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

Paulo Albuquerque, Polykarpos Pavlidis[*], Udi Chatow, Kay-Yut Chen, Zainab Jamal[†]

28th August 2011

Abstract

We measure the value of promotional activities and referrals by content creators to an online platform of user-generated content. To do so, we develop a modeling approach that explains individual-level choices of visiting the platform, creating, and purchasing content, as a function of consumer characteristics and marketing activities, allowing for the possibility of interdependence of decisions within and across users. Empirically, we apply our model to Hewlett-Packard's (HP) print-on-demand service of user-created magazines, named MagCloud. We use two distinct data sets to show the applicability of our approach: an aggregate-level data set from Google Analytics, which is a widely available source of data to managers, and an individual-level data set from HP. Our results compare content creator activities, which include referrals and word-of-mouth efforts, with firm-based actions, such as price promotions and public relations. We show that price promotions have strong effects, but limited to the purchase decisions, while content creator referrals and public relations have broader effects which impact all consumer decisions at the platform. We provide recommendations to the level of the firm's investments when "free" promotional activities by content creators exist. These "free" marketing campaigns are likely to have a substantial presence in most online services of user-generated content.

Keywords: Demand Modeling, User-Generated Content, Online Marketing, 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 and Chen Kay-Yut are principal scientists; Zainab Jamal is a researcher scientist, at Hewlett-Packard 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 U.S. 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 markets have emerged with considerable success, led by user-generated content websites such as Lulu, eBay, and YouTube. According to eMarketer (2009), 82 million people in the U.S. created online content during 2008 at least monthly, with the majority of uploaded materials being related to social network sites, personal videos, and blogs. This number is expected to grow to 114.5 million in 2011.

In these markets, a firm/platform usually plays the role of intermediary that maximizes its own objectives by bringing together content creators, consumers, and in some cases advertisers. Frequently, 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 user 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. The platform may benefit from a wide array of marketing activities that can affect its growth. The firm may engage in advertising, price promotions, or public relations to attract more visitors and influence purchase or creation of content. Additionally, content creators, besides populating the platform with materials, serve as marketing agents, by advertising their own content, or generating referrals and links to the uploaded content in other websites. Given the inter-connectedness and viral community structure of the Internet, the relation between marketing activities by the firm and the decisions of content creators is likely to play an essential role in the development of most user-generated content platforms. 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 a wide range of marketing activities, including content creator actions, on the consumer decisions to visit the platform, and on decisions to create and buy content.

The management of marketing activities for an online platform has a 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 can allocate their marketing budgets to influence each stage. Finally, as previously mentioned, content creators themselves frequently generate significant "free" marketing for the platform's content, in the form of referrals and marketing campaigns. These activities, which are in most cases unobserved by managers, should be taken into account when predicting growth or allocating resources.

To address these challenges, we develop an individual-level model of heterogeneous users making multiple decisions. First, consumers choose to visit the site, given expectations about the utility of actions once at the site. Second, conditional on visit, consumers decide to purchase content, create content, or both.[1] We solve this model backwards. We start by defining and estimating 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 when consumers create 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 and marketing activities, allowing for the possibility of network effects. Once the second stage of content creations and purchases has been estimated, 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. We allow content creators to take into account expectations of short-term future sales when deciding to create content, but we assume that they are making a "yes or no" decision to upload content, and not an inter-temporal one. This is mainly due to the nature of our empirical application, where content is likely to be short-lived and postponing content creation while waiting for the platform to increase its size would make content untimely and very significantly reduce its appeal.[2] This in turn negates the benefits of an inter-temporal

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

[2] We note that this is a reduced-form approach to modeling content creation, since we do not have information about the full process of content creation, but only its outcome in the form of content uploads.

decision, which might not be the case in other applications, where considering content creators as forward-looking and using a dynamic programming approach would be more suitable.

Empirically, we use two different data sets 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 online content and then purchase a printed copy of their own magazine, or purchase someone else's magazines. In the first data set, our main source of data is Google Analytics, containing daily time series about the number of visits, content purchases and creations, as well as information about marketing actions from both MagCloud and creators of content. Google Analytics also tracks the number of new and returning visitors to the site. The data is easy to obtain and free for most managers of online websites. The second data set is collected by HP and it contains individual-level transactional information about content generation and purchase, also at the daily level. By estimating our model separately with the two data sets, we demonstrate the flexibility of our model regarding alternative situations where different types of data are available, as well as the robustness of our model and results.

We provide several substantive insights. First, we find and quantify the significant interdependence between the decisions to create and purchase content, both within and across users. Second, we test a number of different promotional activities, and find that the effectiveness of marketing tools by the company and content creators vary across platform users, when taking into account network externalities and interdependence of decisions. Third, we offer recommendations on different levels of marketing investments in the two sides of the market to improve profits of the user-generated content platform. Finally, we quantify the impact of "free advertising" by content creators to the usage and profits of the online service.

More specifically, our results show that the dependence between the creation of content and purchases has multiple dimensions. Within a user, visitors who have created content are more likely to buy content at the time of creation, mainly content produced by themselves. Across users, visitors are more prone to purchase content when the total amount of recently created content is higher and more likely to create content if they expect a higher number of future purchases of their content. In terms of promotional activities, we compare the impact of three types: price discounts, content creator events and referrals, and events generated by the firm through public relations. We

find that price discounts have strong effects on the number of purchases, while any events originated by public relations and content creator marketing referrals influence all user decisions. Finally, our results show 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 and referrals bring in about 50% of the sales of the platform and we suggest that HP should provide additional incentives to content creators to increase their referral behavior. The impact of content creator activities is likely to be strong in the development of most user-generated content websites and it should 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 describes several managerial implications. Section 8 concludes.

2 Relevant Literature

Given our objective of developing a method that can provide input for managerial decisions of an online service, we 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 some but not all the elements of individual behavior, and we compare its results with the ones obtained using an individual-level data set. While there is some loss of information in web analytics data, it is in general more applicable, given its easier availability. Additionally, these data allow us to quantify the effects of content creators marketing activities, which would be unobservable using individual level data solely from the site navigation records.

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. Kannan, Kline Pope and Jain (2009) propose and implement a model of demand for digital content offered by traditional publisher agencies online. They successfully apply their model to recommend pricing policies for different content formats (i.e. pdf or print), to the National Academies Press. This study is a clear example of the potential that marketing science models can have in formulating, refining and evaluating the marketing actions of digital content providers. Our approach is similar to these two studies, using predicted purchase probabilities to optimize future marketing activities. However, 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. Godes and Mayzlin (2009) report evidence, from both a field and a lab experiment online, that firms can initiate exogenous word-of-mouth that enhances sales. The possibility of encouraging consumers to share their positive view about a product or service, enriches the marketers toolkit with additional options regarding promotion. Our empirical findings that online postings, initiated by the platform and/or its users, generate incremental volume of website visits and transactions corroborates this evidence. We further utilize these findings to make recommendations about additional efforts that aim to induce even greater online word-of-mouth/advertising.

Since our paper deals with a platform that comprises of a two-sided market of creators and purchasers of content, we describe some of the previous related 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. 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. Related to the product category in this paper, 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.[3] In this literature, multi-sided markets imply the existence of indirect network effects, where valuation of a product (or service) by consumers depends on how many other 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. Similar to our paper, the mentioned empirical studies quantify or account for 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 user-generated content 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.

In online settings, recent papers have made advances regarding two-sided markets and user-generated content. For example, Yao and Mela (2008) study an online two-sided market in the context of auctions, presenting a structural model that measures the value of buyers and sellers and providing an empirical analysis of how the two sides should be priced. The creation (uploading) and consumption (downloading) of multimedia content from Internet social networking sites and mobile portal sites is analyzed in two papers by Ghose and Han (2010a, 2010b). In the first paper, the authors build a dynamic structural learning model and find that consumers benefit from experience from content creation and usage behavior, while in the second they find evidence that content creation and content usage are negatively correlated due to time constraints using mobile Internet. Interestingly, we find the opposite result, where consumers that produce content are more likely to buy content. This is due to the different nature of our market, and we develop its analysis in the results section.

Finally, given that consumers can become producers of content in our study, it is important to refer to literature on co-production. For instance, Etgar (2008) provides three drivers of co-production: economic, psychological, and social. Among the social motives, it is important to

[3]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.

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. Lerner and Tirole (2002) also mention ego-satisfaction and signaling incentives as motivation for open source software creation, which complement any possible future monetary and career progression motivations. 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. Additionally, Zhang and Zhu (2011) make use of a natural experiment to study the causal relationship between group size and incentives of users to contribute content to an online platform (Wikipedia). They find that, due to social effects, contributors are more active when the size of the online community increases. 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 and volume of available content in the platform.

3 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 the platform where content is made publicly available for viewing and purchase. Consumer utility is maximized with decisions regarding the visit to the online platform, and subsequently the creation and consumption of content. In our managerial application section, we show how the firm can take the consumer behavior into account to make appropriate marketing investments that improve its profits. We again note that contrary to traditional two-sided markets, users can take on the dual role of content creators and content buyers.

We model the decision process in two stages involving three decisions. Consumers start by

choosing whether to visit the platform. At the visiting decision stage, the utility of consumers is driven by personal preferences for online services and by expectations about the utility of their actions at the site, if they choose to use the platform. In case of a visit, consumers then face two choices. First, whether to produce content, and second, whether to purchase content. We start by describing the second stage decisions of purchasing and creating content, and then move backwards to the initial visit decision, matching the order of our estimation approach.

3.1 Consumption and Production of Content

Conditional on visiting the platform, at each time period $t = 1, ..., T$, a consumer has the possibility of four choices: browsing the site (without purchase nor creation of content), purchasing content, creating content, and both creating and purchasing content.

Visitors draw utility from consuming content. In our application, this corresponds to reading user-generated magazines purchased on an online platform. Thus, when making the decision to purchase a magazine, consumers are anticipating the benefits of owning and reading that magazine, or alternatively of giving that magazine to someone else to read. To make the purchase decision, benefits and costs of acquiring user-generated content are compared, and influenced by the following factors. Users have individual preferences for content usage, online purchasing, and perceptions about the quality of the site and content. This perceived quality is influenced by the firm's activities, such as advertising, by online referrals and comments from content creators, and potentially by independent sources, such as the appearance of an article about the platform in the New York Times. Other marketing activities may also influence the decision to purchase content, such as the price of the available content, or any price promotions offered to the buyers of content.

At the moment of the purchase decision, we assume that consumers are aware of any existing marketing activities. We also assume that the consumer purchase decision is affected by the price of content. Since we do not have detailed information about the price discovery process employed by the consumers, we use the average price over all content existing at the occasion of purchase as a proxy to this effect. Additionally, consumers are also familiar with some of the actions of previous content creators, since previously created content is usually available for browsing. In our model, this is modeled as individual i knowing the cumulative number of content materials created recently, from $t - \tau_1$ until time t-1 or a similar lagged measure. In practice, such information is available in

most websites, as a form of a counter of the number of content available, for instance, in the main page or when users search for an item.

When considering the creation of content, users can have the following two objectives in mind. First, they may want to create content that can be purchased in the following days by other visitors or subscribers, leading to monetary reasons behind content creation. Second, they may want to share the output of their creativity with other users, with similar motivation of artists or volunteers, without a monetary reason in mind. In fact, in our application, we found through surveys that the most important reason to upload content was recreational, and not potential profits from content sales. At the decision time t, the utility from either recreational or monetary reasons is a function of site visits or magazine orders in a small number of the days after upload, since content created at time t will be available for purchase or viewership from period $t+1$ onward until $t+\tau_2$. Following this, we assume that when creating content, users have correct expectations about the benefits of such decision and about the sales of their content, which are based on information available at the website. As in the purchase decision, creators are also aware of marketing activities.

In our model, we assume that consumers do not face inter-temporal decisions because of several reasons. First, the content is in most cases "perishable", relevant only for the time when it is created, and loses its appeal quickly over time. Thus, there is no reason to consider waiting to create content in a future period when platform has a broader base of consumers, since the utility of postponing would be considerably diminished. Second, the monetary value of the magazines and the corresponding sales volumes are small and thus it is likely that the incentives to find out about and respond to future platform changes are weak. Third, we note that we face a different situation from papers that consider forward-looking consumers, such as Song and Chintagunta (2004), where in durable goods categories, consumers are aware of price decreases over time, and they leave the market once a purchase occurs, or Nair and Hartmann (2008), where, in the razor category, consumers are buying tied goods that involve multiple future decisions. Any of these aspects create additional incentives for forward looking behavior that are not frequently observed in the purchase of user-generated content.

Based on these assumptions, users compare the utilities of purchasing content u_{1it} and of creating content 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 for identification This leads to consumer decisions of

purchase d_{1it} and creation d_{2it} defined as follows:

$$\begin{array}{l} d_{1it} = 1 \quad \text{if } u_{1it} > 0,\ d_{1it} = 0 \text{ otherwise} \\ d_{2it} = 1 \quad \text{if } u_{2it} > 0,\ d_{2it} = 0 \text{ otherwise} \end{array}. \tag{1}$$

Formally, we define the utility of purchasing content for individual i at time t as

$$u_{1it} = \xi_{1i} Z_{1it} + \alpha_{1i} X_{1t} - \beta_{1i} p_t + \delta_1 d_{2it} + \lambda_{1i} D_{2\, t-\tau_1} + \epsilon_{1it}, \tag{2}$$

while the utility of creating content is given by

$$u_{2it} = \xi_{2i} Z_{2it} + \alpha_{2i} X_{2t} + \lambda_{2i} E(D_{1\, t+\tau_2}) + \epsilon_{2it}. \tag{3}$$

The vectors Z_{1it} and Z_{2it} include individual-specific (or consumer segment-specific) characteristics that measure the heterogeneity in preferences for buying and creating content at the website, including decisions in previous time periods. The variables X_{1t} and X_{2t} are vectors of observed variables that measure the appeal variation of the platform. In our application, the variables include, for instance, promotional activities from the platform owner and from content creators. They also include exclusion variables that enter one utility function but not the other. In the creation decision, we include the daily number of issues uploaded at MagCloud that are kept private and not displayed at the website for browsing by visitors. This variable captures the appeal of MagCloud as a place to create and print magazines, separated from buyer effects or externalities resulting from the number of visits or purchases in the website. In the purchaser side, we include for example the sales of offline publications to capture aspects that influence buyers but not creators of content. We discuss the identification of possible network effects between the two sides of the market in more detail in the estimation section.

The average price for content sold at the platform at period t is captured by p_t. Consumers purchasing content are charged p_t and so we expect a negative effect of price on their utility. While there might be individual differences in content prices, we use the average price as a general proxy index since we are modeling the purchase of any content available, and not of a specific content title. Although this may lead to some attenuation bias on the price coefficient, it is an unavoidable limi-

tation of most incidence models. Additionally, individual content creators are very small compared to the full market and thus we do not explicitly model pricing decisions, assuming that average prices are exogenous. We control for any remaining correlation in the variation over time between the average price of magazines and unobserved (by the researcher) variation in time varying factors by including temporal fixed effects, such as quarter and weekend dummies. Additionally, the inclusion of the sales of offline magazines as an independent variable should also capture "unobserved" temporal shocks that may be related to the demand of magazines.

The remaining observed variables capture any network effects or relation between the two decisions. The decision of creating content at time t, d_{2it}, may influence the subsequent decision to purchase content, since a creator may want to order his own content to distribute among friends and/or subscribers. The approach to include this structural shift in the utility of purchasing content, operationalized with the inclusion of the content creation decision d_{2it} in the utility u_{1it} of content purchase, has been proposed by Heckman (1978) in his paper on simultaneous equations in the presence of endogenous dummy variables. For identification of the parameters in Equations 2 and 3, 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 users of the platform who create content are likely to have a positive shift in the utility to purchase content, as their self-generated content will be available for purchase. Finally, purchasers can be influenced by the quantity of available content recently created in past periods $t - \tau_1$ to $t - 1$, denoted by $D_{2\,t-\tau_1}$. We expect the increased availability of recent content to have positive effects, since consumers are likely to find a better match for their preferences if they face a larger set of available content or if they have a liking for variety.

On the other hand, since creators of content are motivated by the sales of content, the utility of creating is modeled to be influenced by the expectations about potential revenue obtained in the following days after publication, $E(D_{1\,t+\tau_2})$. Since there is a possible profit motivation, we define $D_{1\,t+\tau_2}$ as the cumulative number of purchases multiplied by price, over $t+1$ to $t+\tau_2$ time periods. To create the expectation of revenue $E(D_{1\,t+\tau_2})$, visitors to the website can use information available at time t. Thus, to construct $E(D_{1\,t+\tau_2})$, we first regress values of $D_{1\,t+\tau_2}$ (which are observed by the researcher) on a subset of observed variables X_{1t} that influence content purchase at time t, and

then use the predicted values from that regression as expectations.[4] The coefficients for revenue expectations are allowed to be heterogeneous across consumers, as we expect users with and without experience on the platform to form different predictions about the popularity of their content. The expectations of future purchases can be discounted to time t using an observed discount rate, but given the short term focus of our application, we present a simpler model with no discounting. The values τ_1 and τ_2 reflect the relevant time periods for past content or future content to influence time t's decisions.[5]

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.,

$$\begin{bmatrix} \epsilon_{1it} \\ \epsilon_{2it} \end{bmatrix} \sim N(0, \Sigma), \tag{4}$$

with

$$\Sigma = \begin{bmatrix} \sigma_1 & \rho \\ \rho & \sigma_2 \end{bmatrix}. \tag{5}$$

These unobservables are assumed to be realized once consumers reach the platform, and need to be integrated out in any decisions occurring before this stage, such as at the decision to visit the platform. Our assumptions lead to a system of equations that forms a multivariate probit model with structural shift (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 been previously created by the creator 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

[4] We use a subset of the variables that are part of the purchase decision of consumer because some of the variables in X_{1t} will only have an immediate effect on purchase utility, such as a one-day promotion or event, and will not provide any signal to build future expectations. We ran an alternative specification including all variables of the purchase decision as regressors of future purchases and the results do not change significantly. In the final specification, the variables that explain most of the variation in future revenue are the number of new issues of a magazine and the number of following issues in a magazine series available at time t. We also tested different values τ_1 and τ_2 and found no significant differences in the results.

[5] We note that we do not solve a dynamic programming problem, since we see the decision to create content as not an inter-temporal one. As previously mentioned, in our case (and in other types of online content, like blogs, news, and tweets), content is in most cases short-lived and it will lose its value (or freshness) if not uploaded on the day or week when it is created. Thus, there is not a decision 'to create today or tomorrow', but instead 'to create or not' at the platform.

creating and then purchasing 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. The parameters to be estimated are $\Theta_1 = \{\xi,\, \alpha,\, \beta,\, \delta,\, \lambda,\, \rho\}$.

Combining Equations 2 and 3, 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., $u_{1it} = v_{1it} + \epsilon_{1it}$ and $u_{2it} = v_{2it} + \epsilon_{2it}$. The implied probability of consumer i making these choices, conditional on visiting the platform, is given by

$$P(d_{1it} = 0,\, d_{2it} = 1) = \int_{-\infty}^{-v_{1it}} \int_{-v_{2it}}^{+\infty} \phi(\epsilon_{1it}, \epsilon_{2it}, \rho) d\epsilon_1 d\epsilon_2, \tag{6}$$

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

$$P(d_{1it} = 0,\, d_{2it} = 0) = \int_{-\infty}^{-v_{1it}} \int_{-\infty}^{-v_{2it}} \phi(\epsilon_{1it}, \epsilon_{2it}, \rho) d\epsilon_1 d\epsilon_2. \tag{7}$$

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:

$$\widehat{S}(d_{1it} = 0,\, d_{2it} = 1) = \sum_{i=1}^{M_t} P(d_{1it} = 0,\, d_{2it} = 1). \tag{8}$$

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 engaging in any content-related action:

$$\widehat{S}(d_{1it}=0,\ d_{2it}=0)=\sum_{i=1}^{M_t}P(d_{1it}=0,\ d_{2it}=0). \tag{9}$$

Since there is no closed form formula for the integrals in these expressions, we use simulation to obtain approximations of the integrals. In our two data sets, we either observe the total number of individuals involved in each of the four decisions at the aggregate level, or the actual choice of each visitor at the individual level. More details about how we obtain the parameters are provided in the estimation section.

3.2 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 or an offline service. For instance, in our empirical application, there are no very 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. To make the visit choice, users compare the utility of the platform with the utility of the outside good. The utility of visiting the online platform is given by

$$u_{3it}=\xi_{3i}Z_{3it}+\alpha_{3i}X_{3t}+\psi_i E\left[max\left(u_{1it},0\right)\right]+\omega_i E\left[max\left(u_{2it},0\right)\right]+\varepsilon_{3it}. \tag{10}$$

Heterogeneity in intrinsic preferences is captured with individual (or segment) characteristics Z_{3it}. The vector X_{3t} contains exogenous variables that influence the utility of visiting the platform, such as marketing actions by the firm related to the quality of the platform. The terms $E\left[max\left(u_{1it},0\right)\right]$ and $E\left[max\left(u_{2it},0\right)\right]$ denote the expected maximum utility over the choices to purchase and create content, conditional on visiting the site. The expected value is over the unobserved components at the time of the decision, ϵ_{1it} and ϵ_{2it}. Our assumption is that, before visiting the platform, users are aware or have correct expectations about the level of utility that they can derive from visiting the platform. In other words, the two expectation terms imply that, before visiting the website, consumers have knowledge about all other components of the utility functions u_{1it} and u_{2it},

such as the number of available content, the average price, and marketing activities, and the only unobserved terms that are integrated out are the errors ϵ_{1it} and ϵ_{2it} in the purchase and creation utility functions. This simplification where consumers are aware of product features before starting search often appears in other search models (e.g., Kim, Albuquerque, and Bronnenberg, 2011).

By including the utility of future actions 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 segment specific). This in turn helps us obtain a structural representation of the utility of a potential visitor that connects the two decision stages.

Finally, we assume that the unobserved part of the visiting utility ε_{3it} follows a normal distribution,

$$\varepsilon_{3it} \sim N\left(0, \sigma_3\right), \tag{11}$$

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 term σ_3 is set to one for identification purposes.

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

$$\begin{gathered} \psi_i E\left[max\left(u_{1it}, 0\right)\right] + \omega_i E\left[max\left(u_{2it}, 0\right)\right] = \\ \int\int \left[\psi_i max\left(u_{1ti}, 0 \mid d_{2ti}\right) + \omega_i max\left(u_{2ti}, 0\right)\right] \phi(\epsilon_1, \epsilon_2, \rho) d\epsilon_1 d\epsilon_2 \end{gathered} . \tag{12}$$

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. Normalizing the utility of the outside alternative to zero for identification purposes, a user decides to visit the platform ($d_{3it} = 1$) if

$$v_{3it} + \varepsilon_{3it} \geq 0 \Leftrightarrow \varepsilon_{3it} \geq -v_{3it}, \tag{13}$$

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:

$$P(d_{3it} = 1) = \int_{-v_{3it}}^{+\infty} \phi(\epsilon_{3it}) d\epsilon_3. \tag{14}$$

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, called MagCloud, 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 a popular online site to create custom magazines among individual or small publishers. The platform is designed for generation of content and its diffusion online and in printed version, 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.[6]

4.1 Data Description

Our model can accommodate both aggregate and individual level data, and we demonstrate its application using two different data sets about the usage of MagCloud. Some variables overlap the two data sets, and some are exclusive to one. By making use of these two alternative data sets on the same market and individuals and estimating the model separately with each data set, it is our objective to show the robustness of our model and illustrate the managerial findings that can be obtained with each type of data.

Our objective is to explain the behavior of users interested in the platform, and we describe some of the characteristics of these agents. Creators of content at MagCloud tend to be individual

[6]For more details, visit www.magcloud.com.

publishers, small groups or organizations, who use the service as a way to complement some other activity, such as a blog, or a sports, recreational, or academic activity. In some cases, these publishers have their own exclusive audience, to whom they purchase and distribute the printed content, or alternatively refer the site. It is also possible that content creators also buy magazine copies for themselves. Content is usually of short-term interest, and sold mostly in the few weeks that follow the upload, with about 60% of content sales done within two weeks, increasing to 70% within three weeks. On the buyers side, a large percentage of users are the followers of the publishers or content creators. In terms of areas of interest, we find that photography, fashion, art, and entertainment are the most important topics, with more than 40% of the magazines. Other themes include sports, lifestyle, technology, and religion.

The first data set is measured at the aggregate market level, for which the main source of information 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.[7] 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 are collected with first party cookies named page tags and have the advantage of not being contaminated by bot visits to the website, not requiring the identification of Internet Protocol (IP) addresses, and being able to measure visits from the computer's cache memory. This 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 relative numbers and trends are measured with acceptable precision (Clifton, 2008). To increase our confidence in 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 the numbers from the two data sets.

From GA, the information is aggregated over website visits at the daily level. This aggregate-level data is a result of all website traffic, without any sampling or selection bias. We collect the number of website visitors over time, total and by consumer segments defined by HP, chosen

[7]The 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).

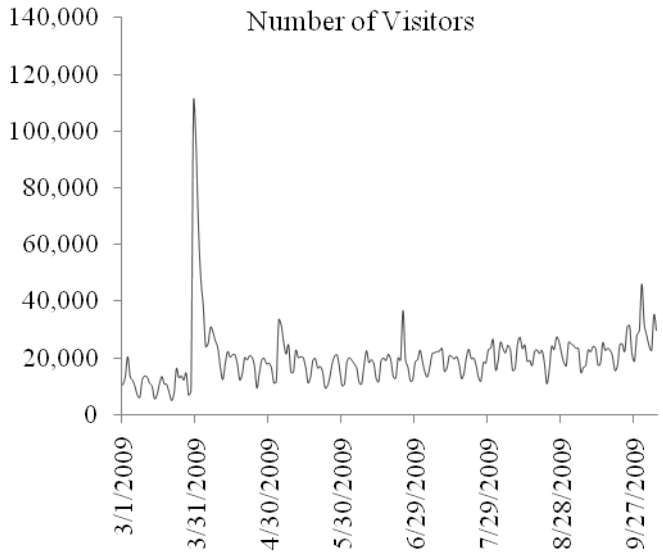

Figure 1: Number of visits, from March 1st to October 5th, 2009

according to 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 will be 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.[8]

Figures 1 and 2 show the time series for our three dependent variables. The first figure displays the daily number of visitors, while the second figure shows the number of content purchases and uploads. We have information about the complete time series of these variables since the service was made available to general online public in the beginning of June of 2008. However, we removed from our analysis the initial months of data, which were dominated by beta versions and software tests that are not the focus of this study and can potentially create biases on the demand parameters. Thus, we focused our study on observations from March 1st, 2009 to October 5th, 2009.

In these figures, we observe some interesting patterns. The visitor numbers show a considerably

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

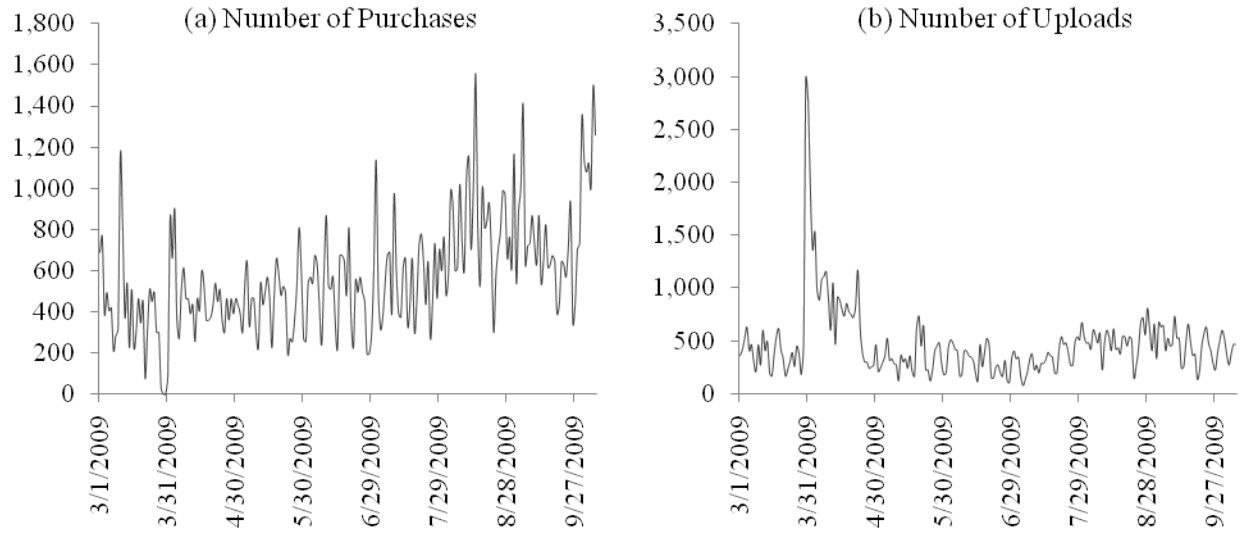

Figure 2: Daily number of content orders and uploads, from March 1st to October 6th, 2009

stable pattern, with higher visits during the weekdays than in the weekends. 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 public relations event when the website was featured and described in the New York Times and in its online site, with a direct link to www.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. We take its impact into account as an explanatory variable included in our model and consider a counterfactual situation later in the paper to quantify the value of such an event to HP.

One of the distinguishing details between the aggregate and individual data sets available to us is the classification of visitors by source, available only at the aggregate level in the Google Analytics data set, and not at the individual level. Visitors that reach the site from referrals 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 new users will need time to understand the requirements to create content, which may lead to multiple visits to the platform.

We present additional statistics in Table 1, where we describe the number of actions in the site

		Potential Actions		
Consumer 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%)
Direct	1,517,370 (97.6%)	17,805 (1.1%)	15,615 (1.0%)	3,780 (0.2%)
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%)	7,440 (1.4%)
Direct	704,655 (91.5%)	18,045 (2.3%)	34,380 (4.5%)	12,870 (1/7%)
Referral	1,072,335 (94.0%)	33,675 (3.0%)	24,660 (2.2%)	9,600 (0.8%)

Table 1: Descriptive Statistics about User Decisions by Segment

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 when 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 31%, 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.

The three source segments present different browsing patterns. Table 2 shows some descriptive statistics. We find that the referral segment spends on average the least amount of time (2:45 minutes) in the website and visits the lowest number of pages (3.9) out of all segments, probably directed to the magazine related to the referral website. Users in this segment are more frequently first time visitors and spend a lower amount of money, mainly because of making individual orders instead of larger orders by direct and search consumers. It is somewhat surprising that the bounce rate, i.e. the percentage of consumers that leave the website after visiting only one webpage, is higher for the referral segment, since it is likely that these users would have a better match with MagCloud. However, what seems to happen is a further dichotomy within this segment, with either a very good match or an immediate recognition that the publishing service is not suitable for these visitors. Additionally, since visitors from referring websites are likely to access the service's website through a different "landing" webpage, namely the page of the magazine they were referred to, they

	Consumer Segment		
	Direct	Search	Referral
Avg Time on Site	03:56	04:07	02:45
Bounce Rate	48.91%	47.99%	53.04%
Pages/Visits	5.09	5.52	3.9
% New Visits	65.51%	64.76%	70.04%
Avg Value	$43.16	$42.91	$20.45
Per Visit Value	$1.26	$0.72	$0.57

Table 2: Statistics about Browsing Patterns per Segment

are perhaps less inclined to browse around. For these visitors, it is more likely that they get all their needed information from the first webpage they face. We also note that the search segment displays clear signs of being involved in browsing behavior, spending considerably more time on the website, visiting more pages, and as a result spending less money per visit. The direct segment, which very likely includes people that are familiar with the website, displays the highest amount spent per visit as well as the highest value per transaction.

A second part of the aggregate level data set is based on Google information 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/researcher 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.[9] 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 "HP Events". The variable "Content Creator Events" counts the daily number of Google Alerts related to 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. On average, there are 4.8 such events per month. The variable "HP 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

[9]Whether 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/)

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 include additional marketing actions related to the platform. As mentioned earlier, 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 longer term effects (or a structural break in the utilities) due to this large event, we include two additional dummy variables for time periods after this activity; a short-term effect lasting eight days after the major event and a long-term effect for all periods after the event until the end of our sample period.

Concerning the individual-level data set, we have access to a number of variables from a transactional database at HP. This database includes several time series containing the date, description, and other details about both content creations and purchases, tagged by the user identification number. With this information, we are able to construct a history of purchases and content creation for any person that decided to do so using MagCloud, which enable us to estimate additional heterogeneity based on previous purchase and/or creation behavior. Unfortunately, the HP system does not track visits at the individual level. In our estimation, we need to augment our individual observations by assigning visits with draws from the observed aggregate-level empirical distributions.

These data also contain the price per page and the numbers of pages of each magazine. 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 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 also have information about the potential market for the MagCloud platform 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 visiting the platform, and assume it to be constant for the time periods in our data. With longer time-series data set, it may be useful to include some dynamics in the total market potential, but we do not do so for this application. Our estimation provides a unified setting for inference and prediction using either of these two types of data.

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. For computational purposes and 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 stage. Additionally, we also assume 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. This does not imply that the two decisions are uncorrelated, since the expectations of the second stage are part of the utility of visiting the platform.

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 description, our model has the form of a bivariate probit with a structural shift. We obtain estimates of the parameters of interest by maximizing the log of the following likelihood function:

$$L = \prod_{t=1}^{T} \prod_{i=1}^{N} L_{it}. \tag{15}$$

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

$$\begin{aligned} L_{it} = P(d_{1it}=0,\, d_{2it}=0)^{I(d_{1it}=0,\, d_{2it}=0)} \times P(d_{1it}=1,\, d_{2it}=1)^{I(d_{1it}=1,\, d_{2it}=1)} \\ \times P(d_{1it}=1,\, d_{2it}=0)^{I(d_{1it}=1,\, d_{2it}=0)} \times P(d_{1it}=0,\, d_{2it}=1)^{I(d_{1it}=0,\, d_{2it}=1)} \end{aligned}, \tag{16}$$

where $I\left(d_{1it}=0,\ d_{2it}=0\right)$ 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, 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. With individual level data, the likelihood is computed for each consumer.

The probabilities in Equation 16 are given by the expressions in Equations 6 and 7 (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 Genz et al. (2009), which provides approximations of integrals from the normal distribution and has been shown to perform well in Monte Carlo simulations. To improve the speed of the integration in the individual-level model, we modified Matlab code made public by Alan Genz based on method described in Drezner and Wesolowsky (1989) to approximate bivariate normal integrals.

5.2 Visiting Stage

Given the estimates of the content purchase and creation stage, we can compute the expected maximum utility of a potential visit and use this, along with the other explanatory variables, to get estimates of the visiting utility function in Equation 10. For each period, to approximate the expectations of visitors for their on-site actions, we start by using 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 variance-covariance matrix. Once in possession of the unobserved draws, the utility of creating content and its simulated decision is obtained before computing the utility of purchase, to account for the structural shift in the utility of purchase. Finally, we can compute the maximum utility for the two decisions, repeat R times, and average the results over the R repetitions to create the expectations that enter the utility of visiting the platform.

We estimate the first stage decision as a single probit equation, using the following likelihood

function:

$$L_V = \prod_{t=1}^{T} \prod_{i=1}^{N} L_{Vit}, \tag{17}$$

with

$$L_{Vit} = P(d_{3it} = 1)^{I(d_{3it}=1)} \times P(d_{3it} = 0)^{I(d_{3it}=0)}. \tag{18}$$

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 segment, which are a sum of $I\left(d_{3it} = 1\right)$ 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}. With individual level data, we directly apply Equation 18 to data.

To obtain correct standard errors that account for simulation error and error from the estimation of the creation and purchase stage parameters, we use a bootstrapping technique. For a number of bootstrap iterations B, we repeat the Monte Carlo draws in each stage for unobserved components, parameters, and 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.

5.3 Consumer Heterogeneity

We briefly discuss here the heterogeneity included in the individual and aggregate level model. We capture consumer heterogeneity in two distinct ways depending on the data set used. For the aggregate level data, we use a discrete segment approach, to make use of the classification of consumers offered by Google Analytics. Since we cannot track individuals over time, we cannot identify individual heterogeneity. Additionally, any identification of random coefficients would come from functional form assumptions and not from data, since we cannot link multiple actions over time for each individual, i.e., each visit is different for each individual and for each time period. However, we have information about discrete classification of consumers into segments while performing actions at the site and use it to include different reactions to marketing activities, and so the coefficients α_{1i}, α_{2i}, λ_{1i}, and λ_{2i} can vary across these segments. Intrinsic preferences for creating and purchasing content also vary across visitors, which we capture using intercepts ξ_{1i} and ξ_{2i} and a dummy

variable included in vectors Z_{1t} and Z_{2t}. The observed heterogeneity has the following formulation:

$$\begin{array}{l} \alpha_{ji} = \sum_{s=1,\ldots,S} \alpha_{js} I\left[i \in s\right],\ j \in \{1,2\} \\ \xi_{ji} = \sum_{s=1,\ldots,S} \xi_{js} I\left[i \in s\right],\ j \in \{1,2\} \end{array} . \tag{19}$$

The indicator variable $I\left[i \in s\right]$ takes the value of 1 if individual i belongs to segment s and 0 otherwise. We define consumer segments based on the available 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.[10] 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.[11] Our final number of observed segments in the empirical application is 6 ($S = 3$ access segments $\times 2$ past usage segments $= 6$).

In the individual level data, we observe the history of consumer interaction with the website. Thus, we include observed heterogeneity based on past decisions of content creation and purchase, and unobserved heterogeneity with a random coefficients approach, for which we identify the variance of the parameter distributions.

5.4 Identification

The identification of similar models has been discussed in the literature before (e.g., Manski, 1993) and it is usually a hard problem to solve empirically. In our case, the main reason for the difficulty

[10] In 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.

[11] We 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.

to identify interactions between decisions is the potential existence of multiple factors that can be confounded with network effects, on both the content creation and purchasing sides of the market. Here we discuss briefly our strategy and data used to provide separation and identification of the several effects included in the utility functions.

In our application, we want to separate out the following effects: (1) factors that influence content creation but not purchase; (2) factors that influence content purchase but not creation; (3) network effects from content creation to purchasing; (4) network effects from content purchasing to creation; (5) person-specific interdependence of decisions; and (6) unobserved factors that influence content creation and purchase simultaneously.

We first discuss the excluded variables that allow us to identify effects (1) and (2). To capture the impact of factors related to content creation but not purchase, our strategy is to include an observed variable that captures the appeal of creating content at MagCloud that does not benefit from purchases. This variable is the daily number of issues uploaded at MagCloud that are kept private and not displayed at the website for browsing by visitors. When creating content, users can choose to make their content available to all visitors in the platform, or keep it private. In the later case, the users are purely interested in participating in the creator side of the market, and are not influenced by the number of potential buyers or visitors attracted to MagCloud. In other words, this variable is very useful at providing identification, since it instruments for the general appeal of MagCloud as a place to create and print magazines, separated from any influence of network effects from buyers or even other visitors to MagCloud.[12] We note that it is possible that there is an indirect effect of consumption of content on the creation of private content, negating the exclusion condition, since more visits to the website to consume content may lead MagCloud to appear higher in search rankings, leading to more visits to create private content. We tested the inclusion of the private content volume in the visits stage, and found it to be insignificant, minimizing this concern.

In the purchasing utility equation, we included the publication numbers of two popular offline magazines, Time and Sports Illustrated (Audit Bureau of Circulations, 2010). Both these magazines have a larger number of pages dedicated to photos, and focus on current events or sports, which

[12]We do not include these private issues as a dependent variable in our model, for two reasons. First, our objective of study is the public part of HP's service and how agents interact in this platform. Second, the private printing service is not the focus of HP's managers, who told us that the creation of a network and online platform was their primary objective with MagCloud.

are some of the main publication areas and topics at MagCloud. The publication numbers of offline magazines capture the general intent to purchase content, in this case magazines, and is not influenced by the effect of content creator activities at the platform. The offline magazine circulation numbers are general enough to move with exogenous events but at the same time, they are not anticipated or known by the smaller publishers at MagCloud for any current day or week. It is possible that given the slightly different focus of the offline magazines compared to the ones included in Magcloud, this variable is not very strongly correlated with the content purchase at MagCloud and is itself a weak instrument. However, we find a significant coefficient of this variable when included in the purchase side, which demonstrates that indeed it picks up some of the time variation of interest in buying magazines.

Additionally, to control for similar interest of buying magazines but in an online setting, we used data from Google Trends. Google Trends provides weekly information about the volume of search for words or terms online. Since a large percentage of content at MagCloud is related to fashion, we used the customization of terms that can be searched at Google to provide us with a variable that would be related to the purchase of magazines, but avoid capturing the interest of creators of magazines. We collected the time series with the search volume for the term that includes the words "Fashion" and "Magazine", but intentionally excludes the words "Create" or "Publish". This search volume index will pick up the general interest in fashion magazines, but will not include searches from users specifically interested in creating or publishing a magazine. Although the search term removes some of the search related to content publication, it is still possible that creators of content use this term to investigate the popularity of fashion magazines, which would influence their content upload decisions, and thus this term would also be correlated with the creation side. This does not appear to be the case for several reasons. First, the outcome of such a search will produce results related to the largest online magazines about fashion, such as Vogue or Elle, magazines that are available for purchase but do not involve any user creation, representing an equivalent measure of the offline publication numbers described in the previous paragraph, but in an online environment. Second, the Google Trends variable captures unexpected or temporary increases or decreases in browsing interest for fashion magazines, which small magazine creators, interested in selling to their communities of followers, are unlikely to anticipate and thus should not influence a planned publication date of content. In fact, we find little seasonality or trend patterns in the search volume,

and so even if content creators have access and use this information, changes in search are hard to anticipate based on previous knowledge of the search results. Finally, we tested the correlation of the proposed term with both creation and purchase of content, by including it simultaneously in both utility equations and found it to be significant in the purchase side but insignificant in the creators side. With this reasoning, we use this variable to complement the offline magazines numbers to capture factors that influence purchases but not the creations. Finally, for identification, we note that there are also other variables that are exclusively included in the purchasing utility, such as the average number of pages per magazine, which are taken into account only by purchasers of content and not by creators of content.

To capture the effect of content creation on purchase denoted by λ_{1i}, we include the lagged number of creations on the utility of purchasing, accounting for the importance of having more and more fresh variety or better match with buyer preferences. The data on the number of purchases and lagged creation numbers identifies the coefficient of interest. Using the lagged variable allows us to have this network effect disentangled from any other unobserved factor at the time of decision. Substantively, the lagged variable captures the information that a consumer receives when he visits the website about the number of the magazines created in recent days or weeks available at MagCloud for browsing. We tested several time lags, from a day and up to a month, between the number of creations and the time of decisions, to test for possible serial correlation in the errors, and the results are statistically equivalent. This approach of using a lagged variable to address the issue of unobserved effects that might impact network agents and be perceived as social networking between agents has been proposed by Manski (1993) and related papers have applied it, such as Van den Bulte and Lilien (2001), Manchanda et al. (2008), and Nair et al. (2010).

To measure the network effects of purchases on content creators λ_{2i}, we include the expected number of future purchases in the utility of creating content. This variable uses data that is not present in other effects, more specifically, the number of purchases after time t, from periods $t+1$ to $t+\tau_2$, where τ_2 is set to 15 days. Thus, the number of expected purchases captures the effect of future revenue from content sales on the decision to create a document today. As before, we tested different time intervals between time t and the start of the future purchases variable, without any significant change in coefficients.

We are left with the two last effects: person-specific interdependence of decisions and unob-

served factors that influence content creation and purchase simultaneously. The person-specific interdependence of decisions is quantified by the structural shift parameter δ_1, and the frequency of simultaneous actions of creating and buying content, compared to the frequency of other actions, provides identification of this parameter. If we observe significant number of purchases and creations happening at time t from the same individuals, then the parameter will be expected to positive. On the other hand, if the largest majority of decisions include creation without purchase, then this parameter is expected to be negative. Finally, with these previous data and variables in place, the error term will then capture any unobserved effects at time t, and we allow for contemporaneous correlation between the two shocks through the correlation parameter in the error term.

6 Results

We start our discussion of the results by presenting some performance measures. 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 of the 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 3 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, for the aggregate level estimation.

We see that the model does a good job explaining the variation of 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. The model performs equally well in the visits stage.

Additionally, for the individual level data, we compare the estimated probabilities of purchase and creation of users that make different decisions at the website. First, we separate the individuals based on the observed decision to create content. For users who do create content, the average estimated probability of creation is 8.3%, while for users that do not create content, the respective average estimated probability is 0.001%, showing that our model discriminates correctly creators

of content. Second, for purchasers, we do a similar analysis. The average estimated probability of purchase is 24.7% for observed purchasers, and 0.04% for non-purchasers.

We also computed the mean absolute percentage error (MAPE), for each of the dependent variables. The MAPEs for both data sets range from 0.16 to 0.28 for the three stages, showing that the model is able to explain a large percentage of the variation in the dependent decisions. Additionally, we evaluate the model's ability to predict future consumer decisions. We form a hold out sample of 60 observations after the last time period included in our estimation, from October 7th to December 5th. We use the real values 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 conditional on actual visits is presented in Figure 4, showing a good predictive ability. In the hold out, the MAPE ranges from 0.16 to 0.26 across the several decisions.

6.2 Content Purchase and Creation

The parameter estimates of the model that captures the behavior of users who visit the website are presented in Tables 3 and 4. As previously mentioned, the model is estimated separately for each data set. The formulation and variables were chosen after careful analysis and comparison with alternative specifications. Most of the independent variables are common across the two data sets, with the main difference coming from how consumer heterogeneity is captured. We discuss the results regarding the decision to purchase content first and continue with the creation decision.

6.2.1 Content Purchase

Table 3 displays the parameter estimates and standard errors related to the decision to purchase content. We observe significant differences in buying propensities between the different segments of users. Visitors arriving from third party websites or through direct access are more likely to consume content, 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 other websites through a link, frequently from sites that advertise content creation at the platform, leading to

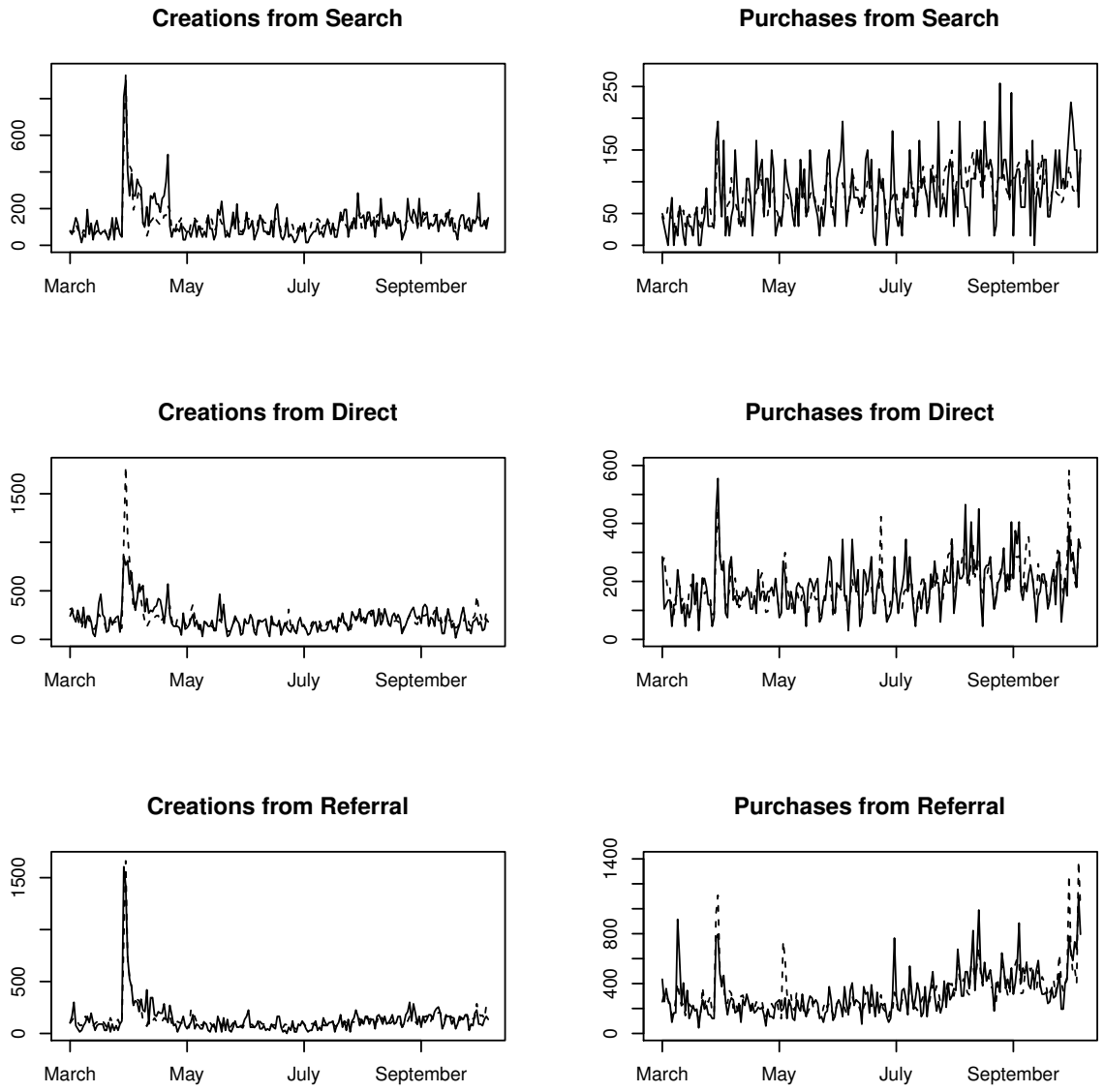

Figure 3: Actual (full line) and predicted (dotted line) number of content uploads and purchases, by segment. Segments are search, direct, and referral by rows from top to bottom.

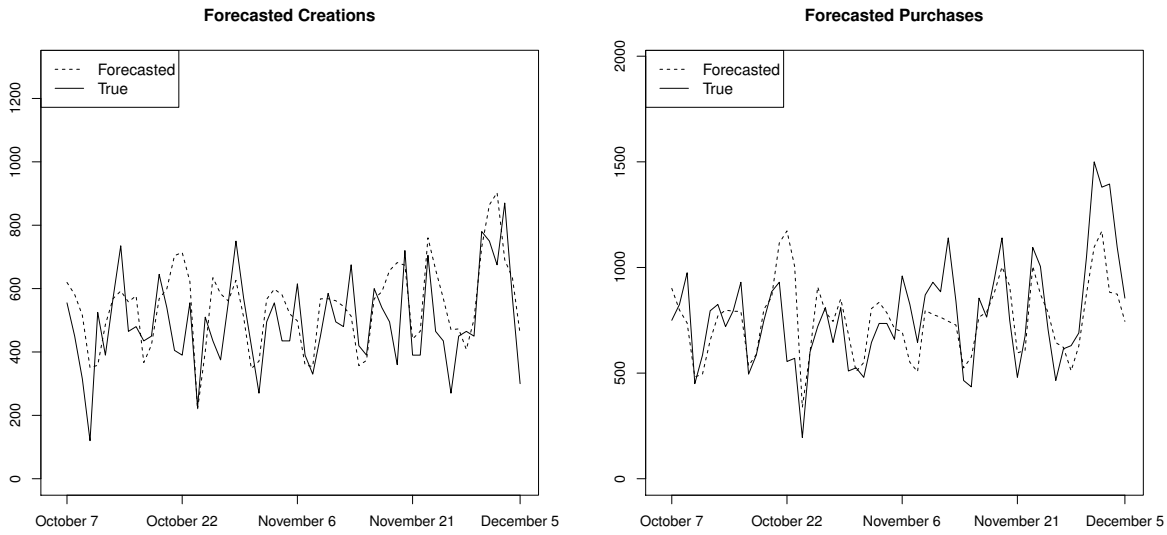

Figure 4: Actual (full line) and predicted (dotted line) number of content creations and purchases for hold out observations.

visits of users who are *a priori* more interested in buying a magazine at MagCloud. On the other hand, people 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.

Additional insights about consumer heterogeneity can be obtained from the individual-level data. We find that users who previously bought a magazine are more likely to buy again, with a coefficient of 0.77, probably due to satisfaction with the obtained magazine. Former creators of content tend not to buy content (-0.57), specializing instead in content creation, as it will be discussed below. Unobserved heterogeneity is limited, likely because of the small number of purchases and uploads compared to the total number of decisions.

We measure the impact of cross-market effects, in three ways. First, we include the number of publicly available magazines (both new issues and issues in a series of a magazine) at the website, a potentially important network effect from the creator side to the purchaser side of the market. We find that the more magazines published in MagCloud, leading to more product variety, the more likely a purchase will take place. Second, an additional cross-market network effect, within user, is reflected in the impact of simultaneously creating content on the utility of purchasing,

		Data			
		Aggregate-Level		Individual-Level	
Content Purchasing	Variable	Mean	Std. Error	Mean	Std. Error
Heterogeneity					
	New Visitors Search	-0.613	0.636		
	New Visitors Direct	-0.095	0.620		
	New Visitors Referral	-0.212	0.615		
	Repeat Visitors Search	-0.002	0.636		
	Repeat Visitors Direct	0.131	0.620		
	Repeat Visitors Referral	0.010	0.614		
	Intercept			-2.243	0.158
	Unob. Heterogeneity			0.010	0.001
	Previous Creator			-0.567	0.061
	Previous Buyer			0.766	0.024
	Weekend	-0.021	0.012	0.007	0.013
Interdependence of Decisions	Available New Content	0.154	0.051	0.200	0.055
	Available Rep. Content	0.130	0.046	0.137	0.069
	Heterog. Avail. Content			0.004	0.020
	Structural Shift	1.561	0.152	3.658	0.228
	Correlation	-0.049	0.065	0.066	0.082
Marketing Variables					
	HP Events	-0.057	0.008	-0.042	0.008
	Creator's event	-0.024	0.014	-0.016	0.016
	NY Times Event 2 days	0.111	0.080	-0.143	0.087
	NY Times Event 8 days	-0.125	0.034	-0.049	0.036
	NY Times Long Term	-0.142	0.024	-0.240	0.025
	Pages per Issue	0.414	0.147	0.696	0.197
Price per Page					
	Search	-0.743	0.730		
	Direct	-0.872	0.507		
	Referral	-0.028	0.398		
	Discount Price	0.125	0.013	0.143	0.014
	Mean Price Effect			-0.800	0.496
	Unob. Heterogeneity			0.070	0.030
Exclusion Variable					
	Offline Magazine Sales	-0.626	0.192	-0.004	0.020
	Search Volume Index	1.305	0.500	1.723	0.554

Table 3: Parameter estimates for the decision to purchase content

denoted by the variable structural shift. Individuals who upload content seem to be more likely to simultaneously consume that content, i.e., buy the magazine they just published, or purchase any other magazine at the same time period. This can be explained by the fact that content creators, once done with the creation process, want to have at least one printed copy of their material as soon as possible. We note that these purchases are likely to be of their own content. Combining this result with the finding in the previous paragraph, we can say that creators of content tend to be more likely to buy their own content (at the time of creation), and less likely than the rest of the users to buy other content available at the website. Third, we find that there is not a significant correlation between unobservables influencing the two decisions. These results are consistent across the two data sets.

In terms of marketing activities, we find that the marketing events originated by HP have a negative impact on purchases. We emphasize however that these effects are at this stage limited to the probability of purchase, conditional on a website visit. Individuals attracted by these events seem to have less interest in buying a title having a lower probability of purchasing, but, as we discuss in more detail in the visits stage, once the impact on visits is also 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 event, while the content creator events are not significant at this stage. Additionally, magazines with more pages provide significantly more utility to consumers.

We see that the average final price of the magazine has a very small impact on purchases of content, with some segments having negative but insignificant coefficients. This is a direct consequence of very small variation of the average price per page over the period of data, with most of the time periods having values between 25 and 27 cents. However, the strongest price variation, driven by the price promotions (in this case, a discount of 20% offered by HP on any purchase during 3 weeks in our data), has a significant positive effect, showing the users are in fact price sensitivity in this market. As a final remark regarding price, we find that consumers who are referred to the website tend to have the most inelastic demand, which is explained by their better fit with the content offered at the platform. The magnitude of these effects is discussed in more detail later in the paper using elasticities.

Finally, offline magazine circulation numbers, included to control for factors that influence con-

tent purchase but not creation, are negatively correlated with purchase probability, but significant only at the aggregate-level data. At the individual-level data, it is likely that some of the heterogeneity resulting from individual previous actions captures the variation over time of the appeal to buy content, which explains the less significant parameter. The negative coefficient seems to indicate that the appeal of the website for purchasing content moves in an opposite direction to the popularity of offline magazines, evidence that they are considered potential substitutes. Quite the reverse, the search volume index of the term "fashion magazine" excluding "publish" or "create" has a significantly positive effect on the choice to purchase a magazine. This seems reasonable, since the term captures the general online interest for browsing and likely purchase of content similar to the one offered by MagCloud, and does not present the substitution nature of offline content.

6.2.2 Content Creation

The results regarding the decision to create content are showed in Table 4. We start by noting that the segments 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 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 is partially revealed by their direct access to the site and the action of previously bookmarking the platform website. The incentive to content creation originated from experience with the website is additionally captured by variables in the individual-level data. Both past purchasing behavior and creation behavior leads to higher probability of creation content in the future.

Next, we focus on the network effects among content creators. We find that future revenue, here captured by the expected revenues in the next 15 days, is significant in the aggregate level data, for both new and returning visitors with coefficients of 2.9 and 1.7 respectively, although

		Data			
		Aggregate-Level		Individual-Level	
Content Creation	Variable	Mean	Std. Error	Mean	Std. Error
Heterogeneity					
	New Visitors Search	-2.413	0.032		
	New Visitors Direct	-2.189	0.029		
	New Visitors Referral	-2.777	0.025		
	Repeat Visitors Search	-1.462	0.025		
	Repeat Visitors Direct	-1.421	0.024		
	Repeat Visitors Referral	-1.744	0.024		
	Intercept			-2.791	0.077
	Unob. Heterogeneity			0.101	0.030
	Previous Creator			1.215	0.088
	Previous Buyer			0.675	0.086
	Weekend	-0.007	0.015	-0.076	0.022
Expectations about Purchases	New Visitors	2.881	0.294		
	Repeat Visitors	1.683	0.226		
	Previous Creator			1.326	1.083
	Previous Buyer			-1.650	1.124
	No Experience			1.265	0.401
	Unob. Heterogeneity			0.310	0.178
Marketing Variables					
	HP Events	-0.041	0.007	-0.059	0.014
	Content Creator Events	-0.006	0.016	-0.237	0.023
	NY Times Event 2 days	0.280	0.077	0.103	0.150
	NY Times Event 8 days	0.267	0.028	0.017	0.069
	NY Times Long Term	-0.321	0.020	-0.045	0.103
Exclusion Variable					
	Private Issues	0.043	0.022	0.085	0.037

Table 4: Parameter estimates for the decision to create content

returning visitors adjust their expectations downwards. In terms of the individual level data, we find an interesting heterogeneity of valuing future purchases. Past creators value the future revenue in a positive way while the expectation of past buyers about revenues has negative impact on the choice to create content, showing a segmentation of the market into users that want to create versus buy content. However, the effect of expected revenues for both these groups is not significant, more likely because it is absorbed by the groups' significant intercepts about creation in general. Interestingly, users without any prior experience, as creators or buyers, value expected revenues in a positive and significant way with a coefficient of 1.3. This intuitive result suggests that evidence on potential future revenues may be useful in attracting new creators but perhaps is not as effective with experienced users.

We included three marketing actions under the direct or indirect control of the firm. HP events show a negative coefficient, similar to the result obtained in the content purchase decision. Content creator events have a similar effect, although significant only at the individual-level data in this case. We also included three variables that capture the New York Times event. At the aggregate level data, it seems that in the short-term the event brought to the site consumers who are more likely to generate content than users arriving before the event. In the period following the event, the probability of creating conditional on a visit reduced (-0.321 for NYT Long Term effect), although less so for the first eight days (0.267 NYT Event 8 days). The interpretation of the negative long term effect is that the New York Times event increased drastically the awareness of the platform but a part of the subsequent visitors were less likely to become creators of content, compared to the first people who discovered and visited the service's web site. The respective coefficients are insignificant in estimation using individual-level data. The differences between individual and aggregate level results are explained by (1) the stronger control for individual heterogeneity at the individual-level and (2) by differences in the tracking of content creation between data sources. Google Analytics tracks uploads, and only a portion of them are translated into actual magazines, which are the numbers tracked by HP in the individual data set.

Finally, we also included the number of private issues to control for unobservable factors that influence content creation but not purchase. As expected, we find a positive and significant result, capturing in this way the impact of effects that influence the creation of public and private magazines, but not the purchase of content.

Given these results, 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 adds) are an effective way to increase content sales. However, these actions are less effective at attracting new magazine creators. These findings will have significant impact on the managerial applications included in the next section.

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 5. 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, we find that repeat visitors have in general positive expectations about the utilities achieved once at website from their actions, which makes them more likely to visit the platform.

There is an interesting matching between consumer segments and action expectations, especially regarding the creation of content. Using the individual level data, we find that in general, users with higher expectations to generate content, tend to visit the website less frequently, with the coefficient of -5.1, and maybe use alternative platforms. However, when the user is a previous creator of content, then his expectations are much more positive than the rest of the users, probably due to the good experience of previous usage, given by the positive coefficient 4.3. This reveals that past creators of content have favorable views about the website.

All marketing events, both originated by HP and by content creators, have a significant and very positive effect on visits. In the visiting stage, the interest for offline magazines moves in the same direction as the browsing behavior to reach the website, capturing the appeal for magazine content in general, without the additional requirement of purchase. We elaborate on these findings in the next subsection by computing elasticities.

6.4 Elasticity Measures

To better evaluate the effect of marketing activities and network effects on sales, we compute some elasticity measures using the model based on the aggregate data. We use simulation to measure the impact of increases in marketing activities on actions of consumers to account for the interactions

		Data			
		Aggregate-Level		Individual-Level	
Platform Visits	Variables	Mean	Std. Error	Mean	Std. Error
Heterogeneity					
	New Visitors Search	-5.055	0.606		
	New Visitors Direct	-5.101	0.603		
	New Visitors Referral	-5.104	0.601		
	Returning Visitors Search	-5.165	0.610		
	Returning Visitors Direct	-5.287	0.612		
	Returning Visitors Referral	-5.202	0.609		
	Intercept			-5.502	0.032
	Past Creators			0.340	0.009
	Past Purchasers			-0.249	0.007
	Unobserved heterogeneity			0.010	0.003
	Weekend	-0.116	0.011	-0.047	0.007
Expected Utility					
	Returning Visitors Creation	9.407	3.382		
	Returning Visitors Purchase	-1.022	2.707		
	Expectation of U(Buy)			-1.865	0.018
	Expectation of U(Create)			-5.131	0.019
	Expect. U(Buy) from Buyers			0.551	0.024
	Expect. U(Create) from Creators			4.703	0.023
Marketing:					
Firm	General Events	0.060	0.016	0.078	0.004
	NY Times Event (2 days)	0.266	0.159	0.307	0.046
	NY Times Event (8 days)	0.093	0.061	0.085	0.052
	NY Times Long Term	0.126	0.027	0.056	0.007
Marketing:					
Creators	Content Creators Events	0.028	0.014	0.502	0.006
	Available Content	0.376	0.123	0.765	0.013
Other Factors	Offline Magazines	0.573	0.184	1.050	0.062

Table 5: Parameter estimates for the decision to visit the online platform

	Visits	Creations	Purchases
Price	0.01%	0.02%	-0.21%
General Events (HP)	0.36%	0.19%	0.11%
Content Creators Events	0.18%	0.18%	0.10%

Table 6: Impact of a change in prices and marketing events on visits, content creations, and purchases.

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 display the results in Table 6.

To obtain these numbers, we compare two scenarios of future values of visits, content creations, and purchases for changes in each variable. 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 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 probabilities, 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 probabilities 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 6 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 0.21% decrease in magazine purchases at MagCloud. It has almost no impact in both the number of visits and content creations. Again, we note that this price sensitivity does not include the response to the price promotion of 20% done by MagCloud, which is evaluated in the next section of the paper. The effects of online marketing events created by HP have the stronger effects on visits and creations, with one event increasing visits by 0.36% and creations by 0.19%, and a lower impact on purchases of 0.11%. Finally, any marketing actions from content creators, which are at zero cost to HP, have impact between 0.10% to 0.18% on each of the three decisions. We note that the HP events have a stronger impact on visits and similar effect of purchases and creations, compared to creator events, but that the later are free for HP. We

provide recommendations for the allocation of investments across these tools in the next section of the paper.

6.5 Within and Across-User Interdependence of Decisions

In this section, we compare the magnitude of the spillover effects from content creation on purchases, looking specifically at two different but co-existent levels: (1) the incentive to purchase content for an individual that created content, which is a "within-person" effect; and (2) the incentive to purchase content if in general there are more available magazines in the website, which is an "across-users" effect. To measure these two effects, we run counterfactual situations and compare to the original estimates. To compute the magnitude of the "within-person" effects, we remove the dummy variable that captures the outcome of the decision to create content from the utility of purchasing content, thus eliminating this interdependence of decisions.[13] To compute the magnitude of the network effect from content creation to purchase across users, we set the parameter for available content, both new and repeated issues in a series, to zero. We then ccalculate and report the difference in the total number of purchases between the actual scenario and each of these two counterfactual situations.

We find that both effects are significantly important. If no "across-users" effect existed from content creation to purchase, content purchases would be lower by 51.4%. This value measures the general improvement appeal of the platform to purchasers by the availability of content, compared to a platform purely based on the printing service of self-produced content, where that content is not made available for public browsing and purchase. The "within-person" effect in our application is also strong. If we do not allow for interdependence of decisions of creation and purchase of content within an individual, the probability of purchasing content goes down by 21.6%. In this platform, it is very frequent to see content creators also buy their own content to distribute to friends and other readers, which justifies this strong effect. In other platforms that do not follow the same business model or have different dynamics across segments the importance of the "within-person" effect may not be as high. At the same time, the "across-users" effect may well be dominant in cases where the two sides interact as much or more than at MagCloud.

[13] This is done by setting all the values of the dummy variable to zero.

7 Managerial Implications

We exemplify the managerial usefulness of our approach with three applications. First, we provide recommendations on investments in the different marketing activities to improve HP profits. Second, we quantify the impact of the New York Times event and compare it to a price promotion. Finally, we quantify the impact of marketing activities and referral effects of content creators.

7.1 Allocation of Marketing Investments between Events from the Firm and from Creator of Contents

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,[14] 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 motivates content creators to advertise or refer MagCloud more frequently by providing monetary incentives (which would put a cost on additional levels of an, until now, free marketing activity).

The profit for HP is given by the following expression

$$\prod = \sum_{t=1,\ldots,T} \delta^{(t-1)} \left[(c_t - c_{0t})\, O_t - c_1(G_t) - c_2(I_t) \right], \tag{20}$$

where c_t represents the average price per page paid by consumers, entering the utility of buying content through the final price p_t, while c_{0t} is HP's variable production cost (e.g., printing). $c_1(.)$ and $c_2(.)$ are functions that translate different levels of HP generated marketing events and additional incentives to content creators, to costs. O_t, G_t, and I_t are respectively the number of pages ordered, the number of marketing events created by HP and by content creators. Finally, the discount rate is denoted by δ.

We evaluate profit variation by changing G_t and I_t. After talks with management, we decided to do the analysis of profits for 180 days[15] and test the following alternative scenarios. For price c_t, initially we set it equal to the average price observed in the market. For the HP marketing activities

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

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

and content creator events, we test six alternative situations: maintaining the same level of events, or increasing them by 20%, 40%, 60%, 80%, or 100%. This analysis creates a grid of $6 \times 6 = 36$ cases to measure profits.

Information about the costs in Equation 20 is provided to us by HP.[16] 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 $150 per event, for the current number of events. For the content creator events, we assume that a similar value would be a reasonable incentive to generate the proposed increases. In both cases, the cost functions are assumed to be convex, increasing in an exponential way with more events, given the need to hire more HP workers in order to generate increased levels of promotional activity online. This assumption can be changed to match the cost structure of the company performing the analysis. We illustrate the results in Figure 5.

The figure shows the level of purchases, creations, and profits, for the alternative cases previously described. We find that both the creation and purchase of content are sensitive to increases in marketing activities originated by either content creators or HP, with slightly higher response to the events created by HP. In terms of profits, 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 60%, and increasing own events by about 40%. Increasing the number of content creator events would lead to more referrals, which would impact the image that potential visitors have of the platform, and likely alter the level of matching of preferences about content in the platform of consumers targeted by these events. 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.

7.2 The Effect of Public Relations and Price Promotions

As previously mentioned, MagCloud benefited from press coverage done by the New York Times at the end of March of 2009. The New York Times "event" involved the publication of an article (and companion slide show) about MagCloud and Do-it-Yourself (DIT) magazines in the Internet section of the New York Times website, and in the technology section of the print edition. The

[16] As previously, the values are scaled for privacy reasons.

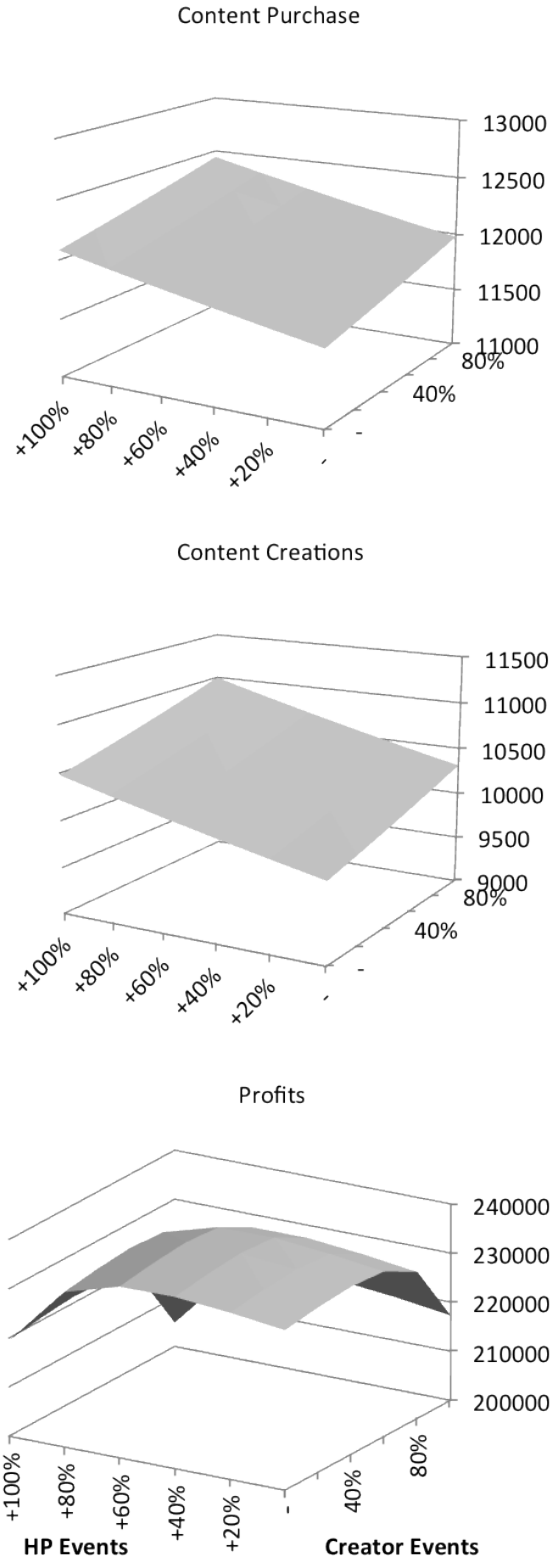

Figure 5: Number of purchases, creations, and profits for different investment scenarios

article happened at a time when there was a lot of ongoing press about the demise of traditional magazine publishing.

Articles like this in which large companies figure prominently are often not 100% developed by the independent firm, but instead corporate public relations departments continuously cultivate ideas with the media in the hopes something will get picked up. In this case, the HP public relations was contacted by a journalist, and a story was created about democratizing magazine publishing through new web services, such as HP's MagCloud service. The public relations department used existing MagCloud magazines, such as BARE (created by U.C. Berkeley students), to show how anyone could now create a magazine thanks to Web technology and high quality on-demand-printing. It is important to note that managers at HP state that the MagCloud business area did not devote any resources to make this happen other than answering questions, and HP Corporate public relations led the interactions with New York Times. HP did not get to see the article before publication, nor were they involved in the partner or customer interviews beyond providing the references. We thus can treat this event as exogenous of the firms pricing decisions and other marketing actions.

To illustrate the impact of the public relations event on the progression of visits and number of content created, we run a simple counterfactual, where the dummy for the New York Times event is set to zero. Figure 6 shows the results of the comparison between the actual and this counterfactual scenario. The top left panel shows the difference between the number of visits in the actual scenario and the number of visits in the counterfactual. We observe the large spike caused by the presence of the article for two days, of more than 100,000 for the first day and about 40,000 visits for the second, and then a permanent shift upwards, of about 8,000 daily visits. The top right panel shows the difference in the rate of creation and we find that it remains almost unchanged. Finally, the bottom panel shows the net effect of creations, considering simultaneously the change in visits and the rate of creation. We observe that this effect is positive and significant, with a permanent increase of 50 to 100 daily new uploads of content.

An alternative marketing action is to offer price promotions to either the content creators or purchasers. In our data set, we observe the occurrence of a 20% discount offered to consumers that ordered a magazine, during three weeks. This results in a 5 cent discount per page printed, and we already discussed that this promotion had a positive effect on the orders received by the platform, given the estimated positive coefficient. We apply our model to quantify this price promotion

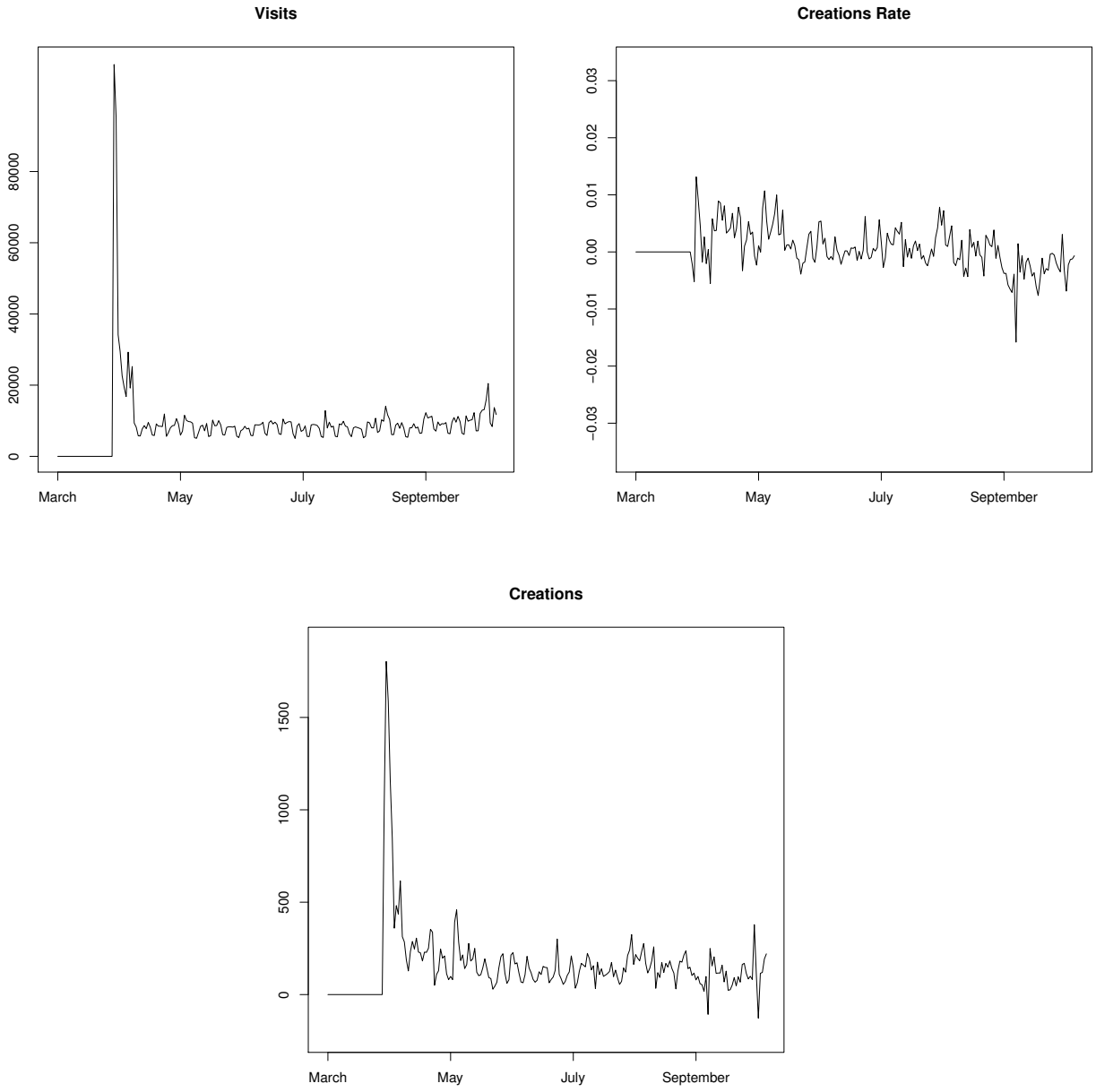

Figure 6: Impact of the New York Times event on content creation and visits

and then compare it to a public relations event of similar magnitude to the one presented above. Our methodology involves comparing the estimated revenues of a scenario with and without these marketing activities. First, in our scenarios with the marketing activities, we include (1) either an hypothetical New York Times event happening at the beginning of September of 2009, or (2) the price promotion that started at the exact same time period, lasting three weeks.[17] These two base scenarios make the alternative marketing activities as comparable as possible, by choosing the same day as a starting point and using the duration and impact observed in our actual data. We compare the results from each of these base situations with the respective results from counterfactual situations where we (1) either do not include the press release from the New York Times, or (2) remove the price promotion. We then compute the difference in total number of purchases for base and counterfactual situations, since the beginning of each action until the end of the data periods.

We find that without the New York Times event, the number of orders would have gone down by 16.3%. In terms of the price promotion, without its occurrence, orders would be down by 4.8%. Thus, in terms of quantity of purchases, a public relations event is able to generate a three-fold larger impact than a price promotion. Additionally, in the case of the promotion, MagCloud is losing five cents per order during the period of time while the promotion is in effect. This increase in quantity and change in price charged makes the promotion have an effect of about 2% in total revenue, for the period of time while the promotion is going on and taking into account longer term effects due mostly to changes in behavior from past purchase or creation of content. This seems to indicate that events generated by HP public relations department are a more attractive investment for HP. However, public relations efforts require a team that consistently connects with the media or other companies, which increases costs not included in this analysis and it has a more risky outcome, since stories and articles about the platform do not always get picked up by outside firms or media. Taking into account this uncertainty, HP can then decide to invest in such marketing activities, only if the cost of sustaining a dedicated MagCloud team is less than implementing price promotions or other marketing activities with more certain outcomes.

[17] We do these by appropriately changing the dummy variables of each of these marketing activities to one during these time periods.

7.3 The Importance of Content Creators - Activities and Referrals

In this section, we quantify the importance of content creators to the platform business. We do this in two ways. First, we quantify the impact of the marketing events originated by content creators. Since our measure of content creator activity includes only the most important events that are captured by Google Alerts, our estimate is a conservative one. Second, we measure the importance of the segment of consumers coming to the platform by referrals on visits, purchases, and creations, and compare these numbers to the segment of consumers reaching the website through search engines.

Our analysis spans for two months after the data included in the estimation, and we project the evolution of the dependent variables of the platform based on different scenarios to obtain the results. To evaluate the impact of content creator events, we start by comparing a scenario where the number of events is at the actual level (about 4.8 events per month) with numbers from a hypothetical scenario where the events from content creators are reduced to zero. All other variables are set at similar values for the two scenarios. For the second case, we quantify the relative importance of two of the consumer segments defined in our model, the referral and the search segment, by computing two additional situations. At the beginning of the projected two months, we make the value of the three decisions - visits, creations, and purchases of content - be equal to zero, in turn, for each of these two segments. We then project the platform evolution for the remaining consumers by forward simulating the consumer decisions starting at the end of our observed data, using current estimates and the empirical distribution of the data. In other words, we are "turning off" the referral segment and measuring how the platform would perform without its presence, and similarly with the search segment. It is unlikely that any of the segments would ever be completed removed from access to the platform, but nonetheless, the two counterfactual situations allow us to compare the aggregate importance of the two segments at the current stage of development. The actual and counterfactual numbers are presented in Table 7, as well as the percent change between scenarios.

We find that, if content creators do not participate in marketing activities, MagCloud would observe losses of about 1.4% in both visits and content creation, and about 0.8% in purchases. We note that currently, the firm is not providing any incentives to content creators to develop their own

		Creator Events		Referral Segment		Search Segment	
	Base	No Events	Diff.	No Referrals	Diff.	No Search	Diff.
Visits	120,207	118,476	-1.4%	61,078	-49%	89,807	-25%
Creations	2,682	2,647	-1.4%	1832	-32%	1,745	-35%
Purchases	3,192	3,169	-0.8%	1,462	-54%	2,517	-21%

Table 7: Impact of free advertising and referrals from content creators

advertising of their magazines, and so, for Hewlett-Packard, this is free support for the platform. Comparing the importance of the two segments, we find that the referral segment is relatively more important with respect to visits and purchases, while the search segment plays a more important role with regards to content creation. Without the referral segment, which is dominated by the behavior of content creators, the platform would have about half of the current customers, with drops of 49% in visits, 32% in creations, and 54% in total purchases at the website. Losing the search segment, dominated by consumers coming from websites such as Google, Yahoo, or Bing, would be reflected in a decrease of about 25%, 35%, and 21% in the respective decisions. For MagCloud, the importance of a smaller referring community of magazine websites reveals itself to be more important than the role played by the most frequented search engines.

From this analysis, we conclude that content creators play an essential role in this platform, either through their own activities or through links and referrals to the platform's website. Given the high importance of content creators in HP's business, combined with our previous finding that it is worthwhile providing monetary incentives to increase their marketing efforts, we believe that managers at MagCloud, and very likely at other similar platforms of user-generated content, should concentrate a significant percentage of marketing investments developing this side of the market to obtain a faster growth of the platform and increase long-term profitability.

8 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 multiple interactions both between the two sides of the market, as well as within user. We measure the impact of a multitude of marketing actions: price promotions, blogs and online events from HP and from content creators, and public relations motivated events, such as an

article on an independent source like the New York Times.

Empirically, we show the wide usefulness of our model using two data sets from the self-publishing magazine site MagCloud, created by HP. The data sets differ mostly on the level of aggregation, but contain similar information about the decisions of users of the online platform of user-generated content. We use the presence of alternative online advertising activities to study the relation between the multifaceted demand of an online intermediary in the market of user-generated content and the different types of marketing activities. We demonstrate that our demand model, based on individual utility maximization, is able to capture the critical elements of the multi-sided platform's demand and predict future demand with adequate accuracy. Nevertheless, our model has limitations. For example, we do not specifically model the process of creating content and assume that content creators as myopic. It is possible that in other two-sided markets of user-generated content, creators go through a lengthy or complex decision process and are motivated to wait before publishing content until the platform has reached a larger size. This would potentially increase the importance of earlier marketing activities, to increase the motivation of consumers to join the platform sooner and increase future growth.

We show that there is significant heterogeneity across consumers, with returning visitors becoming the majority of creators of content, which then refer considerable number of purchasers. 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. We find that applying our model to an individual-level data set can provide additional insights about the impact of heterogeneity resulting from previous decisions on current utility and provide the possibility of individual targeting.

Finally, our in-depth analysis of the relation between a wide range of marketing activities and consumer actions at the user-generated content platform sheds some light on a number of managerial questions. We provide recommendations on how to allocate resources across marketing tools. We offer evidence that marketing actions originated by content creators and their referrals play an essential part in the development of a user-generated website and we are able quantify the results from this free advertising to the firm. Additionally, we found that a particular event originated by the HP public relations department, involving the press coverage of the platform in the New York Times, was worth about three times as much as a monthly price cut of 20% in terms of content purchases. Our model can be used to evaluate the implementation of other hypothetical scenarios,

with the ability to provide recommendations on investments on marketing activities that impact the profitability of user-generated content platforms.

References

[1] Argentesi, E., Filistrucchi, L. (2007), "Estimating Market Power In A Two-Sided Market: The Case Of Newspapers", *Journal Of Applied Econometrics*, 22, 1247-1266.

[2] Armstrong, M., (2006), "Competition in two-sided markets," *Rand Journal of Economics*, 37, 668- 691.

[3] Audit Bureau of Circulations (2010), http://www.acessabc.com.

[4] Baxter, W. (1983), "Bank Interchange of Transactional Paper: Legal and Economic Perspectives," *Journal of Law & Economics*, 26: 541:588.

[5] Berry, S., Waldfogel, J. (1999), "Free Entry And Social Inefficiency In Radio Broadcasting", *The RAND Journal Of Economics*, Vol. 30, No. 3, 397-420.

[6] 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.

[7] 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.

[8] 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.

[9] Clifton, B. (2008), "Increasing Accuracy For Online Business Growth", Whitepaper, Omega Digital Media.

[10] Comscore (2007), *"*Retail E-Commerce Climbs 23% in Q2 Versus Year Ago*"*, July 30, Press Release.

[11] Drezner, Z and G.O. Wesolowsky (1989), "On the computation of the bivariate normal integral," *Journal of Statistical Computation and Simulation*, 35, pp. 101-107

[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] Genz, A., F. Bretz (2009), "Computation of Multivariate Normal and t Probabilities", *Lecture Notes in Statistics*, Vol. 195., Springer-Verlage.

[16] 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.

[17] Ghose, A., S. Han (2010a), "An Empirical Analysis of User Content Generation and Usage Behavior in the Mobile Internet," Working Paper.

[18] Ghose, A., S. Han (2010b), "A Dynamic Structural Model of User Learning in Mobile Media Content," Working Paper.

[19] Gupta, S., Steenburgh, T. (2008), "Allocating Marketing Resources," *Harvard Business School Marketing Research Paper* ,08-069.

[20] Godes D., Mayzlin D. (2009), "Firm-Created Word-of-Mouth Communication: Evidence from a Field Test", *Marketing Science*, Vol. 28, No. 4, p721.

[21] Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System", *Econometrica*, Vol. 46, No. 4, 931-959.

[22] Horsky, D. (1990), "A Diffusion Model Incorporating Product Benefits, Price, Income and Information", *Marketing Science*, Vol. 9, No. 4, Autumn, 342-365.

[23] Ilfeld, J.S., Winer, R., "Generating Website Traffic", *Journal of Advertising Research*, Vol. 2, Issue 5, Sep/Oct 2002.

[24] 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.

[25] Kannan P. K., Kline Pope B., Jain S. (2009), "Pricing Digital Content Product Lines: A Model and Application for the National Academies Press", *Marketing Science*, Vol. 28, No. 4, p620.

[26] Kim, Jun B., Paulo Albuquerque, and Bart J. Bronnenberg (2010), "Online Demand under Limited Consumer Search," *Marketing Science*, Vol. 29, No. 6, November-December, pp. 1001-1023.

[27] Lerner, J. and J. Tirole (2002), "Some Simple Economics of Open Source," *Journal of Industrial Economics*, 52 (June), 197-234.

[28] Maddala, G. S. (1983), Limited Dependent and Qualitative Variables in Econometrics, New York: Cambridge University Press.

[29] Manchanda, P., Y. Xie and N. Youn (2008), "The Role of Targeted Communication and Contagion in Product Adoption," *Marketing Science*, 27(6), 961-976.

[30] Manski, C. (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, Vol. 60, No. 3. (Jul., 1993), pp. 531-542.

[31] Mantrala, M. (2006), Allocating Marketing Resources, *Handbook of Marketing*, Ed. by Weitz B. and Wensley R., SAGE Publications Ltd.

[32] Moe, W., Fader, P. (2004a), "Capturing Evolving Visit Behavior in Clickstream Data", *Journal of Interactive Marketing*, Vol. 18, No. 1, Winter, 5-19.

[33] Moe, W., Fader, P. (2004b), "Dynamic Conversion Behavior at E-Commerce Sites", *Management Science*, Vol. 50, No. 3, March, 326-335.

[34] Nair, H., P. Manchanda, and T. Bhatia (2010), "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research*, Vol. 47, Issue 5, October, 883-895.

[35] 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.

[36] 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.

[37] Rochet, J. and J. Tirole (2005), "Two-Sided Markets : A Progress Report", IDEI Working Papers 275, Institut d'Économie Industrielle (IDEI), Toulouse.

[38] 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.

[39] Rysman, M. (2004), "Competition Between Networks: A Study of the Market for Yellow Pages", *Review of Economic Studies*, 71, 483-512.

[40] 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..

[41] Song, I., Chintagunta, P. (2004), "A Micromodel of New Product Adoption with Heterogeneous and Forward-Looking Consumers", *Quantitative Marketing & Economics*, 1, 4, 371-407.

[42] Van den Bulte, Christophe and Gary L. Lilien (2001), "Medical Innovation Revisited: Social Contagion versus Marketing Effort", *American Journal of Sociology* , 106 (5), 1409-35.

[43] Wilbur, K., (2008), "A Two-Sided, Empirical Model of Television Advertising and Viewing Markets", *Marketing Science*, Vol. 27, No. 3, May-June, 356-378.

[44] Yao, S. and C. Mela (2009), "Online Auction Demand," *Marketing Science,* Vol. 27, No. 5, September-October, pp. 861-885.

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

[46] Zhang, X., Zhu F. (2011), "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia", *American Economic Review*, forthcoming.