. An interesting feature of these studies is that they estimate models derived at the individual level with aggregate data. Our application shares this modeling perspective, taking aggregate data as the result of heterogeneous individual choices. Additionally, the concept of network effects is also present in demand models of two-sided markets or in tied goods categories. For example, Bell and Song (2007) show the existence of neighborhood effects on the pattern of spatial diffusion of an Internet-based grocery store. In this work, we extend these approaches by modeling a two-stage individual-level decision process, in

an environment with network effects and the additional challenge that multiple consumer decisions are inter-dependent.

Given our interest in researching user-generated content, we find that research on co-production is relevant to our study. Past literature has focused on explaining co-production and how it varies across products and consumers. For instance, Etgar (2008) provides three drivers of co-production: economic, psychological, and social. Among the social motives, it is important to mention the desire of users to create social contact values, i.e., enjoyment in sharing activities with people with the same interests (Berthon and John, 2006), which is related to the social networking and feedback effects present in multi-sided markets. An additional advantage of co-production, usually for the firm providing the platform, is that it helps fragment the market, facilitates the development of a one-to-one marketing operation, and provides an expansion of choices to consumers. Thus, it is in many levels related to product customization. Variability in adoption of co-production is justified in most cases by heterogeneity in user availability of resources or ability to participate in co-production, as well as different opportunity cost of time, the main resource used in co-production. (Etgar, 2008). These skills to co-produce are also likely to evolve with experience (Prahalad and Ramaswamy, 2004), which therefore leads more experienced users to be more likely to co-produce. Based on this literature, we allow for correlation in the decisions to produce and consume content, as well as differences in consumer responses based on past experience with the platform.

Finally, since our final objective is to provide some input for managerial decisions of an online service, we briefly refer to literature on online behavior and marketing resource allocation. The closest research includes two papers by Moe and Fader (2004a, 2004b), where the authors analyze the evolution of online browsing and purchasing behavior as a function of browsing and purchasing histories, using individual level data. They find that purchasing propensities change with the frequency of visits and illustrate the need for a segmented structure of Internet users. We demonstrate with our empirical application that aggregate data from Google Analytics have a level of richness that is sufficient to capture most of the elements of individual behavior included in these papers. While there is some loss of information in our approach when compared with the usage of individual data, our approach is more generally applicable given the easy availability of the web analytics data. Additionally, our data allows us to quantify the effects of content creators marketing activities, which would be unobservable using individual level data from the site navigation records.

Finally, in terms of optimizing the effect of promotions, Gupta and Steenburgh (2008) provide a general framework for the problem of allocating marketing resources, while Mantrala (2006) gives a thorough review of literature, contributions and interesting questions on the allocation of marketing resources in the brick-and-mortar world. Zhang and Krishnamurthi (2004) suggest a method to individualize promotion timing for an online grocery retailer. Our approach is similar, using predicted purchase probabilities to optimize future marketing activities. However, we expand this analysis to include the two-sided nature of our market and the interaction between user decisions that characterize it.

3 The Model

We develop a demand model for consumers interacting with an online firm. Consumers choose to engage in the production and/or consumption of content, while the firm serves as a platform where content is made publicly available for viewing or purchase. Consumer utility is maximized with decisions regarding the visit of the online platform, and subsequently the creation and consumption of content. In our managerial applications section, we show how the firm can take the consumer behavior into account to make appropriate marketing investments that improve its profits.

We model the decision process of a consumer in two stages, involving three separate decisions. First, they choose whether to visit the platform or not, where they can produce and/or consume content. At that stage, the utility of the platform to consumers is driven by consumer expectations about the utility of their actions if they choose to use the platform. In case of a visit, they then face two choices: (1) whether to purchase content; and (2) whether to produce content. Matching the steps in our estimation, we start by modeling the second stage decisions of purchasing and creating content, and then move to the initial platform visit decision.

3.1 Consumer: Consumption and Production of Content

At each time period $t = 1, \dots, T$, each consumer has the possibility of four choices once in contact with the platform: just browsing the site, purchasing content, creating content, and creating and

purchasing content. The utility of purchasing content for individual i at time t is defined as

$$u_{1it} = \alpha_{1i}X_{1t} - \beta_{1i}p_t + \gamma d_{2it} + \delta D_{1t-1} + \lambda_1 D_{2t-1} + \xi_{1i} + \epsilon_{1it}, \quad (1)$$

while the utility for individual i of creating content at time t is given by

$$u_{2it} = \alpha_{2i}X_{2t} + \beta_{2i}(p_t - c_t) + \lambda_2 D_{2t-1} + \xi_{2i} + \epsilon_{2it}, \quad (2)$$

where X_{1t} and X_{2t} are variables that affect the utility of purchasing and creating content respectively, including marketing variables such as promotional activities or advertising of the platform. p_t is the average price for the content sold at the platform at period t , while c_t represents the fee charged by the platform when a content transaction occurs. Consumers purchasing content are charged p_t and so we expect a negative effect of price on their utility. We assume that individual content creators are small compared to the market and hence we do not explicitly model pricing decisions, using price as an exogenous variable that affects utilities. On the content provision side, we expect content creators' utility increases as a function of the mark-up $p_t - c_t$.³

Users compare u_{1it} and u_{2it} with the utility of the outside alternative of not doing any of these options and just browsing the site, which is normalized to zero, leading to consumer decisions d_{1it} and d_{2it} defined as follows:

$$\begin{aligned} d_{1it} &= 1 && \text{if } u_{1it} > 0, && d_{1it} = 0 && \text{otherwise} \\ d_{2it} &= 1 && \text{if } u_{2it} > 0, && d_{2it} = 0 && \text{otherwise} \end{aligned} \quad (3)$$

We allow for a structural shift in the utility of purchasing content, which is operationalized with the inclusion of the content creation decision d_{2it} in the utility u_{1it} of content purchase. This approach has been proposed by Heckman (1978), in his simultaneous equations paper in the presence of endogenous dummy variables. For identification of the parameters in Equations 1 and 2, we restrict our model to have only one structural shift (Maddala, 1983). We choose to include the impact of content creation in the decision to purchase content, since a user of the platform who creates content

³An alternative formulation is to model content creators as profit maximizers, instead of utility maximizers. In the case of our application, through surveys, we found that the main reason for users to upload content was not to obtain profits, but instead some other recreational reason. Thus, we chose to model their decision as a result of utility maximization.

is likely to have a positive shift in the utility to purchase content, as his self-generated content will be available for his purchase. In our empirical study, discussions with managers of the service and observations in the data provide evidence that some creators do purchase their self-generated content.

We include network effects from past visitors to current visitors. D_{1t-1} and D_{2t-1} represent cumulative number of purchases or content creations for the previous time periods from τ to $t-1$,

$$\begin{aligned} D_{1t-1} &= \sum_{k=\tau, \dots, t-1} \sum_{\forall i} d_{1ik} \\ D_{2t-1} &= \sum_{k=\tau, \dots, t-1} \sum_{\forall i} d_{2ik} \end{aligned} \quad (4)$$

where D_{1t-1} captures any possible word-of-mouth effects between previous purchasers and users deciding to purchase. D_{2t-1} measures two different effects. In the purchase utility, it quantifies the number of content alternatives available for purchase, which we expect to have positive effects, since consumers have a large set of content to choose from. In the upload utility, it captures simultaneously a possible word-of-mouth effect between creators about the service and a negative competitive effect, since with more content available at the platform, each individual content is less likely to be browsed and purchased by the final users. The coefficient λ_2 measures the net effect of these two forces.⁴

We include consumer heterogeneity in two ways. First, consumers may differ in their reaction to marketing activities, and so the coefficients α_{1i} and α_{2i} can vary across users. Intrinsic preferences for creating and purchasing content also vary across visitors, which we capture using intercepts ξ_{1i} and ξ_{2i} . We include observed heterogeneity and use finite segments,⁵ with the following formulation:

$$\begin{aligned} \alpha_{ji} &= \sum_{s=1, \dots, S} \alpha_{js} I[i \in s], \quad j \in \{1, 2\} \\ \xi_{ji} &= \sum_{s=1, \dots, S} \xi_{js} I[i \in s], \quad j \in \{1, 2\} \end{aligned} \quad (5)$$

⁴Although it is possible to include the effect of past purchases D_{1t-1} on the utility of creating content, in our application the current content creators do not know how many transactions have occurred in the past. Thus, we do not allow for the number of past purchases to influence current content creation. This will of course limit the cross-side effects from buyers to creators. The model can be easily changed to account for this additional relation in cases where the number of purchases is public information (for example, in YouTube.com, the viewership numbers are shown in the respective video-content pages).

⁵Additional heterogeneity can be included in our formulation, for instance, unobserved heterogeneity using random coefficients. However, in our empirical application, we found that our current model fits the observed patterns of the aggregate data very well and any additional heterogeneity parameters would be very hard to be estimated precisely. Thus, we chose the more parsimonious model.

The indicator variable $I[i \in s]$ takes the value of 1 if individual i belongs to segment s and 0 otherwise. We define consumer segments based on two criteria: (1) how users first access the site and (2) based on past actions. Consumers can reach the platform directly, through a search engine, or a referral site. Additionally, users are classified as new and returning visitors.⁶ We believe that our segmentation scheme captures both the level of involvement and experience of consumers with the platform. Consumers that have more interest in using the platform are likely to know the web address, have a direct link saved in their computers, or come from a related site, and thus be in the segment of consumers that reach the site directly or by referrals. Consumers with less interest in the site are likely to come from search engines, when searching for services in the platform industry. Information and experience from past usage is captured by the new and returning visitor heterogeneity.⁷ Our final number of observed segments in the empirical application is 6 ($S = 3$ access segments \times 2 past usage segments = 6).

Finally, we assume that ϵ_{1it} and ϵ_{2it} are independent over time and normally distributed, with mean zero and variance-covariance matrix Σ , where the off-diagonal parameter ρ in Σ represents correlation in the unobserved components of the utility, i.e.

$$\begin{bmatrix} \epsilon_{1it} \\ \epsilon_{2it} \end{bmatrix} \sim N(0, \Sigma), \quad (6)$$

with

$$\Sigma = \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}. \quad (7)$$

Our assumptions lead to a system of equations that forms a multivariate probit model with structural shifts (Heckman, 1978). Given the two decisions, each consumer falls into one of four possible outcomes: neither purchase nor produce content $\{d_{1i} = 0, d_{2i} = 0\}$; purchase content (which has

⁶In the direct access, we include consumers who write the website address in the web browser or have a previously saved link to the site in their computer; users who use a search engine, e.g. google.com, to get to the online service site are classified in the search site segment; finally users that are referred to the online site by a different website are in the last group. In terms of returning visitors, we include visitors that have at least visited the platform once. This classification is done and captured by Google Analytics in our empirical application.

⁷We tested additional segments in term of past actions, by looking at users visits and actions, such as past purchases and content creation, and in previous week and month, instead of the full period. The alternative formulations did not change the results significantly. Classification in repeat content creators vs. new creators can be obtained with minor programming changes done by managers in Google Analytics, if the firm is interested in targeting these consumers.

been previously created by him or other consumers) but not create new content $\{d_{1i} = 1, d_{2i} = 0\}$; create content and not purchase $\{d_{1i} = 0, d_{2i} = 1\}$; and do both actions of purchasing and creating content $\{d_{1i} = 1, d_{2i} = 1\}$. This framework is general enough to cover a wide spectrum of multi-sided platforms where user-generated content is exchanged and it can easily be extended to include more actions within the platform.

Combining Equations 1 and 2, and the assumption of normality of the error distribution, consumer i chooses, for example, to create content but not purchase content $\{d_{1i} = 0, d_{2i} = 1\}$ if

$$\begin{aligned} v_{1it} + \varepsilon_{1it} \leq 0 &\Leftrightarrow \varepsilon_{1it} \leq -v_{1it} \\ v_{2it} + \varepsilon_{2it} \geq 0 &\Leftrightarrow \varepsilon_{2it} \geq -v_{2it} \end{aligned},$$

where v_{1it} and v_{2it} are the deterministic portion of the utility, i.e., $u_{1it} = v_{1it} + \varepsilon_{1it}$ and $u_{2it} = v_{2it} + \varepsilon_{2it}$. The implied probability of consumer i making these choices, conditional on visiting the platform, is given by

$$P(d_{1it} = 0, d_{2it} = 1) = \int_{-\infty}^{-v_{1it}} \int_{-v_{2it}}^{+\infty} \phi(\varepsilon_{1it}, \varepsilon_{2it}, \rho) d\varepsilon_1 d\varepsilon_2, \quad (8)$$

with ϕ representing the bivariate normal probability density function. We obtain similar expressions for the remaining decisions. To obtain the probability of consumer i just browsing the site and not purchasing nor creating content, we use

$$P(d_{1it} = 0, d_{2it} = 0) = \int_{-\infty}^{-v_{1it}} \int_{-\infty}^{-v_{2it}} \phi(\varepsilon_{1it}, \varepsilon_{2it}, \rho) d\varepsilon_1 d\varepsilon_2. \quad (9)$$

At any period t , the fraction of M_t website visitors who will choose one of the four options is given by the aggregation of these probabilities across individuals. For instance, the following expression provides the estimated number of content creators who do not purchase at time t :

$$\widehat{S}(d_{1it} = 0, d_{2it} = 1) = \sum_{i=1}^{M_t} P(d_{1it} = 0, d_{2it} = 1). \quad (10)$$

The expected number of individuals choosing one of the remaining three decisions can be computed using similar equations, including the estimated number of users choosing to browse the site but

not engage in any content-related action:

$$\widehat{S}(d_{1it} = 0, d_{2it} = 0) = \sum_{i=1}^{M_t} P(d_{1it} = 0, d_{2it} = 0). \quad (11)$$

Since there is no closed form formula for the integrals in these expressions, we use simulation to obtain approximations of the integrals. In our data, we observe the actual number of individuals involved in each of the four decisions, denoted for instance by $S(d_{1it} = 0, d_{2it} = 1)$ or $S(d_{1it} = 0, d_{2it} = 0)$. More details about how we obtain the parameters is provided in the estimation section.

3.2 Consumer: Platform Visit

Before having the opportunity to create or purchase content, online users must decide whether to visit the online platform or choose an outside alternative, such as visiting a website that offers similar service to the platform.⁸ To make this choice, users compare the utility of the platform with the utility of the outside good. At time t , users visit the website with expectations of utility from the usage of the platform. We define the utility of visiting the online platform as

$$u_{3it} = \alpha_{3i}X_{3t} + \psi_i E[\max(u_{1it}, 0)] + \omega_i E[\max(u_{2it}, 0)] + \xi_{3i} + \varepsilon_{3it}. \quad (12)$$

The vector X_{3t} contains exogenous variables that influence the utility of visiting the platform, such as marketing actions by the firm. The terms $E[\max(u_{1it}, 0)]$ and $E[\max(u_{2it}, 0)]$ denote the expected maximum utility over the choices to purchase and create content conditional on visiting the site. Our assumption is that, before visiting the platform, users have correct expectations about the level of utility that they can derive from visiting the platform. By including the utility of future actions when at the platform as a covariate in the utility of visiting the website, we allow for users who are more inclined to purchase and/or upload content to have a higher than average probability of visiting the website. We also account for heterogeneity in these preferences by setting the coefficients ψ_i and ω_i to be individual specific (or more exactly, segment specific in our specification). This in turn helps us obtain a structural representation of the utility of a potential visitor that connects the two decision stages. Additional heterogeneity in intrinsic preferences is captured with intercepts for

⁸In our empirical application, there are no close competitors to the service offered by HP, but visitors can choose to create their magazine and upload it in social network or blogging sites, for instance.

each consumer segment ξ_{3i} . Finally, we assume that the unobserved part of the visiting utility ε_{3it} follows a normal distribution,

$$\varepsilon_{3it} \sim N(0, \sigma_3), \quad (13)$$

which reflects independent shocks which are known to the users when they make the decision to visit the website, but unobserved by the researcher. The variance σ_3^2 is set to one for identification purposes in estimation.

The expectations of utilities from actions must take into account the correlation in the unobserved shocks of the two decisions, as well as the structural shift in the utility of purchasing content, if content is created. We use the following formulation

$$\begin{aligned} & \psi_i E[\max(u_{1it}, 0)] + \omega_i E[\max(u_{2it}, 0)] = \\ & \int \int [\psi_i \max(u_{1ti}, 0 | d_{2ti}) + \omega_i \max(u_{2ti}, 0)] \phi(\epsilon_1, \epsilon_2, \rho) d\epsilon_1 d\epsilon_2 \end{aligned} \quad (14)$$

This expression involves two-dimensional integrals of the bivariate normal distribution and does not have a closed form. We use simulation to compute these expectations, conditional on estimates of purchase and content creation utilities and their correlation coefficient. We describe the necessary steps for this simulation in more detail in the estimation section. Normalizing the utility of the outside alternative to zero for identification purposes, a user decides to visit the platform ($d_{3it} = 1$) if

$$v_{3it} + \varepsilon_{3it} \geq 0 \Leftrightarrow \varepsilon_{3it} \geq -v_{3it}, \quad (15)$$

where v_{3it} defines the determinist part of the utility obtained from visiting the platform. This leads to the following expression for the probability of user i visiting the online platform at time t :

$$P(d_{3it} = 1) = \int_{-v_{3it}}^{+\infty} \phi(\epsilon_{3it}) d\epsilon_3. \quad (16)$$

4 Data

Our empirical application uses data provided by HP, more specifically by their research division Hewlett-Packard Labs (HP Labs). It relates to an online platform created by HP where users can buy and sell custom and niche magazines with print-on-demand fulfillment. According to HP,

“MagCloud offers an innovative alternative to bring consumers and publishers together in a web-based marketplace where choice, flexibility and print-on-demand are the cornerstones of the community.”

The service was launched in June of 2008, and has consistently grown to become one of the most popular online sites to create custom magazines by individual or small publishers. The platform is designed for generation of content and its diffusion online, similar to other websites such as Lulu.com for books, or YouTube.com for videos. Once users access the online platform they are offered the opportunity to browse, create and upload content, usually in the form of a Portable Document Format (PDF) document, and purchase existing content in the form of printed magazines, which are then shipped by HP.⁹

The main source of information about MagCloud purchase and generation of content used in this paper is Google Analytics (GA). GA is a leading online service of website traffic statistics and is provided free of charge to managers of websites by Google.¹⁰ Its output is user-friendly and oriented for managerial usage, especially to measure the performance of website traffic. Any website administrator can register his website with GA and start extracting customized reports, in text or spreadsheet format, with almost real-time website traffic information. The data from GA are collected with first party cookies named page tags. They have the advantage of not being contaminated by bot visits to the website, do not require the identification of Internet Protocol (IP) addresses, and being able to measure visits from the computers cache memory. The data also has some limitations. Since it is collected from page tags and computer cookies, the absolute numbers reported from web analytics may not be completely accurate, although the aggregate numbers are measured with acceptable precision (Clifton 2008). To increase our confidence on the data, we cross-validated the accuracy of the GA data by comparing some of the collected numbers for site actions with internal accounting data that were retained separately in a transactional database. We found a close match in numbers from the two data sets.¹¹

The collected information is at the daily level aggregated over website visitors. This aggregate-level data is result of all website traffic, without any sampling or selection bias. From GA, we collect

⁹For more details, visit www.magcloud.com.

¹⁰The free version has an upper limit of 5 million page views per month. This limit is lifted if the user has an active AdWords account with Google (www.google.com/support/googleanalytics).

¹¹Although we could have used some information of the transactional data set from HP to complement our data, we instead chose to maintain our analysis supported by GA data, so that our approach has a wide applicability by any website manager that decides to freely opt to gather information from GA.

the number of website visitors over time, total and by consumer segments defined by HP, given the limitations of the GA software. For example, we observe the daily number of new and returning visitors, as well as how they accessed the site (i.e., search engine, referring sites, or direct access). Returning visitors are defined as users that have accessed the site at least once in the past. The tracking of website traffic includes the rates of conversion for any user defined goal. In the case of MagCloud, the specified goals are magazine orders and magazine uploads that are transformed into magazine titles ready to be printed and shipped automatically through the platform after purchase. In other words, besides visits, we observe the daily number of content purchases and number of content creations (magazines), for each segment.¹²

Figure 1 shows the time series for our three dependent variables: the daily number of visits, purchases of content, and creations of content. Our data ranges from July 1st, 2008 to October 6th, 2009. We observe almost all history of the site, since MagCloud was made available to general online public in the beginning of June of 2008. The data shows some interesting patterns. We see a continuous growth of the three variables, as the site became more popular. We also note that at the end of March of 2009, the number of site visits presented a large spike. This was driven by an important and free advertising that the website obtained by being featured in the New York Times and described at its online site, with a direct link to magcloud.com. This link remained visible in the online site of the New York Times for a few days, which explains why the spike lasted for more than one day. This event impacted the evolution of the purchases and content creation considerably, and we take this into account in the explanatory variables included in our model.

Figures 2 and 3 display the evolution of content purchases and creations by segment. Visitors that reach the site from referrals, mainly new but also returning visitors, are the main source of content purchases over the observed time periods, followed by users that reach the site by direct access. For content creation, we see a different pattern. Most creators are returning visitors, with the exception of the time periods around the spike caused by the New York Times advertisement. The fact that the majority of creators are users that have previously been in contact with the site is reasonable, since it is likely that most users will need to understand the requirements to create content, which may lead to multiple visits to the platform.

¹²For privacy reasons, actual numbers are masked, but we use a consistent scale so that all effects keep their substantive meaning.

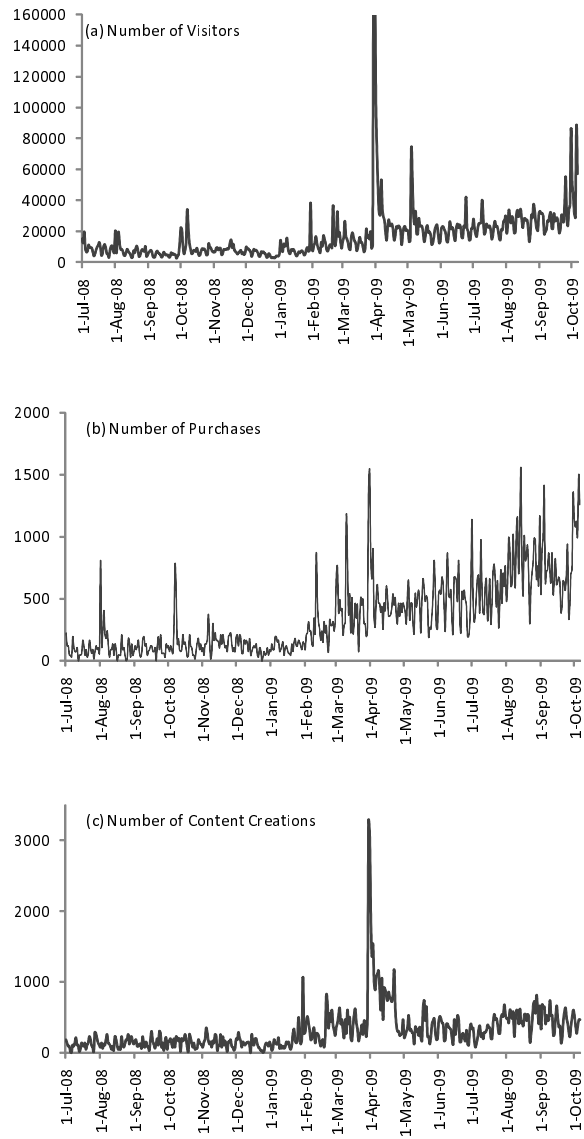


Figure 1: Number of visits, content purchases, and content creations.

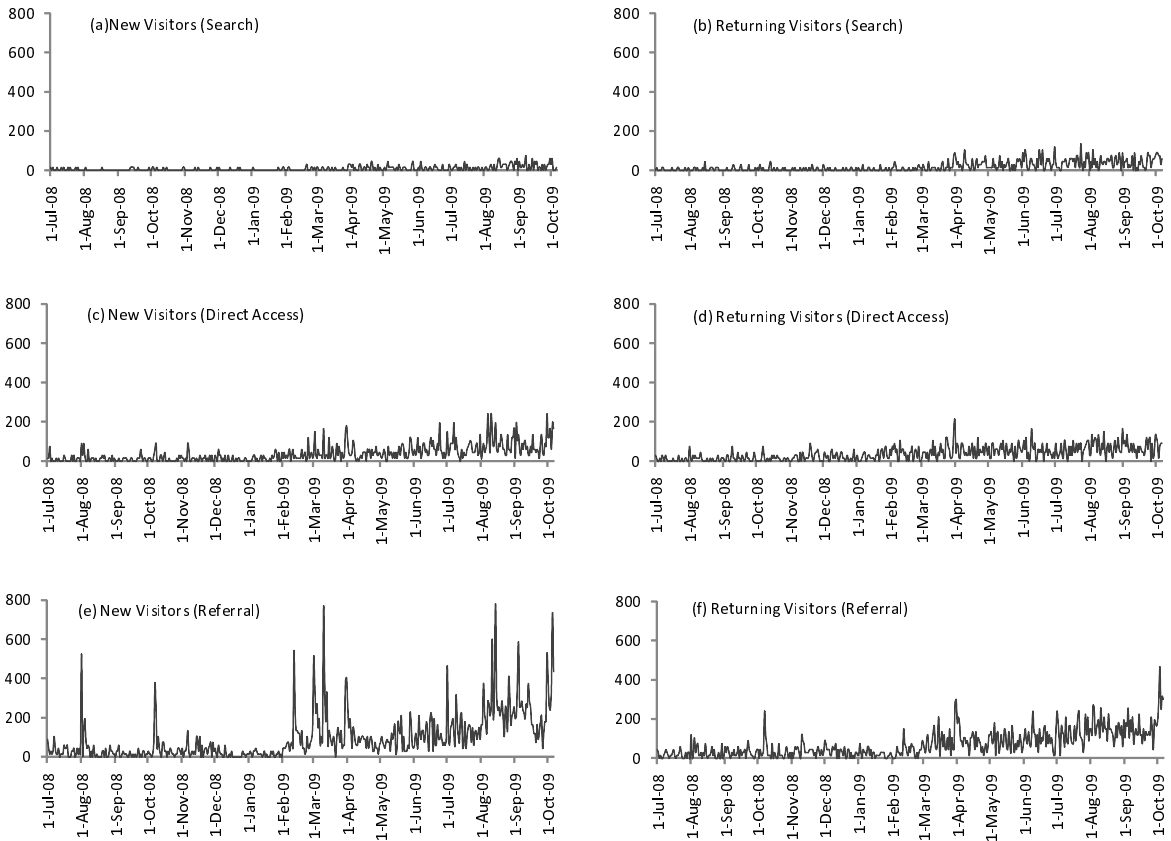


Figure 2: Number of content purchases by segment.

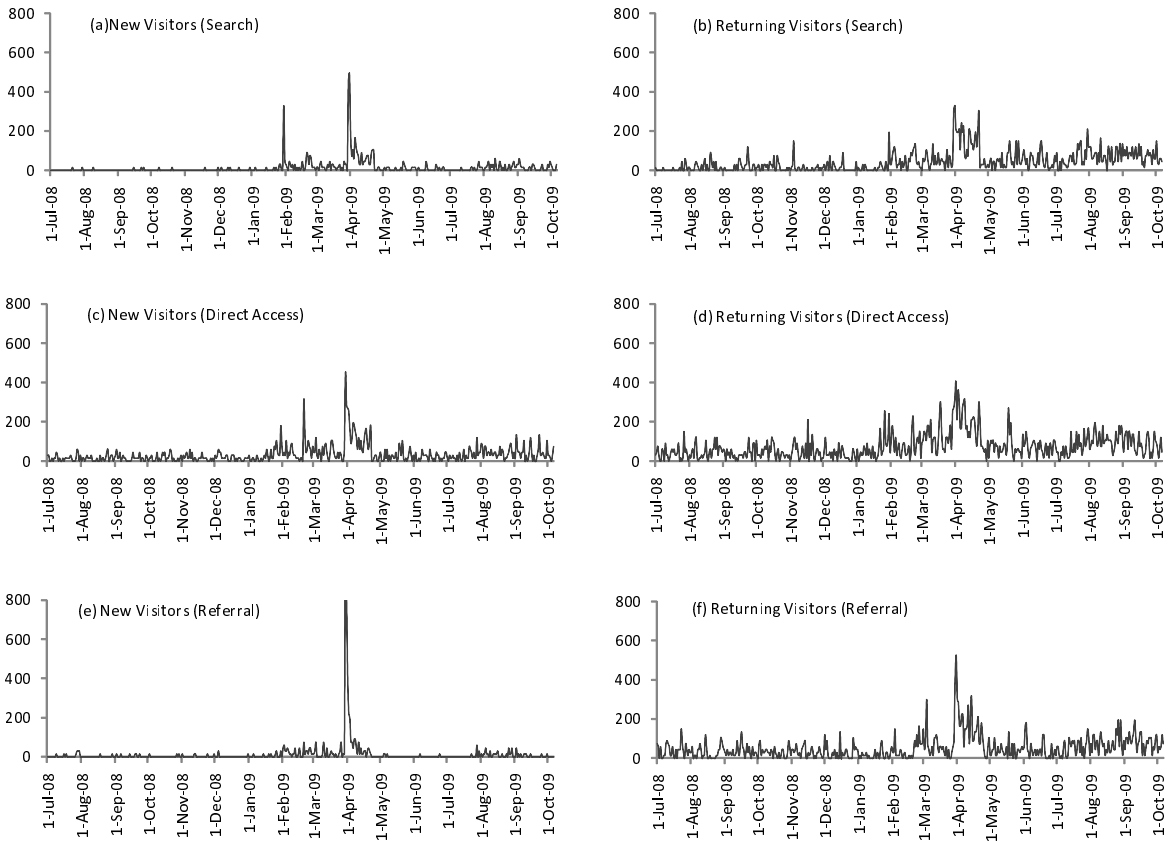


Figure 3: Number of content creations by segment.

We present additional statistics in Table 1, where we describe the number of actions in the site based on the two defined dependent variables for the second stage, which examines the behavior of users conditional on a website visit. As previously described, each consumer makes two decisions, which combined lead to the four outcomes in the table. In brackets, we show the percentage of users in each segment for each outcome. New visitors are responsible for about 69% of all platform visits with a total of 5.3 million, while returning visitors constitute the remaining 39%, with 2.4 million. The conversion rates, from visits to each action is 1.7% for purchase without content creation, 2.1% for creation only, and 0.7% for both content creation and purchase. In general, and as previously shown, returning visitors have higher conversion rates, while a large majority of browsing visits comes from new visitors.

A second data set based on Google information provides us with the records containing marketing activities originated by both the platform firm, in this case HP, and the users that generate content. This type of information is collected through Google Alerts, also a free service from Google, that sends automatic emails with alerts at a pre-specified time interval (daily in this case) about any key terms that the user sets as criteria. Google Alerts notifies the website manager each time that a new web page appears in the top ten or top twenty results from a Google search on the key term.¹³ This allows us to control the frequency of appearance of the term “MagCloud” in blogs, social networks and personal web pages. We manually code the information from Google Alerts in the form of two count variables, “Content Creator Events” and “General Events”. The variable “Content Creator Events” counts the daily number of websites that directly advertise magazines published with MagCloud. The initiators of these events are mainly the creators of the magazines who usually include an active link that generates web traffic to MagCloud, and represent free advertising for the platform. In average, there are 4.8 such events per month. The variable “General Events” counts the daily number of articles/posts that refer to issues like on-demand publishing, magazines, cloud computing, and new web services, and explicitly mention MagCloud. These events are initiated mostly by HP, which makes it a decision variable for managers. We observe a monthly average of 12 marketing actions by HP in our data set.

We also include additional marketing actions related to the platform. As mentioned earlier,

¹³Whether it is on the top ten or on the top twenty results depends on the type of the alert; web alerts check the top twenty results while blog alerts check the top ten. (Source: <http://www.google.com/support/alerts/>)

Segments	Browsing Only	Purchase, No Creation	Creation, No Purchase	Purchase & Creation
New Visitors Search	970,875 (98.9%)	3,630 (0.4%)	6,240 (0.6%)	690 (0.1%)
New Visitors Direct	1,517,370 (97.6%)	17,805 (1.1%)	15,615 (1.0%)	3,780 (0.2%)
New Visitors Referral	2,792,670 (98.2%)	45,315 (1.6%)	6,030 (0.2%)	165 (0.0%)
Returning Visitors Search	493,260 (92.9%)	9,390 (1.8%)	21,000 (4.0%)	9,390 (1.4%)
Returning Visitors Direct	704,655 (91.5%)	18,405 (2.3%)	34,380 (4.5%)	12,870 (1.7%)
Returning Visitors Referral	1,072,335 (94.0%)	33,675 (3.0%)	24,660 (2.2%)	9,600 (0.8%)

Table 1: Descriptive Statistics

MagCloud appeared in the New York Times during two days at the end of March of 2009, which led to a spike in visits at the action. We code this event as a dummy variable for the two days when there was an article online about MagCloud and a link to the website. Additionally, to control for possible long term effects (or a structural break in the utilities) due to this large event, we include a dummy variable for time periods after this activity. We also control for two marketing events that changed the interaction of users with the site: the introduction of a publishing user-friendly software at MagCloud and the possibility of orders from outside of the U.S.. Both are coded as dummy variables for time periods after these two actions occurred.

Finally, we complete our data with additionally information collected directly from HP. First, we have data on the price of content. There are three prices of importance related to the platform performance. First, HP charges a base price per page printed, usually \$.20. Second, the content creators can set their own markup per magazine. Third, each magazine sells at a price of \$.20 times the number of pages in the magazine, plus the markup. Through most of time in our data set, HP did not change the price charged per printed copy, keeping the \$.20 as an everyday price, except in September of 2009, when it offered a promotional discount of 20%. Across all periods, average mark-up per page is about \$.08, which leads to an average final price per page of about \$.28 to content buyers. We observe a higher variance at the beginning of the time series for the price, since less content was available at that time. As more content is presented in the platform, prices and mark-up tend to stabilize around the \$.25 and \$.5 respectively. Finally, we have also information about the potential market for the MagCloud platform by source given research studies done before the introduction of the service. HP predicts that 15 million magazine interested users are potential targets for the print-on-demand service. We use this number as the total market potential for the platform, and assume it to be constant for the time periods in our data. Our estimation ties all these sources of data together and provides a unified setting for inference and prediction.

5 Estimation

Our estimation has two stages: obtaining the parameters related to the purchase and creation of content, and estimating the parameters related to the decision to visit the online platform. To break down the estimation into these stages, we assume that the parameters in the utility of content purchase and creation do not depend on those from the visits model, and that conditional on the data and expectations about the maximum utility that a visit can offer, the unobserved components of the visiting utility ε_{3it} are not correlated with the unobserved parts ε_{1it} and ε_{2it} in the content purchase and creation utility.

5.1 Purchase and Creation of Content Stage

We start by estimating the parameters that relate to the decisions of creating and purchasing content. According to our previous description, our model has the form of a bivariate probit with structural shift. Using aggregate daily data on visits, content creations, and purchases by (observed) segment as dependent variables, we obtain estimates by maximizing the log of the following likelihood function:

$$L = \prod_{t=1}^T \prod_{i=1}^N L_{it}. \quad (17)$$

The individual likelihood L_{it} is based on data and the probabilities of each pair of actions presented in the modeling section:

$$L_{it} = P(d_{1it} = 0, d_{2it} = 0)^{I(d_{1it}=0, d_{2it}=0)} \times P(d_{1it} = 1, d_{2it} = 1)^{I(d_{1it}=1, d_{2it}=1)} \\ \times P(d_{1it} = 1, d_{2it} = 0)^{I(d_{1it}=1, d_{2it}=0)} \times P(d_{1it} = 0, d_{2it} = 1)^{I(d_{1it}=0, d_{2it}=1)}, \quad (18)$$

where $I(d_{1it} = 0, d_{2it} = 0)$ is an indicator function if individual i chooses not to create nor purchase content, and similarly for all other alternative actions. When using aggregate-level data and in the case when only observed heterogeneity is included using a finite number of discrete segments, such as in our empirical application, all individuals belonging to the same segment display the same deterministic utility v_{1it} and v_{2it} , implying that the probabilities of actions are equal for all individuals i of segment s . This simplifies the estimation considerably, since we need only to compute $S \times 4$ (S segments, 4 outcomes) different likelihood values for each time period instead of

$I \times 4$, and exponentiate each segment and outcome probability to the respective observed number of individuals to obtain the final likelihood expression.

The probabilities in Equation 18 are given by the expressions in Equations 8 and 9 (and by similar equations for the other pairs of decisions), which do not have a closed form. In our estimation routine, we use a simulator in Genz and Bretz (2009) and Gentz et al. (2009), which provides approximations of integrals from the normal distribution and has been shown to perform well in Monte Carlo simulations.

5.2 Visiting Stage

Given the estimates of the content purchase and creation stage, we can compute the expected maximum utility of potential website visitors and use it along with additional explanatory variables to get estimates of the visiting utility function in Equation 12. Since the unobserved shocks of the utilities in the previously estimated stage follow the bivariate normal distribution, we use Monte Carlo integration to approximate the expectations of visitors for their on-site actions. For each period, we use the parameter estimates and explanatory data of the second stage model to compute the deterministic part of the upload and purchase utilities. We then draw unobserved shocks from a bivariate normal distribution with the estimated correlation coefficient and compute the maximum utility for each second stage decision.¹⁴ We repeat R times and average the results to create the expectations.

After we obtain the expected utility from the second stage, we estimate the first stage decision as a single probit equation, using the following likelihood function:

$$L_V = \prod_{t=1}^T \prod_{i=1}^N L_{Vit}, \quad (19)$$

with

$$L_{Vit} = P(d_{3it} = 1)^{I(d_{3it}=1)} \times P(d_{3it} = 0)^{I(d_{3it}=0)}. \quad (20)$$

As in the previous stage, with aggregate level data and observed heterogeneity, we do not observe I decisions. However, we know and use the observed number of visits and non-visits at time t by

¹⁴Once in possession of the unobserved draws, the utility and a simulated decision to create content is obtained before computing the utility of purchase, due to the structural shift in the utility of purchase.

segment, which are a sum of $I(d_{3it} = 1)$ and $I(d_{3it} = 0)$ across individuals, to obtain the correct number of individual likelihoods in each segment and outcome, which can then be combined with the probability expressions to obtain L_{Vit} .

To obtain correct standard errors that account for simulation error and error from the estimation of the creation and purchase stage parameters, we use the bootstrap technique. For a number of bootstrap iterations B , we repeat the Monte Carlo integration for each stage with different draws of unobserved components and of the deterministic part of the utilities. We use the bootstrap samples to obtain a series of parameter estimates, which we then use to compute standard errors. In our implementation, we use $R = 1000$ in the expectations integration and $B = 200$ for the bootstrap.

6 Results

We start our discussion of the results by presenting some performance measures of our model. We then analyze the parameter estimates regarding the stage of content purchase and creation, and follow by analyzing the results regarding the visit decisions. Lastly, we discuss a number of managerial applications from our modeling approach.

6.1 Model Fit and Hold-out Measures

To evaluate the fit of the proposed model, we compute the estimated number of visits, purchases, and uploads. Figure 4 displays the actual and estimated values for actions at the online site, for the time periods in our data set divided by three consumer segments. We see that the model does a good job explaining the variation of visits, content creation and purchase. It is particularly interesting to see that the model is able to capture the spikes in content and purchases, which coincide with marketing actions from HP and from the creators of content advertising their magazines.

Additionally, we evaluate the model's ability to predict future consumer decisions. We use 60 observations after the last time period included in our estimation, from October 7th to December 5th. We use the real value for marketing decisions that are in control of the firm, and take draws from the empirical distributions of any of the other variables. The variables quantifying network effects, such as the number of content purchases in the previous week, are obtained using realizations of the choice probabilities. As an illustration, the predicted and actual content purchases and creations

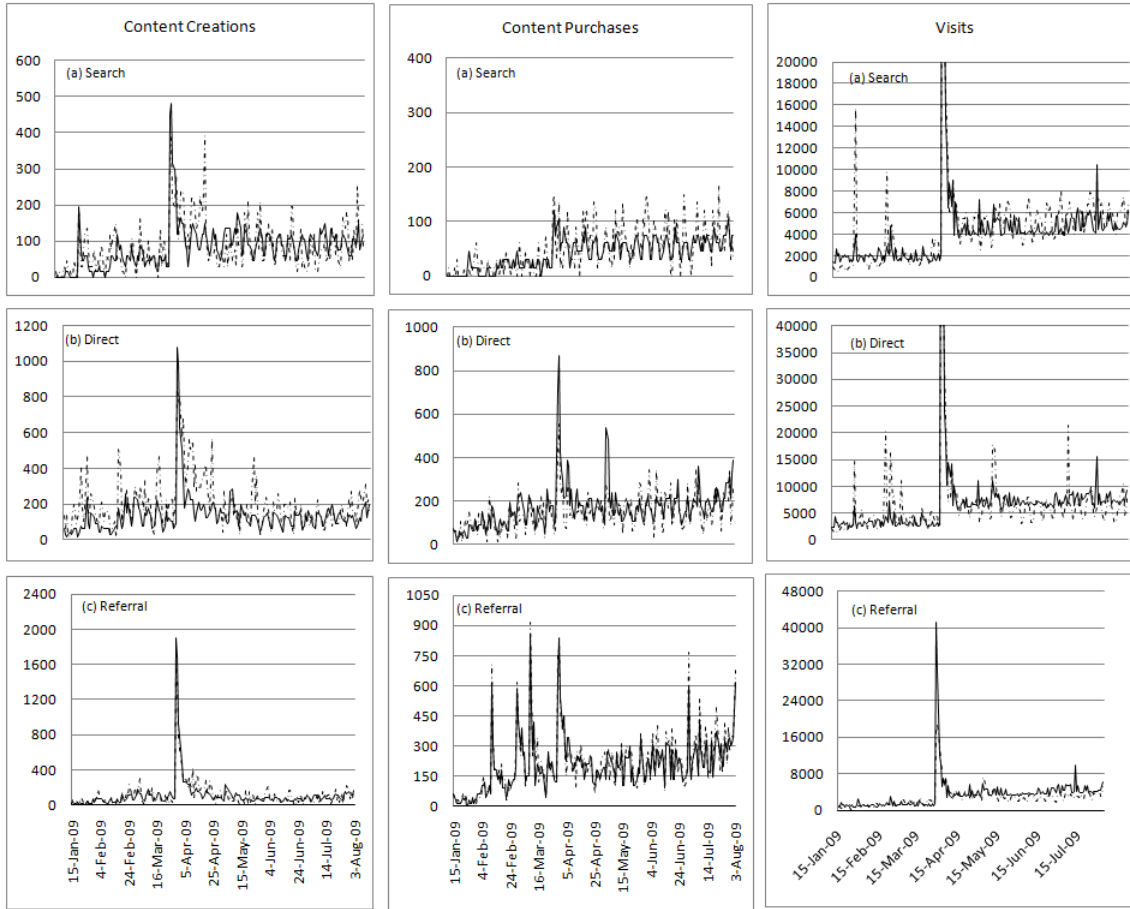


Figure 4: Actual (dotted line) and predicted (full line) number of visits, content creations and purchases

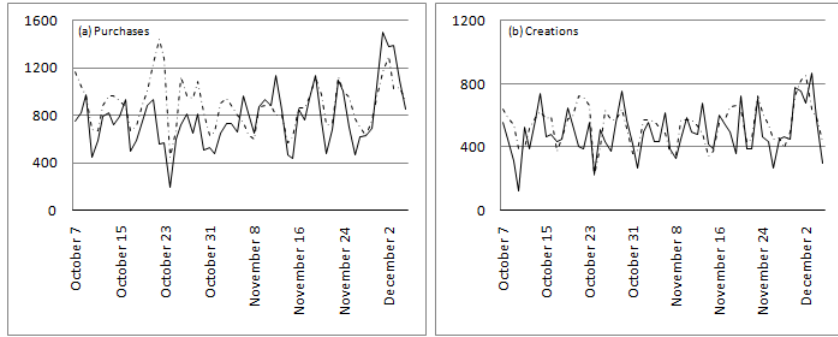


Figure 5: Actual (dotted line) and predicted (full line) number of content creations and purchases for hold out observations.

conditional on actual visits is presented in Figure 5, showing a good predictive ability. Good results are also obtained for the visits stage.

6.2 Content Purchase and Creation

The parameter estimates of the bivariate probit model that captures the behavior of users who visit the website are presented in Tables 2 and 3. The presented formulation and variables included in the model were chosen after careful analysis and comparison with alternative specifications. We discuss the results regarding the decision to purchase content first and continue with the creation decision.

6.2.1 Content Purchase

Table 2 displays the parameter estimates and standard errors related to the decision to purchase content. We observe that there are significant differences in buying propensities between the different segments of users. Comparing the intercepts, we see that visitors arriving from third party websites are more likely to consume content, irrespective of whether they are first time users or returning, while users directed from search engines to the website have the lowest purchase likelihood. For all three web sources, returning visitors are more inclined to buy compared to new visitors. These results have face validity. Referral traffic is driven to MagCloud from another website through a link, frequently from sites that advertise content creation at the platform, leading to visits of users who are a priori more interested in buying a magazine at MagCloud. On the other hand, people

Content Purchase	Variables	Estimate	Std. Error
Segment Intercepts	New Visitors Search	-2.624	0.062
	New Visitors Direct	-2.196	0.059
	New Visitors Referral	-2.082	0.059
	Returning Visitors Search	-2.036	0.060
	Returning Visitors Direct	-1.925	0.060
	Returning Visitors Referral	-1.833	0.059
Own-Side Network Effects	Number of Past Purchases (previous week)	0.135	0.030
Cross-Side Network Effects	Number of Public Issues (previous week)	0.493	0.068
	Creator Decision Effect on Purchase	1.654	0.119
	Correlation of Unobservables	-0.131	0.051
Marketing: Firm	General Event	-0.068	0.005
	Price Promotion (20% off)	0.096	0.013
	NY Times Events (2 days)	0.228	0.060
	NY Times Long Term (dummy after event)	-0.085	0.018
Marketing: Content Creators	Content Creators Events	0.022	0.010
	Average Price per Page	-0.657	0.157
	Average Magazine Size (pages, avg. week)	0.108	0.099
Fit Measures	Log Likelihood	-88899	
	BIC	178095	

Table 2: Parameter estimates for the decision to purchase content

using search engines may have more general objectives, for instance seeking for an online magazine not specific to MagCloud. In this context, search engine advertising is likely to be a less effective online marketing medium to increase the volume of magazine sales when compared to investment in content related websites.

In terms of network effects, we measure the impact of both own and cross-side effects. The number of magazine purchases in the previous week, which captures own-side network effects from past purchasers, has a significantly positive effect on purchases. This can be explained by positive word-of-mouth between past and current users. We also quantify the impact of three cross-market effects. First, we measure the impact of the number of public issues available for purchase. The more magazines published in MagCloud, leading to more product variety, the higher the probability that a purchase will take place. An additional cross-market network effect is reflected in the impact of deciding to create content on the utility of purchasing, for a specific user. Individuals who upload content seem to be more likely to consume that content, i.e., buy the magazine they just published, or purchase any other magazine. This can be explained by the fact that content creators have invested time in knowing the site well and so are more motivated to interact with the platform at all levels. Finally, we find that there is a small negative correlation between unobservables influencing the two

decisions, suggesting that unobserved site characteristics that motivate purchase are likely to lower the intentions to create content and vice-versa. For example, if the site contains an extensive user-friendly tool to create a new magazine, it is likely that visitors focus on creating a new magazine, leaving less time and effort to search and purchase magazines.

In terms of marketing activities, we find that decreases in the per-page printing price (price promotions) have a positive coefficient as expected, as well as the large advertising event at the New York Times. Interestingly, the marketing events originated by HP, denoted as general events - which include online communications from HP in numerous websites - have a negative impact on purchases. We emphasize however that these effects are limited to the utility and probability of purchase, conditional on a website visit. Individuals attracted by “General Events” from HP seem to have less interest in actually buying a title. However, as we discuss in more detail later, once the impact on visits of these events is taken into account, they increase website traffic and have a positive net effect on the number of purchases. The same interpretation can be applied to the negative coefficient of the long term effect of the New York Times advertising.

Concerning marketing activities related to content creators, we find that the average final price of the magazine (which includes a mark-up defined by the creator) has a negative impact on purchases of content, as expected. On the contrary, marketing events from content creators increase the likelihood of purchasing content for users, measuring for example the positive effect of referrals. Finally, larger magazines also provide more utility to consumers. A more detailed quantification of the magnitude of the effects discussed in this section is presented later using elasticities.

6.2.2 Content Creation

The results regarding the decision to create content are showed in Table 3. We start by noting that the segments that were found to be more likely to purchase content are less interested in creating content. Users coming from referring sites are now the least likely segment to upload content, while direct visitors are the most promising ones. Similarly to the purchase decision though, past experience with the website increases the utility of creating content, since the intercepts of the returning visitor segments are significantly larger than those of new visitors. Again, these results are reasonable. On the one hand, visitors referred from other websites are likely to be motivated (by the referral party) to buy content, thus explaining their higher utility from purchase, but not

Content Creation	Variables	Estimate	Std. Error
Segment Intercepts	New Visitors Search	-2.541	0.026
	New Visitors Direct	-2.341	0.022
	New Visitors Referral	-2.964	0.025
	Returning Visitors Search	-1.682	0.022
	Returning Visitors Direct	-1.629	0.021
	Returning Visitors Referral	-1.964	0.021
Network Effects	Number of Repeat Public Issues (prev. week)	0.119	0.050
	Number of New Public Issues (prev. week)	-0.502	0.059
Marketing: Firm	General Events	0.008	0.005
	New Publishing Processes	0.001	0.013
	New Markets (open sign-up for other countries)	0.197	0.019
	NY Times Events (2 days)	-0.062	0.053
	NY Times Long Term (dummy after event)	-0.144	0.020
Marketing: Content Creators	Content Creators Events	0.012	0.011
	Average Markup (previous week)	0.512	0.052

Table 3: Parameter estimates for the decision to create content

from creating content. On the other hand, direct returning users have previous experience with the platform and are likely to know how to create and upload material. The platform is a better match for their content creation interests, which are partially revealed by their direct access to the site and the action to previously bookmark the platform website.

Next, we focus on the network effects among content creators. The existence of more content is expected to have two opposing effects. First, more content is a sign that the site is popular and provides confidence that the content will be seen and purchased in the future. Second, more magazines also create more competition among creators, leading to lower number of purchases for each individual magazine when more content is available in the site. We cannot isolate each of these effects, but instead measure the dominant net effect. Empirically, we measure these network effects of two types of content. Each publication at MagCloud is either a new magazine title or a new issue of an existing magazine. We find that the higher numbers of repeat public issues available have a positive impact on the utility of content creators, implying that they serve as a sign of quality of the platform. This seems reasonable, since in general, our data shows that magazines with multiple issues are likely to be of higher quality and finish, and attract more purchases. On the contrary, the number of new magazine titles has a negative effect on content creation, reflecting negative impact of competition for purchases.

We included five types of marketing actions under the control of the firm. HP general events

have a positive though not significant impact on content creation. HP improved the performance of the platform twice in our data set, when a user-friendly publishing method was made available and when users from more markets (e.g., Canada and the United Kingdom) were allowed to make purchases.¹⁵ We observe that only the latter event had a significant positive effect on content creation, possibly due to the expectation of an increase in platform popularity with the geographic expansion. Finally, we included two variables that capture the New York Times event. In the short term, measured by the day of the event and the following day, the impact is insignificant. In the long term, it appears that the event brought to site consumers that are less likely to generate content than users arriving before the event. We conclude that the event created large awareness but a relatively lower rate of content creations.

Given these results, we find that different marketing campaigns are appropriate depending on the manager’s objectives. For example, referred visitors are the most attractive targets for purchase, indicating that marketing actions at the referral sites (e.g., links, banner ads) are an effective way to increase content sales. However, these actions are less effective at attracting new magazine creators.

6.3 Estimates of Visits Model

Once in possession of the parameter estimates from the site actions, we move to the analysis of the decision to visit the site. The results are listed in Table 4. Our estimates show that there is little heterogeneity in the base utility of visiting the site across the different segments, as the intercepts are insignificantly different. However, there are substantial differences in utility expectations from visiting the site for returning and new visitors. There is an interesting matching between consumer segments and action expectations: new visitors are driven to the site by a strong positive utility expectation of buying content, with a coefficient of 20.4, while returning visitors are motivated by the utility of creating content, with a value of 9.8 for the respective coefficient.

All marketing events, both originated by HP and by content creators, have a significant positive effect on visits. We elaborate on these findings in the next subsection by computing elasticities.

¹⁵Although currently HP allows for international orders, the larger majority of purchases has remained from the United States.

Platform Visits	Variables	Estimate	Std. Error
Segment Intercepts	New Visitors Search	-3.308	0.023
	New Visitors Direct	-3.336	0.046
	New Visitors Referral	-3.405	0.038
	Returning Visitors Search	-3.322	0.075
	Returning Visitors Direct	-3.422	0.086
	Returning Visitors Referral	-3.329	0.053
Expected Utility	New Visitors Creation	-14.321	7.475
	New Visitors Purchase	20.439	6.949
	Returning Visitors Creation	9.823	4.129
	Returning Visitors Purchase	-1.610	3.712
Marketing: Firm	General Events	0.091	0.010
	NY Times Events (2 days)	0.169	0.096
	NY Times Long Term (dummy after event)	0.251	0.015
Marketing: Content Creators	Content Creators Events	0.033	0.016
Fit Measures	Log Likelihood	-3994464	
	BIC	7989040	

Table 4: Parameter estimates for the decision to visit the online platform

	Visits	Creations	Purchases
Price	-0.15%	0.02%	-0.78%
General Events (HP)	0.61%	0.70%	0.55%
Content Creators Events	1.13%	1.01%	2.60%
Creations	0.5	0	0.1
Purchases	5	0	0.7

Table 5: Impact of a change in prices and marketing events on visits, content creations, and purchases, and measures of cross-market effects.

6.4 Elasticity Measures

In order to better evaluate the effect of marketing activities and network effects on sales, we compute some elasticity measures.

6.4.1 Marketing Activities

We use simulation to measure the impact of increases in marketing activities on actions of consumers to account for the interactions and temporal effects across users. We compute the effect of changes in three variables - page price, marketing activities by HP, and content creators events - and list the results in the first three rows of Table 5.

To obtain these numbers, we compare two scenarios of future values of visits, content creations, and purchases for changes in each variable for 60 days.¹⁶ In the "base" scenario, realizations of marketing variables are drawn from their empirical distributions from the last days of our sample, while in the counterfactual scenarios, we change the variables in the following way: for price, we increase price by 1% for all 60 days; for general events and content creators events, we increase the number of events by 1. With the parameter estimates and exogenous variables, we start by obtaining creation and purchase stage utilities, which can be used to simulate the expected maximum utility of uploading and purchasing. We then predict the number of visits per segment. Finally, we combine the number of visits with the probability of content purchase and creation to obtain the final number of predicted uploads and purchases. For each variable, we use ten iterations with different draws and we average the results over iterations and sum over the 60 days.

The results in Table 5 are percent changes from the base to the counterfactual situation for each of the dependent variables of our model. We find that the marketing variables have different effects on consumer actions. Variation in price affects purchases the most, in a negative way as expected, with 1% increase in price leading to a 0.8% decrease in magazine purchases at MagCloud. It has almost no impact in both the number of visits and content creations. The effects of online marketing events created by HP have the stronger effects on visits and creations, with one event increasing visits by 0.6% and creations by 0.7%. Finally, any marketing actions from content creators, which are at zero cost to HP, have the stronger impact on purchases, with one event leading to an increase of 2.6% of purchases. Visits and creations go up by about 1% with one more of these events. We note that the content creator events have a stronger impact on all actions than HP's own events, which suggests that it may be optimal for HP to provide incentives for such activities. We note that in terms of total impact on platform sales, these content creator events happen with lower frequency than marketing actions from HP, making them less important at the aggregate-level.

6.4.2 Market-side Elasticities

Next, we evaluate the responsiveness of consumer actions in the site to increases in purchases and creations of content. We construct these measures by adding one purchase or one upload to the last

¹⁶Any other number of days could be chosen to evaluate the impact of marketing activities. However, longer time periods will allow for a more complete picture, since some of the effects are a result of the interaction of users over time.

day of our data set and comparing the number of visits, purchases, and creations over the next 60 days between the original and this modified scenario. This is an exercise that aims at evaluating the general magnitude and relative proportion of the network effects, and we present the results in the last two rows of Table 5.

First, both content purchase and creation have a significant impact on the number of visits. This effect comes through the expected maximum utilities component, since once the service becomes more attractive to its users, more visitors will flow to the website. For each additional purchase, five future visits are likely to occur. For each additional magazine uploaded, 0.5 more visits are expected. These positive effects seem reasonable, since users deciding to visit the platform will expect higher utility from a site that has a larger amount of content available for purchase or that has a larger number of sales. Second, the purchasing side is responsive with respect to changes to its own numbers, while the creators side is not. More specifically, one purchase leads to an additional 0.7 purchases in the future, while one magazine upload leads to no additional creations in the future. Our results suggest that the content creator side of the market is less responsive to network effects than the purchaser side and that it needs more incentives to be developed. This is consistent with the larger impact of firm marketing actions on creator content than on purchase.

6.5 Managerial Implications

We exemplify the managerial usefulness of our approach with two situations. First, we provide recommendations on investments in the different marketing activities to improve HP profits. Second, we quantify the impact of the “free” advertising obtained by HP from the content creators.

6.5.1 Allocation of Marketing Investments

We measure the allocation of marketing investments on three events using estimates from our model and input from MagCloud management. First, HP can choose the number of marketing events,¹⁷ such as online advertising. Second, HP can choose to do price promotions, by offering a per-page discount to buyers of each magazine copy. Third, we consider the possibility that HP motivating content creators to advertise or refer MagCloud more frequently by providing monetary incentives

¹⁷HP can also choose to change the timing of the events. In our analysis, we use the same timing as observed in the data.

(which would put a cost on an until now free marketing activity).

The profit for HP is given by the following expression

$$\Pi = \sum_{t=1, \dots, T} \delta^{(t-1)} [(c_t - c_{0t}) O_t - G_t c_1 - I_t c_2], \quad (21)$$

where c_t represents the average price per page paid by consumers, which enters the utility of buying and creating content (see Equations 1 and 2), while c_{0t} , c_1 , and c_2 are respectively the HP's variable production cost (e.g., printing), the cost of one HP generated marketing event, and the cost of additional incentives to content creators. O_t , G_t , and I_t are the number of pages ordered, the number of marketing events created by HP and by content creators. The discount rate is denoted by δ .

We evaluate profit variation by changing c_t , G_t , and I_t . After talks with management, we decided to do the analysis of profits for 180 days¹⁸ and test the following alternative scenarios. For price c_t , we chose two alternatives: (1) no price promotion case; (2) one price promotion activity that lasts the first 30 days, with a price discount of 20%, similar to a price promotion activity observed in our data set. For the HP and content creators' marketing activities, we test six alternative situations: maintaining the same level of events, or increasing them by 20%, 40%, 60%, 80%, or 100%. These three dimensions create a grid of $2 \times 6 \times 6 = 72$ cases to measure profits.

Information about the costs in Equation 21 is provided to us by HP.¹⁹ For the variable production costs, the company has costs of 50% of the per-page price (i.e., $c_0 = 50\%$ of the current c_t). The costs of events are mainly justified by the time of HP personnel allocated to MagCloud and cost of online advertising, which amounts to \$30,000 per event (i.e., $c_1 = \$30,000$). For the content creator events, we assume that an incentive similar to the HP event cost is a reasonable number (i.e., $c_2 = \$30,000$).

When comparing the two price promotion cases, we find that across all measures - content creation, purchases, and profits - having a price promotion at the beginning of the quarter is beneficial, for all levels of marketing investments. Thus, we focus our analysis on the case where a price promotion is done and illustrate the results in Figure 6.

¹⁸We use a discount rate of $\delta = 1$, given the short time span of our analysis.

¹⁹As previously, the values are scaled for privacy reasons.

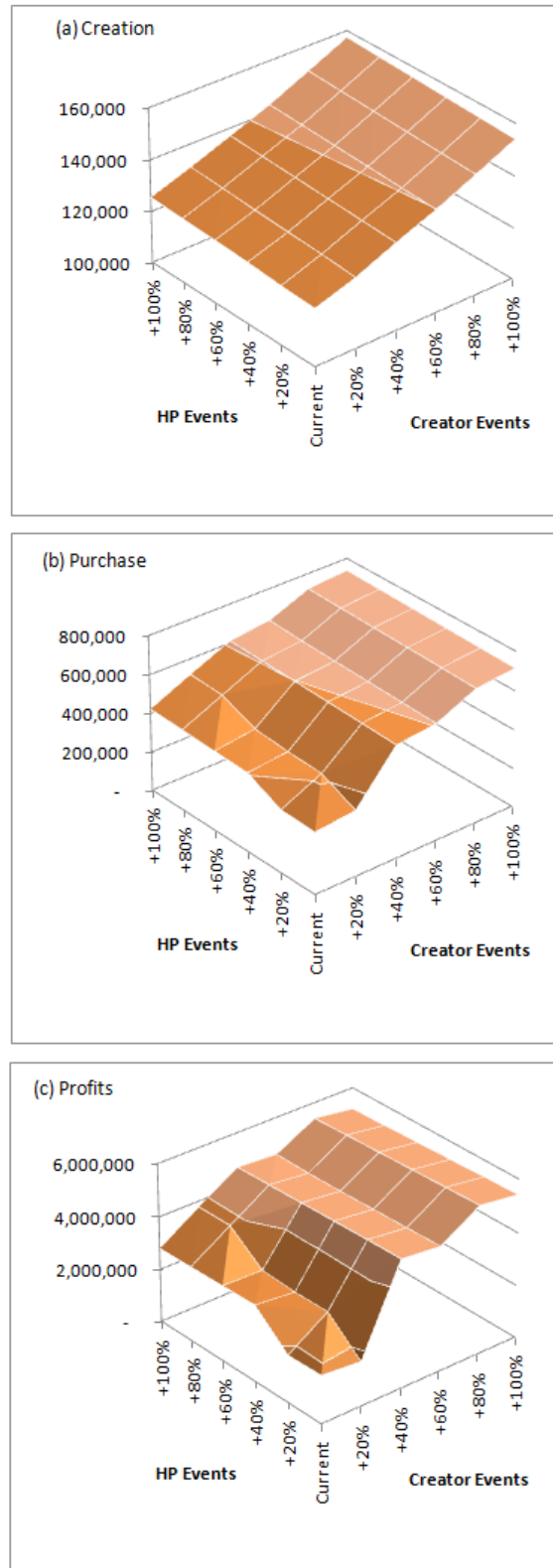


Figure 6: Number of purchases, content creations, and profits for several marketing investment scenarios

	Scenarios		% Change
	Base	No Creators'	
Visits	5,645,992	5,512,725	-2.4%
Content Creation	122,948	120,574	-1.9%
Purchases	204,565	192,407	-5.9%
Profit	\$ 2,171,480	\$ 1,989,107	-8.4%
HP Events	60	60	
Creator Events	23	0	

Table 6: Impact of free advertising from content creators

The figure shows the level of purchases, creations, and profits, for the alternative cases previously described. We find that both the provision and the purchase of content are more sensitive to increases in marketing activities originated by content creators. Profits also increase steeper when there is an increase in creator events, when compared to a similar percentage increase in HP events. Given the observed data and marketing costs, we find that the combination of investments that results in the highest profit for the 180 days is for HP to provide incentives to content creators that increase their marketing actions by 80%, while keeping own events at the current level. This exercise provides evidence that offering additional monetary incentives to individual content creators would increase profits for HP, when compared to the current level of investments and to the decision to invest solely on actions that do not involve content creators.

6.5.2 The Impact of Free Advertising from Content Creators

In this section, we quantify the impact of the marketing events originated by content creators on the profits of the firm managing the platform. Our analysis spans for six months, as in the previous section. We compare the profits from a scenario where the number of events is at the observed level with profits from a hypothetical scenario where the events from content creators is reduced from 23 to zero over a period of six months. All other variables are set at similar values for the two scenarios. For example, the number of HP events is set to be 60 for both situations. The results for the number of visits, content creation, purchases, and profits are displayed in Table 6, as well as the percent change between scenarios.

We observe losses of about 2% in visits and content creation, and about 6% in purchases. Visits for instance drop from 5.6 million to 5.5 million. This translates into a drop of 8.4% in profits,

which represents about \$200,000 for the platform over six months. We note that with lower sales (content purchases), the fixed costs incurred by the company when supporting marketing activities increases in relative terms, which justifies the larger hit on profits than on sales. Consistent with the results in the previous section, the benefit of \$200,000 provides reasons for HP to add monetary incentives to content creators for marketing actions related to MagCloud.

7 Conclusion

In this paper, we present a model for a two-sided online market of user-generated content. We explain decisions to visit the platform, purchase and create content at the individual level. Our model accounts for interaction between the two sides of the market, as well as the relation between the visit decision and actions while in the site. We also measure the impact of marketing actions by the firm and by the content creators.

We combine easily available sources of data on web analytics and online advertising activities to study the multi-faceted demand of an online intermediary in the market of user-generated content. Empirically, we estimate the model using data from the self-publishing magazine site MagCloud, created by HP. We demonstrate that our demand model, based on utility maximization, is able to capture the critical elements of the multi-sided platform's demand and predict future demand with adequate accuracy.

We show that there is significant heterogeneity across consumers, with new visitors being more likely to purchase from the site, while returning visitors becoming the majority of creators of content. One of the limitations of our aggregate-level data is the reduced level of detail about the action history of each individual user with the website. Google Analytics has added more levels by which the data can be disaggregated, that can be used to track the state dependence of past buyers and past creators of content. These data has recently become publicly available to managers and we leave its implementation for future research.

Additionally, we provide evidence that marketing actions of content creators can play an essential part in the development of a user-generated website. Interestingly, we find that these marketing actions can have higher per-event impact on visits and content purchases than similar marketing events originated by the platform. Our model quantifies the results from this free advertising to

the firm on the online market sales, and we implement a number of counterfactual situations that exemplify the model's ability to provide recommendations on investments on marketing activities and price promotions to increase the platform's profitability.

References

- [1] Akerberg, D., Gowrisankaran, G. (2006), "Quantifying Equilibrium Network Externalities in the ACH Banking Industry", NBER Working Paper.
- [2] Argentesi, E., Filistrucchi, L. (2007), "Estimating Market Power In A Two-Sided Market: The Case Of Newspapers", *Journal Of Applied Econometrics*, 22, 1247-1266.
- [3] Armstrong, M., (2006), "Competition in two-sided markets," *Rand Journal of Economics*, 37, 668- 691.
- [4] Baxter, W. (1983), "Bank Interchange of Transactional Paper: Legal and Economic Perspectives," *Journal of Law & Economics*, 26: 541:588.
- [5] Bell, D., Song, S. (2007), "Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing", *Quantitative Marketing & Economics*, 5, 361-400.
- [6] Berry, S., Waldfogel, J. (1999), "Free Entry And Social Inefficiency In Radio Broadcasting", *The RAND Journal Of Economics*, Vol. 30, No. 3, 397-420.
- [7] Berthon, P. and J. John (2006), "From Entities to Interfaces: Delineating Value in Customer-Firm Interactions," in R. F. Lusch & L. S. Vargo (Eds.), *The Service Dominant Logic of Marketing*, 196-207.
- [8] Brynjolfsson, E., Hu, Y., Smith, M., "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers", *Management Science*, Vol. 49, No. 11, 2003.
- [9] Chen, Y., Xie, J. (2007), "Cross-Market Network Effect with Asymmetric Customer Loyalty: Implications for Competitive Advantage", *Marketing Science*, Vol. 26, No. 1, Jan - Feb, 52-66.
- [10] Clifton, B. (2008), *Increasing Accuracy For Online Business Growth*, Whitepaper, Omega Digital Media.
- [11] Comscore (2007), "*Retail E-Commerce Climbs 23% in Q2 Versus Year Ago*", July 30, Press Release.

- [12] eMarketer (2009), January, Press Release.
- [13] Etgar, M. (2008), "A descriptive model of the consumer co-production process," *Journal of the Academy of Marketing Science*, Vol. 36 (March), 97:108.
- [14] Evans, D., (2003), "The antitrust economics of multi-sided platform markets," *Yale Journal on Regulation*, 20, 352-382.
- [15] Gandal, N., M. Kende, and R. Rob (2000), "The Dynamics of Technological Adoption in Hardware/Software Systems: The Case of Compact Disc Players," *The RAND Journal of Economics*, Vol. 31, No. 1 (Spring), pp. 43-61.
- [16] Genz, A., F. Bretz (2009), *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics*, Vol. 195., Springer-Verlage.
- [17] Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, T. Hothorn (2009), "mvtnorm: Multivariate Normal and t Distributions," R package version 0.9-7. URL <http://CRAN.R-project.org/package=mvtnorm>.
- [18] Gupta, S., Steenburgh, T. (2008), "Allocating Marketing Resources," *Harvard Business School Marketing Research Paper* ,08-069.
- [19] Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System", *Econometrica*, Vol. 46, No. 4, 931-959.
- [20] Horsky, D. (1990), "A Diffusion Model Incorporating Product Benefits, Price, Income and Information", *Marketing Science*, Vol. 9, No. 4, Autumn, 342-365.
- [21] Ilfeld, J.S., Winer, R., "Generating Website Traffic", *Journal of Advertising Research*, Vol. 2, Issue 5, Sep/Oct 2002.
- [22] Kaiser, U. and J. Wright (2006), "Price structure in two-sided markets: Evidence from the magazine industry," *International Journal of Industrial Organization*, Volume 24, Issue 1, January, Pages 1-28.
- [23] Maddala, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, New York: Cambridge University Press.

- [24] Mantrala, M. (2006), Allocating Marketing Resources, *Handbook of Marketing*, Ed. by Weitz B. and Wensley R., SAGE Publications Ltd.
- [25] Moe, W., Fader, P. (2004a), "Capturing Evolving Visit Behavior in Clickstream Data", *Journal of Interactive Marketing*, Vol. 18, No. 1, Winter, 5-19.
- [26] Moe, W., Fader, P. (2004b), "Dynamic Conversion Behavior at E-Commerce Sites", *Management Science*, Vol. 50, No. 3, March, 326-335.
- [27] Nair, H., Chintagunta, P., and Dubé, JP (2004), "Empirical Analysis of Indirect Network Effects in the Market for Personal Digital Assistants", *Quantitative Marketing & Economics*, 2, 23-58.
- [28] Parker, G., Van Alstyne, M. (2005), "Two-Sided Network Effects: A Theory of Information Product Design, *Management Science*, Vol. 51, No. 10, 1494-1504.
- [29] Prahalad, C.K. and V. Ramaswamy (2004), "Co-creation experiences: The next practice in value creation," *Journal of Interactive Marketing*, Vol. 18, N. 3 (Summer), 1-10.
- [30] Rochet, J. and J. Tirole (2005), "Two-Sided Markets : A Progress Report," IDEI Working Papers 275, Institut d'Économie Industrielle (IDEI), Toulouse.
- [31] Rysman, M. (2004), "Competition Between Networks: A Study of the Market for Yellow Pages", *Review of Economic Studies*, 71, 483-512.
- [32] Rosse, J.N. (1979), "The Evolution of One Newspaper Cities. Proceedings of the Symposium on Media Concentration," Washington DC: Federal Trade Commission, Vol. II, 429-71.
- [33] Sismeiro, C. and R. E. Bucklin. (2004). Modeling Purchase Behavior at an E-Commerce Web Site: A Task Completion Approach. *Journal of Marketing Research* (August), 306-323..
- [34] Song, I., Chintagunta, P. (2004), "A Micromodel of New Product Adoption with Heterogeneous and Forward-Looking Consumers", *Quantitative Marketing & Economics*, 1, 4, 371-407.
- [35] Wilbur, K., (2008), "A Two-Sided, Empirical Model of Television Advertising and Viewing Markets", *Marketing Science*, Vol. 27, No. 3, May-June, 356-378.

- [36] Zhang, J., Krishnamurthi, L. (2004), "Customizing Promotions in Online Stores", *Marketing Science*, Vol. 23, No. 4, Fall, 561-578.