Incorporating Profit Margins into Recommender Systems: A Randomized Field Experiment of Purchasing Behavior and Consumer Trust

Umberto Panniello, Michele Gorgoglione, Shawndra Hill and Kartik Hosanagar

Abstract

A number of recent studies have proposed new recommender designs that incorporate firm-centric measures (e.g., the profit margins of products) along with consumer-centric measures (e.g., relevance of recommended products). These designs seek to maximize the long-term profits from recommender deployment without compromising customer trust. However, very little is known about how consumers might respond to recommender algorithms that account for product profitability. We tested the impact of deploying a profit-based recommender on its precision and usage, as well as customer purchasing and trust, with data from an online randomized field experiment. We found that the profit-based algorithm, despite potential concerns about its negative impact on consumers, is effective in retaining consumers' usage and purchase levels at the same rate as a content-based recommender. We also found that the profit-based algorithm generated higher profits for the firm. Further, to measure trust, we issued a post-experiment survey to participants in the experiment; we found there were no significant differences in trust across treatment. We related the survey results to the accuracy and diversity of recommendations and found that accuracy and diversity were both positively and significantly related to trust. The study has broader implications for firms using recommenders as a marketing tool, in that the approach successfully addresses the relevance-profit tradeoff in a real-world context.

Key words: recommender systems, profit-based recommenders, content-based, trust, personalization

1. Introduction

Recommender systems attempt to predict items of interest to their users based on information about the items and users. They are widely used in online retail by major firms such as Amazon, Wal-Mart, and Netflix, and they are known to exert a significant influence on consumer choice (Fleder et al. 2010). Recommender systems offer benefits to both consumers and firms. They help consumers become aware of new products and help them select desirable products from a myriad of choices (Pham & Healey, 2005). Recommender systems have the potential to help firms increase profits by converting browsers into buyers, allowing cross-selling of products, and increasing loyalty (Schafer et al., 1999).

The vast majority of researchers who have studied recommender systems have evaluated designs using consumer-centric measures, such as relevance of recommended items. Although these designs consider user satisfaction, researchers only imply that firms deploying these systems benefit from increased customer satisfaction and higher purchasing rates from users. However, a stream of papers in recent years (e.g., Bodapati, 2008; Hosanagar et al., 2008; Das et al., 2010) have suggested that firms can do better by combining firm-centric measures, such as profit margins of products, with consumer-centric measures, such as relevance of recommended products. These designs seek to help those using recommender systems to maximize long-term profits without compromising customer trust. However, it is unreasonable to expect consumer behavior to remain the same if recommenders are modified to incorporate firm-centric measures such as profit margins. In particular, if consumers perceive that recommendations are even slightly biased, they may not pay as much attention to recommendations or may distrust recommendations altogether. In such a scenario, the net impact on consumer trust and firm profits may even be negative.

Although the issue of consumer response to biases in recommenders is a relatively new one, related issues have come up in the context of other information filters. For example, even though the current business models for search engines rely on providing sponsored results alongside organic results, the founders of Google had famously stated that "we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers … search engine bias is particularly insidious (and) … we believe the issue of advertising causes enough mixed incentives" (Brin and Page 1998). Although search engines, including Google, eventually incorporated sponsored results without a drastic impact on consumer trust or usage, it is not clear if the experience of search engines will apply to recommenders and, hence, a similar question has arisen in the context of recommenders. On the one hand, a number of research papers (e.g. Bodapati 2008; Chen et al. 2008) have advocated incorporating firm-specific measures into recommender design and, consistent with that, several commercial recommenders do bias recommendations in practice.[1] However, other academics and practitioners have argued that manipulations and biases will hurt recommender credibility and erode consumer trust, leading to a net reduction in firm profits (Resnick and Varian 1997; Simonson 2003). There have been no controlled experiments evaluating these designs in the field or the lab, so very little is known about how consumers might in fact respond to recommender algorithms that account for product profitability. In this paper, we help reconcile the opposing views and report the results of a randomized field experiment testing the impact of deploying a profitability-based recommender design on website usage, purchasing behavior, and trust.

Our field experiment randomly assigned users of an online comic store to three different types of treatments: random recommendations, content-based recommendations (based only on item relevance), and a profit-based recommender system. We found, as expected, that a simple content-based recommender is able to drive a meaningful increase in consumers' usage and purchase volume relative to a control group receiving randomized recommendations. We also found that the profit-based algorithm, despite concerns about its potential negative impact on consumers, generated higher

[1] Recommenders licensed by vendors such as Oracle and Adobe include features that allow deploying firms to assign a higher weight to high margin items.

profits for the firm than the content-based recommender. In a post-experiment survey of users, we did not find any significant reduction in their trust in the profit-based recommender. Furthermore, we found that our results can be explained in terms of the accuracy and diversity of the recommendations.

The study makes several key contributions to the literature. First, the paper provides the first real-world empirical evidence regarding the impact of profit-based recommender designs. Second, a firm's choice of recommender design and the response of its users are typically endogenous, making the issue of identification challenging. Existing work on profit-based designs have evaluated their impact assuming user behavior remains the same under these new designs, which ignores the endogeneity in design and response. As a result, the true impact of these designs has not been identified. We addressed the issue using a randomized field trial with a real-world online comic bookstore. Further, to the best of our knowledge, this is the first study on recommender algorithms to report both purchase and trust measures from the field. These measures not only allowed us to evaluate whether the new designs indeed drove the purchase outcomes but also allowed us to directly address concerns regarding consumer distrust of the proposed designs.

2. Literature Review

Recommenders are widely deployed on the Internet by media firms and retail websites. Recent research has shown that recommender systems offer significant value to consumers and firms. Recommenders help consumers become aware of new products and help them select relevant products from a myriad of choices (Pham & Healey, 2005). Recommenders help firms increase sales by converting browsers into buyers and increasing their ability to cross-sell products and retain customers (Schafer et al., 2001). Dias et al. (2008) measured the business value of a personalized recommender system, showing that its effect extends beyond the direct revenues generated from the purchase of recommended items and that it generates substantial additional revenues by introducing shoppers to new categories from which they then continue to purchase. Ansari and Mela (2003) studied how to customize the design and content of e-mail marketing and found that the consumer response rate could be increased by 62% if the e-mail's design and content are customized.

Given the potential impact of personalized recommendations for both firms and consumers, the issue of recommender design and its impact on consumer choice has been a topic of much interest. In the subsections that follow, we describe prior work on recommender design and consumer choice.

2.1 Design of Recommender Systems

There are many studies in the literature in several fields, including computer science, information systems, and marketing, which have looked at the problem of designing effective recommender systems that can infer user preference and recommend relevant items. A survey of various approaches was provided by Adomavicius and Tuzhilin (2005), who classified these systems into content-based, collaborative filtering, and hybrid approaches. In content-based systems, items that have a high degree of similarity to users' preferred items (inferred through ratings or purchases) are recommended

(Mooney & Roy, 1999; Pazzani & Billsus, 2007). An advantage of using content-based designs is that even a small set of users can be addressed effectively. As highlighted by Balabanovic and Shoham (1997) and Shardanand and Maes (1995), a major limitation of content-based methods is that one must be able to parse items using a machine, or their attributes must be assigned to items manually. Unlike content-based recommendation methods, collaborative filtering systems recommend items based on historical information drawn from other users with similar preferences (Breese et al., 1998). Collaborative recommender systems do not suffer from some of the limitations of content-based systems; in fact, since collaborative systems use other users' recommendations (ratings), they can deal with any kind of content and even recommend items from product categories other than the ones rated or purchased by a user. However, collaborative filtering suffers from the "new item problem", namely the difficulty of generating recommendations for items which have never been rated by users. Both collaborative filtering and content-based systems suffer from the "new user problem", namely the difficulty of generating meaningful recommendations for users who have never expressed any preference (Balabanovic & Shoham, 1997; Lee, 2001). Finally, those who use hybrid approaches are trying to avoid the limitations of content-based and collaborative systems by combining collaborative and content-based methods in different ways (Claypool et al., 1999; Nicholas & Nicholas, 1999).

Researchers have also studied how to include other information besides customers' demographic data, past purchases and past product ratings, in order to improve the accuracy of recommendations. In particular, Bettman et al. (1998) demonstrated that context induces important changes in a customer's purchasing behavior. Other experimental research suggested that including context in a user model in some cases improves the ability to predict behavior (Palmisano et al., 2008). Adomavicius et al. (2005) described a way to incorporate contextual information into recommender systems by using a multidimensional approach in which the traditional two-dimensional (2D) user/item paradigm was extended to support additional contextual dimensions, such as time and location. Similarly, Ansari et al. (2000) studied how to use expert evaluations in addition to traditional users' evaluations and item characteristics.

An interesting research stream has studied how recommender designs should be altered to increase diversity of recommendations. McGinty and Smyth (2003) investigated the importance of diversity as an additional item selection criterion and demonstrated that the gains can be significant, but also show that its introduction has to be carefully tuned. Fleder and Hosanagar (2009) demonstrated that recommender systems that discount item popularity in the selection of recommendable items may increase sales more than recommender systems that do not. Adomavicius and Kwon (2010) showed that while ranking recommendations according to the predicted rating values provides good predictive accuracy, such a system provides poor performance with respect to recommendation diversity. Therefore, they proposed a number of recommendation ranking techniques that can provide significant improvements in recommendation diversity with only a small amount of loss of accuracy.

In recent years, a number of studies have attempted to incorporate firm-specific measures into the item selection process and have attempted to place the issue of recommender design within a profit-maximizing context. This stream of research is particularly important, in our opinion, because recommender systems were conceived as a tool to help consumers select relevant information when browsing the Web but soon became a tool for improving the effectiveness of companies' marketing actions. Bodapati (2008) studied the relevance-profitability tradeoff in recommender design. The author modeled recommendations as marketing actions that can modify customers' buying behavior relative to what they would do without such an intervention. He argued that if a recommender system suggests items that are most relevant, it may be of little value if those items might eventually be bought by consumers in the absence of recommendations. He showed that the system should recommend items with a purchase probability that can be best influenced by the recommendation, instead of recommending a product that is most likely to be purchased. Chen et al. (2008) integrated the profitability factor into a traditional recommender system and compared four different systems (obtained by using a personalized/non-personalized system and by including and not including profitability). They showed that including profitability increases the cross-selling effect and revenues and that it does not cause recommendation accuracy to drop. Hosanagar et al. (2008) also investigated how to recommend products to help firms increase profits rather than recommend what is most likely to be purchased. The authors identified the conditions under which a profit-maximizing recommender system should suggest an item with the highest margin and those under which it should recommend the most relevant item. This paper highlighted two main tradeoffs a company faces in designing a profit-enhancing recommendation policy. The first tradeoff is between a product's purchase probability and the firm's margins from selling that product. The second tradeoff is between increasing near-term profit versus maintaining consumer trust to increase future profits. Similarly, Das et al. (2010) developed a model that uses the output of a traditional recommender system and adjusts it based on item profitability. The authors applied a model of consumer response behavior to show that their proposed design can achieve higher profits than traditional recommenders. Other studies on profit-based recommenders include works by Akoglu and Faloutsos (2010), Brand (2005), and Iwata et al. (2008).

2.2 Users' Response to Recommender Systems

Several researchers have studied how users respond to personalized recommendations, focusing on factors that influence perceived usefulness and ease of use, trust, and satisfaction. Pavlou (2003) and Gefen et al. (2003) integrated trust, risk, and perceived usefulness and ease of use and empirically confirmed the links from trust to perceived usefulness and adoption intention. Wang and Benbasat (2005) extended the integrated model to online recommender adoption and demonstrated the link between trust and adoption intention. Liang et al. (2007) demonstrated that both the number of items recommended to the user and the recommendation accuracy, as measured by the number of recommended items accepted by the user, had a significant effect on user satisfaction. Bharati and

Chaudhury (2004) showed that the recommender's information quality (i.e., relevance, accuracy, completeness, and timeliness) had a significant effect on users' decision-making satisfaction.

An interesting research stream includes studies of how the diversity and familiarity of recommendations can affect the effectiveness of recommender systems. Most researchers agreed that consumers generally prefer more variety when given a choice (Baumol and Ide 1956; Kahn and Lehmann, 1991). McGinty and Smyth (2003) empirically demonstrated that there may be significant gains from introducing diversity into the recommendation process. Simonson (2003) proposed that the purchase type and degree of variety-seeking affect customers' acceptance of recommended "customized" offers; in particular, higher rates of variety-seeking decrease a consumer's receptivity to customized offers. In addition, several researchers (Broniarczyk et al., 1998; Dreze et al., 1994; , Hoch et al., 1999, van Herpen & Pieters, 2002) showed that consumers' perception of variety can be influenced not only by the number of distinct products offered but also by other features (such as the repetition frequency, organization of the display, and attribute differences). Cooke et al. (2002) studied how customers respond to recommendations of unfamiliar products. Their analysis demonstrated that unfamiliar recommendations lowered users' evaluations but additional recommendations of familiar products serve as a context within which unfamiliar recommendations are evaluated. Further, additional information about a new product can increase the attractiveness of an unfamiliar recommendation. Xiao and Benbasat (2007) also showed that familiar recommendations increase users' trust in the recommender system, and recommender systems should present unfamiliar recommendations in the context of familiar ones. They go on to show that the balance between familiar and unfamiliar (or new) product recommendations influences users' trust in, perceived usefulness of, and satisfaction with recommender systems. Several other researchers (Komiak and Benbasat, 2006; Sinha & Swearingen, 2001; Swearingen & Sinha, 2001) showed that familiar recommendations play an important role in establishing user trust in a recommender system. Further, a user's trust in a recommender system increases when the recommender provides detailed product information.

Gershoff et al. (2003) showed that higher rates of agreement led to greater confidence in a system and a greater likelihood of a user accepting a recommender's advice; Hess et al. (2005) showed that a high similarity between users and the recommendations contributed to an increased involvement with the recommendations, which in turn resulted in increased user satisfaction with the recommendations. The similarity in attribute weighting between users and a recommender system has a significant impact on users' perceptions of the utility of the recommendations generated by the system (Aksoy & Bloom, 2001).

In addition, Simonson (2003) proposed that customers are more likely to accept recommendations to choose a higher-priced, higher-quality option than a lower-priced, lower-quality option and that this tendency is negatively correlated with the level of the customer's trust in the marketer. In addition, customer preferences are often constructed rather than revealed, and this practice has important implications with respect to the effectiveness of customizing offers to match individual tastes. This theory is also confirmed by psychological studies (Payne, 2003) showing that customers do not exactly know their preferences when confronted with a set of product alternatives. On the contrary, preferences are constructed while learning about the choices offered.

In summary, the discussion on recommender design and consumer choice reveals that (i) there is considerable recent interest in placing recommender design in a marketing context, and (ii) while the impact of recommendation diversity, familiarity, similarity and size on purchasing behavior and trust are known, we know very little about the impact of recommendation biases on these outcomes. Specifically, all the papers on profit-maximizing recommenders develop theoretical models of consumer behavior and evaluate their systems within the framework of their consumer model. It is not clear whether the proposed designs offer the suggested benefits in the field. For example, Hosanagar et al. (2008) did not provide any empirical evidence about the effectiveness of their models whereas, as suggested by the authors themselves, the issue calls for an empirical approach. Further, several researchers have warned that pursuing a firm’s goal of increasing profits when making recommendations can decrease consumer trust in these systems (Das et al., 2010; Resnick and Varian 1997). Simonson (2003) warned that a “learning relationship” that is used as a basis for producing customized offers is perceived by customers as an indicator of good service but also as a restriction of their freedom of choice and can be a source of discomfort. In particular, customers may also perceive a marketer’s personalization efforts as attempts to persuade and manipulate. Garfinkel et al. (2007) stated that the credibility of recommender systems is also an important factor in determining the strength of the impact of recommendations on sales and that recommendations can influence shoppers’ decisions only when they are perceived to be objective and credible. The authors suggest that since retailers have full control of what recommendations to make and how to present them, it is natural for shoppers to discount the credibility of online recommender systems because of potential manipulation by retailers. Indeed, in another environment, Chen et al. (2008) showed that user satisfaction suffers when ratings are manipulated, probably due to lower accuracy, and that users can detect systems that manipulate ratings. As a result, it is important to empirically investigate how the use of profit-based design affects a firm’s profits and consumers’ trust in these systems.

Our study’s objective is to fill this key gap in the literature. To this end, we deployed a recommender algorithm in the field, based on the model by Hosanagar et al. (2008) and compared its performance to a more traditional algorithm. In addition to evaluating the impact on consumer purchasing behavior and the deploying firm’s profits, we also studied the effect of the recommender design on trust. The contributions of this work cover both the streams on recommender design and the users’ response to recommendations, thus providing input to both Information Systems and Marketing fields.

3. Research Methodology

We conducted the randomized experiment in a real-world setting in partnership with an online firm. The company involved in the experiment is a very well-known, medium-sized Italian firm operating in the publishing industry. The company's Web division mainly sells comic books and related products, such as DVDs, stickers, and t-shirts. As part of its normal business, the company sends a weekly non-personalized newsletter to approximately 23,000 customers and planned to send personalized notices recommending comic books via e-mail to the same customer base. We partnered with the firm to augment its evaluation of appropriate recommendation strategies while answering our research question: *Do firm-centric recommendations impact performance (profitability) and/or trust?*

The study was conducted in two stages. First, we invited extant customers of the comic site to a randomized field experiment. Second, upon completion of the experiment, we helped the firm survey the participants regarding their trust in the particular recommendation engine to which they were assigned. We describe both of these stages below.

3.1 Field Experiment

3.1.1 Experimental Design

At the beginning of the experiment, the firm asked customers via e-mail if they wanted to participate in the project. The study was presented as a "collaboration between the firm and the [university name] to improve the newsletter service to customers." The firm continued to send its traditional weekly newsletter and provided our personalized newsletter as an additional service. The final number of customers involved in the experiment was 260, corresponding to a participation rate slightly higher than 1%. The experiment participants were then randomized to three experimental treatment conditions: random recommendations, content-based recommendations (based only on item relevance), and a profit-based recommender system. The items recommended were comic books sold by the online store of the publishing company.

Each subject who participated in our study received 10 comic book recommendations per week via a weekly e-mail newsletter. The experiment ran for 10 consecutive weeks. The newsletter was similar to the type of correspondence the customers were accustomed to receiving from the firm prior to our study.

The first newsletter asked the users to rate a representative set of comics to be used to gauge the users' preferences. That is, the comics in the first newsletter were the same for everyone and were selected by the firm. These comics were both diverse in terms of content category and popularity. We considered the first newsletter to be a pre-experiment questionnaire needed to manage the "cold start problem." Only 5% of the customers made more than one order per year, which made it impossible to create a meaningful profile from past purchases.

After the pre-experiment, the weekly e-mails contained a link to a personal recommendation page. The page was composed of two parts, one containing "recommended brand new items" (selected from brand new arrivals at the firm) and one containing the "recommended old items" (selected from arrivals in the past two months). A total of 10 recommendations were made per newsletter. Each recommended product was presented to include the following information: title, cover image, description, and a "see more details" link. Below the product description, a question was presented asking the customer to rate the recommended product by clicking an opinion on a (1-5) point scale. The feedback was returned to the server and used to update the user profiles. The direct feedback was the rating given to each product on the recommendation page. The indirect feedback was the tracked "click" data on the link, representing the customer's interest in a product on the recommendation page.

The content of the last 9 newsletters in the experiment was determined for each individual based on his or her historical ratings of comics, and the users were asked each time to rate the comics. Past ratings did not matter for the group that received random recommendations. On the other hand, content-based and profit-maximizing recommendations considered an individual's past ratings. We next discuss the three recommendation engines.

3.1.2 Recommendation Engines

The experiment included three recommendation treatments: a content-based recommender, a profit-based recommender, and a random recommender. We used the content-based recommender as a "benchmark," and we chose a content-based recommendation (as opposed to collaborative filtering) because the experiment was carried out with a relatively low number of participants. Given the sparsity of the user/item matrix, it would be very difficult to generate meaningful recommendations by using a collaborative engine. The profit-maximizing recommender system is based on the algorithm presented by Hosanagar et al. (2008). In selecting specific content-based and profit-based algorithms, we note that our goal was neither to propose a new algorithm nor to test every possible recommender design. Instead, we chose reasonable designs that are easy to understand and implement and that have previously been proposed in the literature.

3.1.2.1 Content-based.

The content-based algorithm simply recommends items based on the description of the new items (title, sub-title, and description) and the user profile, which includes items that have been rated high in the first pre-experiment survey as well as items with high ratings as the experiment progressed. In our study, the descriptions of items are treated as a weighted vector representing a weighted bag of words.

Let *ItemProfile(s)* for item *s* and *UserProfile(i)* for user *i,* be two vectors representing the item characteristics and the customer preference, respectively. *ItemProfile(s)* is computed by extracting a set of keywords from a description of item *s*. The description simply describes the item and its contents, including author and publisher details. *UserProfile(i)* is computed by analyzing the content of the items previously seen and rated by user *i*. In particular, the vector is defined as a vector of weights (w_{i1}, …, w_{ik}), where each w_{ik} denotes the importance of keyword *k* to user *i*. We computed w_{ik}

as an “average” of the ratings provided by user *i* to those items that contained the keyword $k \in$ K. For example, suppose that user *i* rates two comic books, and her ratings for these books are “5” and “1,” respectively. If the descriptions of both comic books contain the word “romantic,” the weight for the word “romantic” for that user is 3.

We computed the relevance *u(i, s)* of item *s* for user *i* by matching the *UserProfile(i)* and the *ItemProfile(s)*. The following score was computed:

$$u(i,s) = \frac{\sum_{k} w_{i,s,k}}{k} \qquad (1)$$

where $w_{i,s,k}$ are weights of the words in common between the *UserProfile(i)* and the *ItemProfile(s)*, and *k* is the total length of profiles. The top 10 items with the highest score were presented to the user in the newsletter. This algorithm was based on the content-based recommender engine described by Pazzani & Billsus (2007).

Since we adopted a content-based engine, which uses item features, we checked to ensure that each item had the same amount of information (i.e., title, sub-title, and description) in order to avoid the introduction of any bias (e.g., recommending items with long descriptions more than items with short descriptions or the converse). To this end, we restricted the vector length in the profiles to 80 words.

3.1.2.2 Profit-based.

The key design goal of the profit-based recommender is to take the relevance measures from the content-based system and complement it with information about their profitability before making a recommendation. To do so, we operationalized the key insights from the model proposed in the paper by Hosanagar et al. (2008). Although several other researchers have developed heuristics-driven profit-based recommender algorithms, we chose to build on Hosanagar et al.’s model because their recommendation strategies are derived from a clear theoretical model and their setting is closest to our application context. We first describe the key features of their model and then describe the manner in which the insights were adopted in our setting.

Each recommended item *s* has a profit margin *M(s)* and an expected relevance *u(i, s)* to customer *i*. In each period, the firm shows a set of recommendations to the customer. The purchase probability of recommended items is temporarily boosted by the recommendation (Hosanagar et al. refer to it as the salience effect). The magnitude of the boost depends on the consumer’s trust in the recommender, modeled through a state variable *S*. This state *S* reflects the reputation of the recommender with the customer, and if a recommender system recommends the “right” products, then the recommendations influence consumer choice to a greater extent than would be the case if it often recommends irrelevant products. Hosanagar et al. (2008) modeled two states: *H* and *L* (high reputation and low reputation, respectively). The salience of the recommendation differs across these states: $H > L > 0$. According to Hosanagar et al. (2008), the customer’s state is determined by the past performance of the recommender, i.e., poor past recommendations lead to low trust (*L*) and good past recommendations increase trust (*H*). In our study, we proxied a customer's current state by measuring the average rating provided by the customer during his last visit to the recommendation page.

According to their model, if a customer is in a state *H* in a period and in that period purchases at least one recommended product or if the average rating provided by the customer in that period is over a threshold (e.g., 2.5 on a 0-5 scale), then he remains in the state *H* in the next period. If he does not purchase any of the recommended items and the average rating is below the threshold, then he will transition to state *L* with probability p_1 in the following period. If a customer is in state *L* and in that period does not purchase a recommended product and if the average rating provided the last time is below the threshold, then he remains in the state *L* in the next period. But if he purchases one of the recommended products or if the average rating is above the threshold, then he will transition to state *H* with probability p_2 in the following period. Hosanagar et al. (2008) show that, based on the customer's state, the recommender can determine whether to recommend items with the highest expected profit, defined by the items' purchase probability times margin, or the most relevant items.

We based our implementation on the above model. Given the fact that the company had never used a recommender system before, there was no historical data from which to compute the probabilities of transition. Moreover, the company wanted to limit the intrusiveness of the experiment so we did not run real-time surveys to determine customers' trust at each recommendation instance. Therefore, we adjusted the model described above as follows:

- If the average rating provided by the customer after reading the last newsletter is above a threshold (2.5 on a 0-5 scale) or if the customer purchased at least one recommended item the last time (i.e., last week), then assume state *H* and recommend the item with the highest $M(s)*u(i, s)_j$.
- If the average rating provided by the customer the last time is below the threshold and he did not purchase the recommended item last time, then assume state *L* and recommend the item with the highest $u(i, s)$.
- If a customer has not received any recommendations previously, then toss a fair coin to recommend either the highest $u(i, s)$ or highest $M(s)*u(i, s)$.

In this approach, as stated above, $u(i, s)$ is the expected relevance of the product *s* to the customer and is measured using the content-based recommendation engine, while $M(s)$ is the profit margin of the product *s* and is provided by the firm. In other words, the relevance calculation across our two treatments is done the same way making it meaningful to compare the treatments. The profit-based treatment differs only in its accounting of the profit margin but otherwise uses the same relevance measures.

3.1.2.3 Random.

Unlike the content-based and profit-based approaches, the random approach does not take the user profile into consideration when recommending new products. Instead it randomly selects, without replacement, a set of items to recommend from the products that have not been recommended or purchased in the past.

3.1.3 Performance metrics.

We were able to obtain direct feedback from the participants in the form of ratings (scale 1- 5) assigned to each product on the recommendation page. The second direct feedback was related to purchases made by customers after viewing the recommendation page, as the company provided us with access to the purchase data. In addition, we were able to measure indirect feedback by tracking when users clicked on the "more information" link for a particular comic, representing the customer's interest in a product on the recommendation page. We used these forms of feedback to compare performance of the recommendation systems in terms of accuracy, response rate, and profit.

3.1.3.1 Accuracy.

Referring to customer's explicit ratings, we considered traditional accuracy metrics used in the recommender system literature (e.g., precision, recall, and F-measure). However, in this case, we could only compute precision since it is not possible to know the ratings of the unseen items needed to compute the recall and F-measure. The precision *P* measures the fraction of recommendations that are labeled as "good recommendations" (Herlocker et al. 2004). Precision is computed as:

$$P = \frac{N_{rs}}{N_s} \qquad (2)$$

where N_s is the number of the items recommended to the customer ("selected" by the recommender system as items to be recommended), and N_{rs} is the number of items that proved to be "relevant" for the customer among those selected by the recommender system ("good recommendations"). In our application, we considered an item "relevant" or a "good recommendation" if it was rated as 3, 4 or 5, meaning that the customer liked the recommended item. In addition, we measured the average rating provided by users of recommended products over time, as follows:

$$Average\ rating_{z,u} = \frac{\sum_n rating_{n,z,u}}{n} \qquad (3)$$

where $Average\ rating_{z,u}$ is the average rating provided by user u in period z, $rating_{n,z,u}$ is the rating provided by user u to item n in period z, and n is the total number of items rated by u in period z. We also measured the overall average rating over time, the percentage of users with average rating over 3, and the percentage of users with an average rating less than or equal to 3 as additional measures of recommendation accuracy.

3.1.3.2 Response rate

The response rate is simply the ratio of users that respond to the total number of customers invited to participate. We measured the response rate over time for each of the three treatment groups as follows:

$$Re\,sponse\ Rate_z = \frac{Responders_z}{Number\ of\ \ customers\ in\ the\ group} \qquad (4)$$

where *Response Rate*$_z$ is the response rate for period *z,* and *Responders*$_z$ is the number of users who responded to the newsletter in period *z*. We computed that value for each period and for each treatment group.

3.1.3.3 Profit

We measured the profit generated by recommendations, using actual purchases. We measured the revenues and the number of orders for all participants in the pre-experiment period and in the two months following the beginning of the experiment. Although the firm did not provide the actual profit margin for the products, it assigned all the products into one of several bins, indicating different levels of profitability depending on a products' price. This resulted in a profit margin index on a scale of 1-5 that identified the relative profitability of all items in our study. The conversion of a products' price to a profit index was carried out using the following scheme:

$$\Pr ofit\ Index_i = \begin{cases} 1 & \text{if} \quad \text{price}_\text{i} < 5{,}00\ € \\ 2 & \text{if} \quad 5{,}00\ € < \text{price}_\text{i} < 10{,}00\ € \\ 3 & \text{if} \quad 10{,}00\ € < \text{price}_\text{i} < 17{,}00\ € \\ 4 & \text{if} \quad 17{,}00\ € < \text{price}_\text{i} < 25{,}00\ € \\ 5 & \text{if} \quad \text{price}_\text{i} > 25{,}00\ € \end{cases} \tag{5}$$

where *Profit Index*$_i$ is the profit index for product *I,* and *Price*$_i$ is the price of product *i*. Note that Eqn (5) is based on item price and not margin. The firm did not share the actual cost information for the comics but informed us that the margins closely tracked prices and that the two indices were equivalent for the purposes of calculating relative profit levels.

3.2 Post-Experiment Survey

3.2.1 Survey design

In addition to evaluating user feedback in terms of ratings and purchases, we gathered additional feedback from participants to study whether there were differences in trust across the treatments. The constructs for trust were derived from prior studies. We limited the scope of the survey to testing trusting beliefs (Beldad et al., 2010; Mayer et al., 1995), which are comprised of three constructs: ability, benevolence, and integrity (Mayer et al., 1995). We adapted a previously used set of questions and scales from Wang et al. (2005), Mayer et al. (1995), Schoorman et al. (2007), and Doney et al. (1997) where trusting beliefs were also linked to recommendations.

The survey was given with the last newsletter by e-mail to all participants in the study. The firm was keen to run a very brief survey, so the survey contained only 11 questions, which can be accessed in Appendix A. The first question evaluated the user's overall propensity to trust. Questions 2 through 9 were directed at measuring the users' perceptions of the recommender system's ability, integrity, and benevolence. In addition, we asked users about purchases of recommended products on other channels (Q_{10}). This was done because the company is not purely an Internet player. It mainly sells products via traditional channels. Therefore, it is possible that many customers use the website much more in the early stages of the purchasing process (e.g., when collecting information), preferring to buy products from newsstands or comic book shops in order to avoid shipping costs. Q_{10} allowed us to evaluate such spillovers. Finally, we also tested whether the users thought that the recommended items were expensive (Q_{11}). Because the profit-based recommender explicitly accounts for profit margins, we expected that users assigned to the profit-based recommender would perceive that the recommended products were relatively more expensive than other users.

Lastly, one concern related to the survey is that a number of subjects did not respond to the survey in the last week. In order to avoid any bias that might arise from non-response, we employed propensity score matching (Rosenbaum et al., 1983) on the demographic attributes that we had available to ensure that the subjects compared in the three populations were similar in terms of their demographics to control for any confounding factors when estimating the treatment effects We discuss propensity score matching in greater detail in Section 4.

4. Data Description

In this section, we provide summary information about the users in our study and the treatments administered to the users as part of the study.

4.1 User Characteristics

We analyzed the historical data owned by the firm in order to better understand the behavior of the customers involved in the experiment. The historical data available to us was collected from September 2008 to April 2010. Figure 1 plots the frequency histograms of orders and revenues generated in the past by the users in our study.

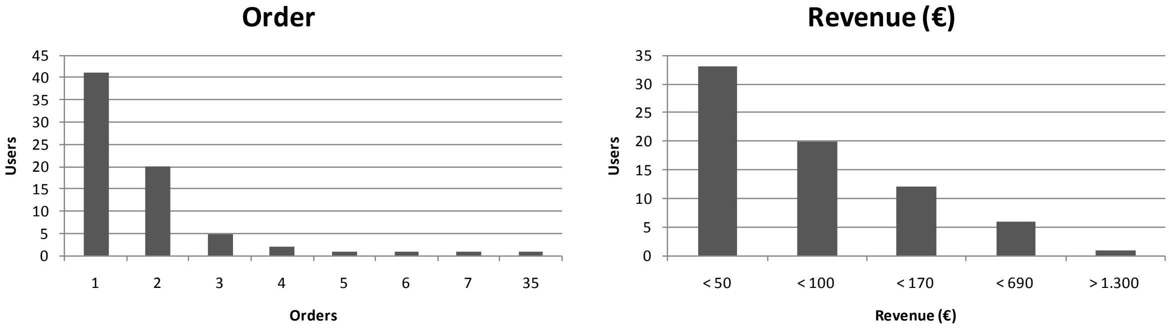


Figure 1. Details of the number of orders and revenue generated in the past by users involved in the experiment.

As Figure 1 shows, only 72 users out of 260 involved in the experiment purchased products during the period before the experiment. Almost all of those 72 users made only one or two orders, spending less than 100 Euros. Thus, the vast majority of users involved in the experiment were not "heavy users". Therefore, we expected few purchases from these users during the experiment. These users were randomly assigned to one of the three treatments. Figures 2 shows the demographic distributions of the users in each group.

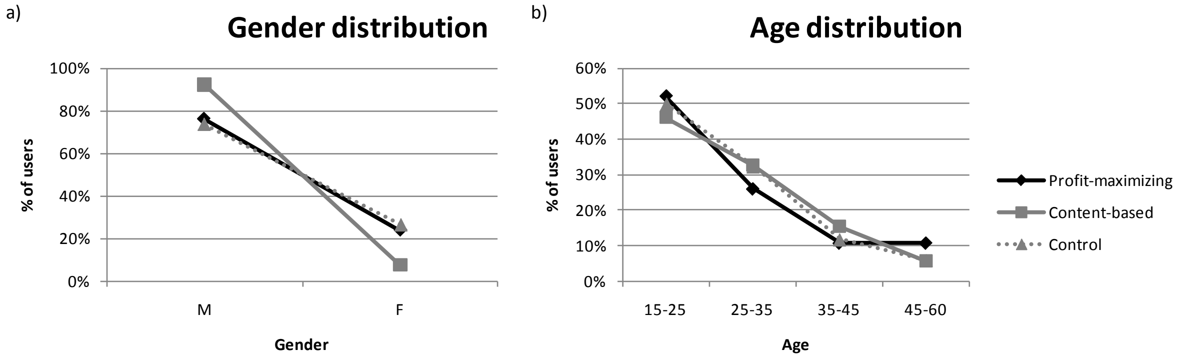


Figure 2. Distribution of users by gender (a) and age (b) in each treatment group.

We performed a t-test to explore whether there were any statistically significant differences across each treatment group for the aforementioned variables (namely, age and gender). The results showed that there were not any significant differences in terms of age or gender across the treatment groups.

4.2 Treatment Characteristics

To ensure that the recommendation systems were indeed providing different types of recommendations, we evaluated two important characteristics of the recommendations made by the algorithms: entropy and profitability of recommended items. We calculated the entropy in recommendations using Shannon's entropy (Shannon, 1948) based on product categories. We used four comic book categories to measure how consistent the comic recommendations were per person. The most consistent scenario occurred when a person received recommended comics from only one category. The least consistent scenario is when a person received recommended comics from each category equally. The four categories were based on the main classification the company uses to present its products on the website: 1) Marvel Comics (including the well-known comic books popularized by the American publisher); 2) Manga Comics (including all comic books published in Japan); 3) Other Comics (including all comic books popularized by either European publishers or American publishers other than Marvel); 4) Bundled Comics (including any kind of comic books sold in association with a DVD or other media contents). We evaluated the variability of the content of the comics:

$$H(X) = \sum P(X_i)\log_2(P(X_i)) \qquad (6)$$

where entropy, or H(X), is the uncertainty (or inconsistency) of variable X (or the categories), and P(X_i) is the probability that a comic X belongs to category i.

Again, in this context, entropy reflects the amount of diversity in the recommendations. Figure 3 reports the entropy distributions for users in the three treatment groups. As expected, the random control group is the one characterized by the highest entropy, whereas the recommendations received by the content-based group are those with the lowest entropy.

In addition, we measured the distribution of product profit margins (in terms of profit index) of recommended items in order to investigate whether there was any difference across the different treatment groups, given that our expectation was that the items recommended by the profit-maximizing recommender system had higher values than those recommended by the other two engines. Figure 4 shows the percentage of recommended items with a specific profit index (which ranges from 1-5) for each treatment group.

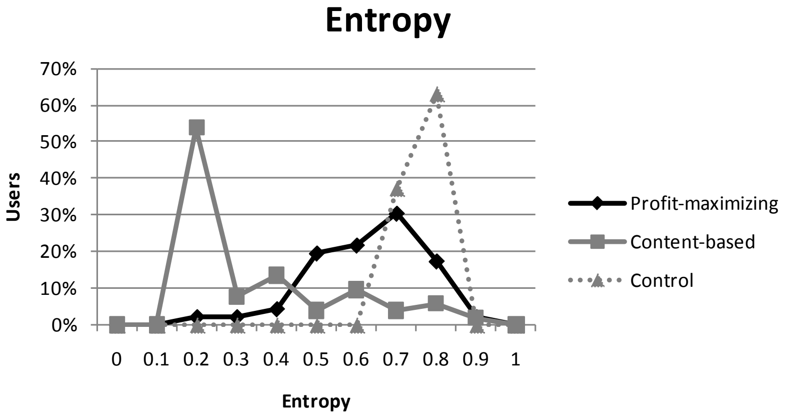


Figure 3. Entropy distributions for users in the three treatment groups.

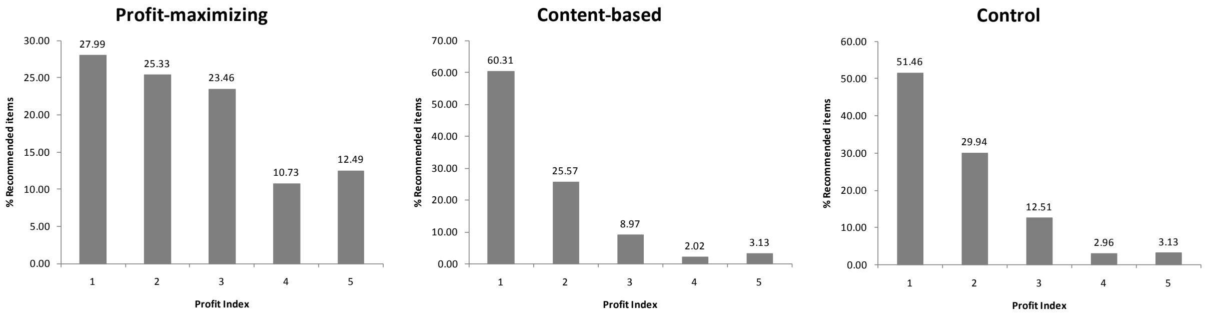


Figure 4. Distributions of profit index for recommended items.

As Figure 4 shows, the users assigned to the profit-maximizing group received recommendations for items with profit indexes higher than those suggested to members of the other two treatment groups. Thus, the three treatments appear to be functioning as expected.

5. Results

In this section, we present the results of the randomized field experiment as well as the results of a follow-up survey. We evaluated the relative performance of the different recommender systems based on the results of the experiment, and we compared user trust across treatment groups based on responses to the follow-up survey.

5.1 Experiment Results

As part of the experiment, we collected explicit feedback from users in the form of comic book ratings. In addition, the firm collected explicit purchase data on its customers. In this section, we report recommender system performance using 1) recommendation accuracy, measured by precision and average product ratings, 2) responses to the newsletters, and 3) profitability measures in terms of actual profit and proxies for profit. All performance measures were compared across the recommender system treatment groups over time.

5.1.1 Recommendation Accuracy

We measured recommendation accuracy in two ways: first, by comparing recommendation precision over time (Equation 2), and second, by comparing average ratings provided by users over time (Equation 3).

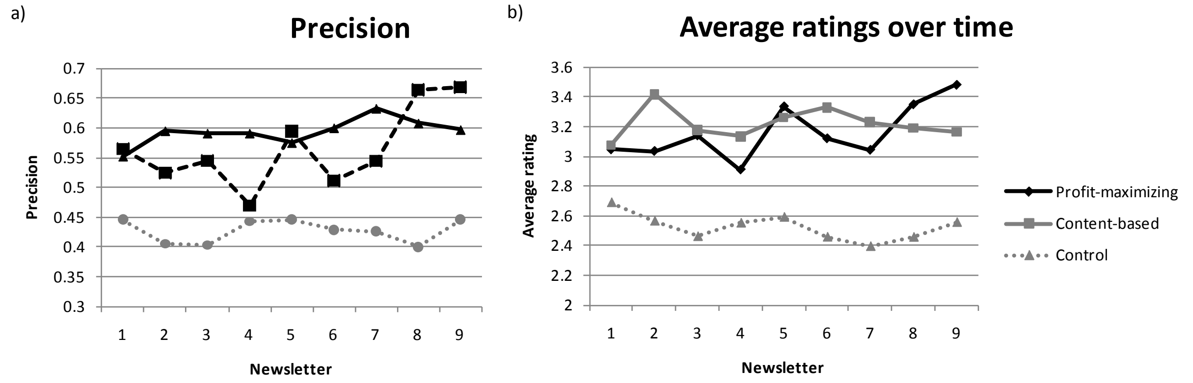


Figure 5. Precision (a) and overall average ratings (b) over time for control, content-based, and profit-maximizing recommenders.

Figure 5(a) reports the average precision of the three treatments over time. To compare the precision curves, we performed a Mann-Whitney significance test. We found statistically significant differences between the precision of recommendations made by the content-based system and those made by the random control system, as well as between the profit-maximizing system and the random control system (at the level of $p < 0.001$). In contrast, there was no significant difference with regard to precision between the content-based system and the profit-maximizing system. We concluded that the levels of precision for both the content-based recommendation system and the profit-maximizing recommender system are significantly higher than the precision level of the recommender system that generates random recommendations. However, the precision of the content-based recommendation system and the profit-maximizing recommender system were the same statistically.

We also compared the overall average ratings (Equation 3) over time. As Figure 5(b) shows, the ratings provided by customers who received content-based or profit-maximizing recommendations are higher, on average, than those provided by customers who received random recommendations. The differences between the average ratings provided by those in the content-based group and those provided by people assigned to the control group are statistically significant, as are the differences between the profit-maximizing group's ratings and the control group's ratings, according to a Mann-Whitney test (at the level of $p < 0.001$). Similar to the precision results above, there is no significant difference between the average ratings provided by the customers who received content-based recommendations and those provided by the customers who received profit-maximizing recommendations. Figure 6(a) plots the percentage of users who provided ratings higher than 3, on average, while Figure 6(b) plots the percentage of users who provided ratings lower than or equal to 3. The figures show that approximately 56% of the users in the content-based and profit-maximizing groups provided ratings above 3, on average, and the remaining 44% provided ratings below 3, on average, whereas only 31% of the users who received random recommendations provided ratings above 3.

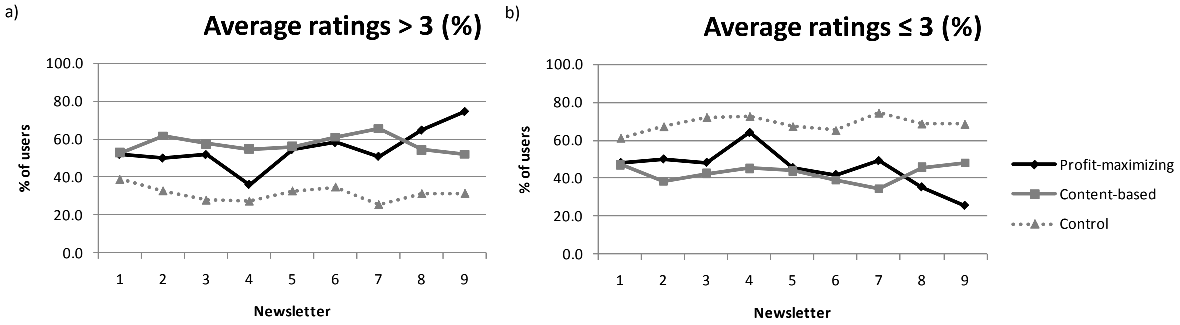


Figure 6. Percentage of users who rated the items recommended by each recommender system above 3 (a) and at or below 3 (b).

The implication of these results with respect to recommendation accuracy is that even though profit-based recommendations are based on firm-centric measures, the performance with respect to precision and average ratings is statistically the same as that of a traditional content-based system. In addition, both the profit-based and content-based recommender systems outperformed the control recommender system with respect to recommendation accuracy.

5.1.2 Response Rate

In addition to tracking recommendation accuracy, we tracked the response rates over time for each of the treatment groups. For each treatment group, customer response rate refers to the percentage of users in that group who responded to the newsletters containing the recommendations. Figure 7 shows an unremarkable reduction in the response rate, over time, for all of the treatment groups. Further, we found that there were no statistically significant differences across different recommendation engines when comparing the response rate curves over the entire study. However, for the last newsletter, in period 9, the control response rate of approximately 44% was lower than in other periods, and also lower than the content- and profit-based response rates of around 56% and 52%, respectively. Therefore, because of this notable difference in response rates, we made sure to check for biases in the final sample of participants when evaluating survey results discussed in Section 5.2.

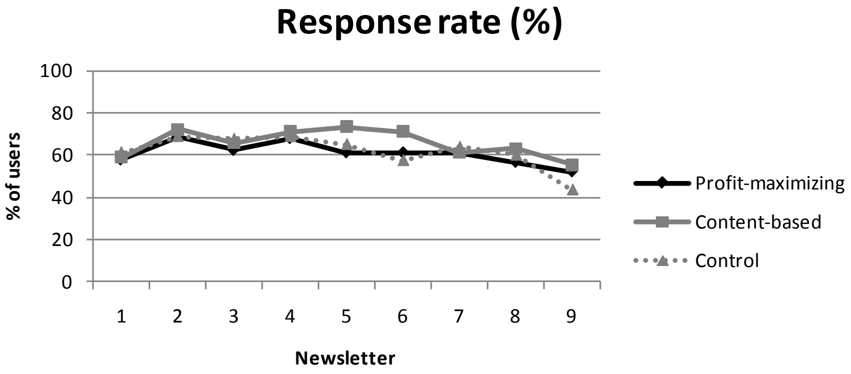


Figure 7. Response rate over time to different recommender systems.

The implication of our results with respect to response rates is that even though the recommendations in the newsletters are generated by very different recommendation processes, users responded to the newsletter at about the same rate over time.

5.1.3 Profit

As explained in Section 3.1.3, we measured the profit generated by each recommendation system in a number of ways, using both explicit purchases and proxies.

Figure 8 reports the mean number of purchased items per customer and the mean revenue per customer generated during the experiment. The number of items purchased (Figure 8(a)) was calculated by dividing the overall number of items purchased by members of each group by the number of users included in the group (i.e., 90 users for the profit and content groups, and 80 users for the control group). The mean revenue per customer (Figure 8(b)) is the overall revenue generated by each group divided by the number of users in the group. These graphs show the values averaged by customers and all the values referred occurred in a time period of two months. The overall number of purchased items is 81 for the content, 114 for the profit, and 16 for the control group. The overall revenues generated during the experiment are 428 euros for the content, 832 euros for the profit, and 151 euros for the control group. As expected, the random recommendations generated the lowest number of purchased items and the least revenue. The fact that members of the profit-based group purchased more products than the content-based group is surprising. This may be because customers are more likely to accept higher-priced recommendations than lower priced recommendations. This is in fact an argument provided in the personalization literature (e.g., Simonson, 2003). Another possible reason is that the customers in the profit-based group received recommendations of higher diversity. This hypothesis is discussed later on in this section. The higher revenue per user for the profit-maximizing group is because of its higher purchase count and higher margin per purchased item.

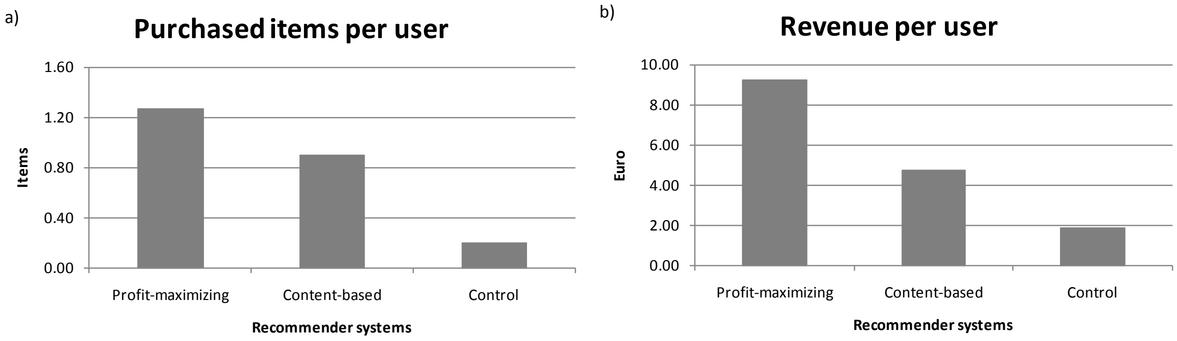


Figure 8. Mean number of purchased items (a) and revenue generated (b) during the experiment for different recommender systems.

We also computed the mean monthly revenue per customer and the mean price of items purchased by users in each treatment group in the past (i.e., during the 20 months prior to the experiment) and during the two months of the experiment. As shown in Figure 9(a), the profit-maximizing group witnessed the highest increase in the average revenue per user. The content-based group had a relatively smaller increase. Finally, there was almost no impact on the control group. Looking at the profit-maximizing group and control group, the difference-in-difference (DiD) of the average revenue per user is 2.846. A non-parametric permutation test indicates that this difference is significant and rejects the null hypothesis that the DiD is equal to zero (p-value < 5%). Similarly, the DiDs are significant for profit versus content. The DIDs are not significant for content versus control. These differences in profit and content were driven, in part, by the differences in purchase counts across groups, that was reported in Figure 8. Another key driver was the difference in the prices of purchased items. Figure 9(b) plots the price of items bought by customers in the three groups. The price of purchased items tended to increase for members of the profit-maximizing group whereas this was not the case for those in the content-based group. The group receiving random recommendations also witnessed an increase in the price of purchased items but very few recommended items were purchased by the users due to the lower accuracy.

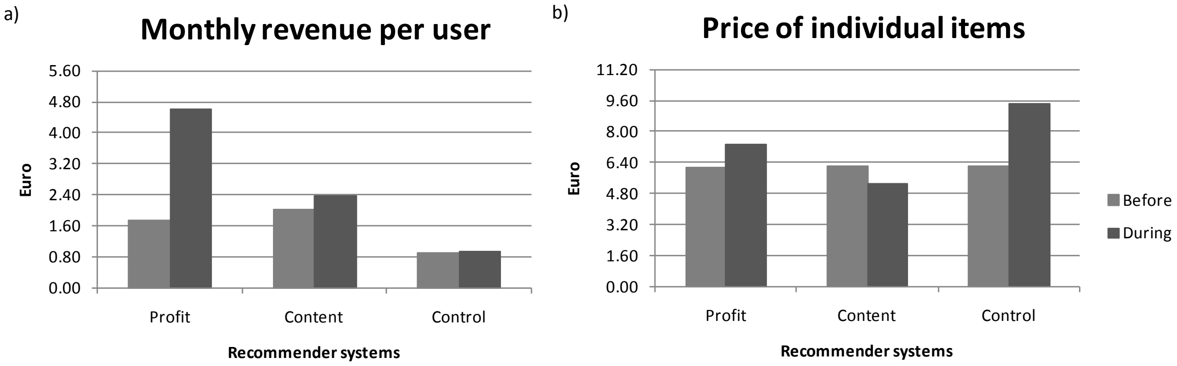


Figure 9. Mean monthly revenue per customer (a) and price of items purchased (b)before and after recommendations for different recommender systems.

One of the challenges in measuring orders and revenues associated with the treatments is that purchases were infrequent. Further, newsstands or comic bookstores sell products at the same prices available on the firm's website but without any shipping costs; often, consumers use online sites just to learn about new products but purchase offline. These offline purchases cannot be connected to online identities of experiment participants. Therefore, in addition to explicit purchase data from the firm's online site, we used a couple of proxies for offline purchases. First, in the post-experiment survey, we directly asked the users if they purchased any recommended products offline (Question 10). Figure 10 summarizes the responses to Question 10 and indicates that other purchases were generated by the recommendations but could not be tracked using only the website transaction data. However, we did not find any statistically significant differences among the responses to Question 10 of the three treatment groups. Therefore, the results in Figure 10 do not change the observations that we provided when discussing Figure 9.

In addition to using the answer to Q_{10} as a proxy for offline purchases, we also treated the comics rated 5 with a click on the "see more details" link as purchases. We then measured the estimated profit based on this assumption. This proxy assumes that items that are rated high and additionally researched by consumers are eventually bought by the users either in the online or offline channel. Figure 11 plots the mean (Figure 11a) and the cumulative (Figure 11b) profit using that proxy. The results again suggest that the profit-maximizing recommender generated significantly higher profits than the other two systems. Interestingly, there were no statistically significant differences between the content-based and control recommendations under this proxy. The reason is that although the content-based recommender gained higher ratings compared to the random recommendations, members of the control group clicked more frequently on the recommended items, and the items recommended to those in the control group were a bit more expensive than those recommended by the content-based system (see Figure 4). The reason for the higher click rate exhibited by the control group is probably related to the fact that these customers discovered more new items than those in other groups (see Section 5.2 on survey results that confirm this). The results shown in Figures 10 and 11 show that accounting for offline purchases is unlikely to reverse the primary findings discussed above.

Our experiment findings led us to conclude that the profit-maximizing algorithm, despite accounting for firm-centric measures that might be presumed to be negatively impact consumer engagement, can generate precise recommendations and effectively retain consumers' usage at the same level as the traditional content-based recommender. The profit-based algorithm used in our field experiment generated significantly higher profits.

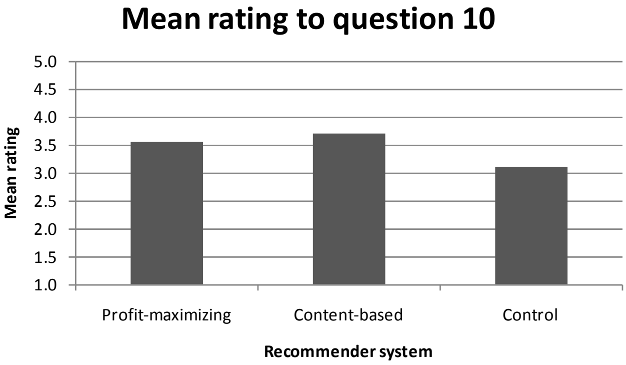


Figure 10. Mean rating for question 10 for different recommender systems, as a measure of purchases made on channels other than those online.

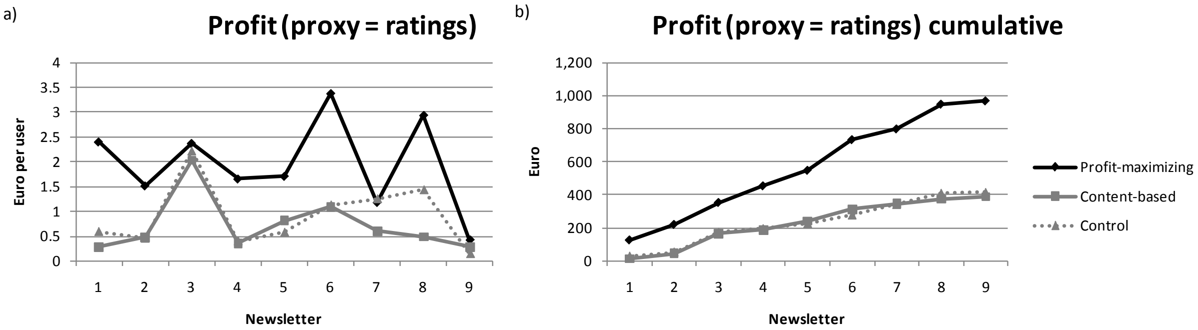


Figure 11. Estimate of the potential profit generated during the experiment, if items rated 5 would become purchases, instantaneous (a) and cumulative (b).

5.2 Survey results

As explained in previous sections, we distributed a final online survey at the end of the experiment to those who participated, attempting to measure the impact of different types of recommender systems on customers' trust. (See Appendix A for the list of survey questions.)

5.2.1 Summary of survey results.

We noted that the response rate to the survey among members of the control group was lower than for those in the other two treatment groups for the last newsletter. Therefore, we used propensity score matching (Rosenbaum, 1983) to ensure that the members of the groups in each of our treatments were consistent with respect to demographics. We used two variables, namely "age" and "gender," to match users in the control group to users in the other two groups; after matching, we found that there were no statistically significant differences between the average ages of group members (verified by a t-test). Next, we checked that the distributions of the propensity scores looked the same and that there were no statistically significant differences among group members (verified by a t-test). After matching, we were left with 37 participants in each treatment group. The results of the survey after propensity score matching, in terms of average ratings for each question and each treatment group, are reported in Table 1.[2]

	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8	Q_9	Q_{10}	Q_{11}
Profit-maximizing	2.8	3.3	3.4	3.0	2.8	3.2	3.1	3.6	4.1	3.6	3.2
Content-based	2.9	3.1	3.5	2.8	2.7	3.2	3.1	3.4	4.1	3.9	3.0
Control	2.9	2.9	3.4	3.4	3.1	3.5	3.4	3.4	3.8	3.2	3.0

Table 1. Average ratings for the questions in the final survey for each group of customers.

We began by evaluating differences between the treatments in terms of trust measures (Q_1 - Q_9). First, we checked for biases in the treatment groups with respect to the propensity to trust. Q_1 proposes: “My tendency to trust a person/thing is high.” We found no statistically significant differences in the general trust levels of the users in the different groups, using Q_1 responses. Thus, users in the different treatment groups were similar in terms of propensity to trust. Comparing the content-based recommender to the profit-based recommender in terms of trust measures (Q_2 - Q_9), we did not find any statistically significant differences. The absence of differences reveals that customers in both groups perceived the ability, integrity and benevolence of the recommender systems similarly and trusted the systems to the same extent.

Looking at the control group, it is surprising to note that the ratings are higher for questions Q_4 - Q_7. We found statistically significant the differences in the responses to Q_4 and Q_7 between the control group and the content-based group at the level of $p < 0.05$. Q_4 asks: “Personalized newsletters recommended comic books that I did not know about.” The significance of Q_4 can be explained by the algorithm used for each group. In particular, the control group received random recommendations that are likely to be more diverse. In contrast, the content group provided the lowest average rating on Q_4, given that its members received recommendations from a traditional content-based recommender algorithm that reduces the recommended items’ diversity. This is consistent with Figure 3, which showed that the random recommendations generated the highest entropy followed by the profit-based recommender and then the content-based system.

The fact that the control group also generated higher ratings for Q_7 is surprising and may have been driven by the system’s better performance related to the discovery dimension. We discuss this in greater detail in Section 5.2.3.

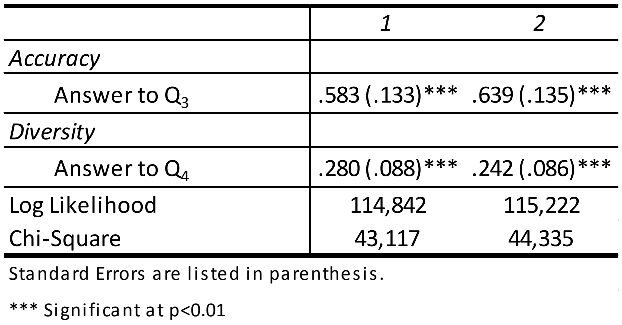
We now turn to Q_{10} and Q_{11}, which were asked primarily as sanity/robustness checks. We did not find significant differences for Q_{10}: “I bought some of the recommended comic books through channels other than Panini’s website”. This result was discussed above. Finally, Q_{11} tested whether users were able to perceive differences in the prices of the recommended items. Again, we did not find statistically significant differences. The higher rating for the profit-based group suggests that

[2] A number of the answers are obviously correlated. We evaluated the survey results using Principal Components Analysis (PCA). PCA shows that the four primary components are related to propensity to trust (Q_1), actual trust in the system (Q_2 - Q_9), offline purchases (Q_{10}), and cost of recommended items (Q_{11}). The full results of the PCA analysis as well as the correlation matrix for the survey responses are available upon request.

some of these users may have picked up on the fact that the profit-based system recommended more expensive items but, interestingly, this did not impact their trust in the system.

5.2.2 Trust drivers.

If biasing recommendations towards higher profit margin items does not impact trust, what does? To better understand the drivers of trust, we built ordered probit models, using measures of trust as dependent variables (i.e., responses to Q_7 and Q_6), with three measures of accuracy (i.e., precision, average ratings, and Q_3) and two measures of diversity (i.e., individual entropy and Q_4) as independent variables. We used responses to Q_7 and Q_6 as dependent variables because looking at the principal components extraction matrix, we noticed that the first main component is related to all of the questions about trust and the responses to Q_6 and Q_7 showed the highest correlation to this component. We found that accuracy and diversity were both positively and significantly related to trust, using the answers to Q_3 and Q_4 as independent variables. The results are shown in Table 2, with model 1 using Q_6 as the dependent variable and model 2 using Q_7 as dependent variable.



	1	*2*
Accuracy		
Answer to Q_3	.583 (.133)***	.639 (.135)***
Diversity		
Answer to Q_4	.280 (.088)***	.242 (.086)***
Log Likelihood	114,842	115,222
Chi-Square	43,117	44,335

Standard Errors are listed in parenthesis.

*** Significant at p<0.01

Table 2. Linking accuracy and diversity to trust (measured by dependent variables Q6 and Q7).

The results suggest that even when accuracy is accounted for, diversity is a significant driver of trust. The customers in the control group might have trusted the newsletter, despite the low accuracy, as much as the customers in the content-based group because they discovered more products they had never seen before, thanks to a higher diversity of recommendations. In contrast, the customers in the content-based group received more homogenous recommendations during the weeks in which the experiment was conducted. This might have lowered their perception of the reliability of the newsletter.

6. Conclusions

A number of recent studies have attempted to place recommenders within a profit-maximizing framework. While these studies have made notable advances in developing analytical models or heuristic strategies for deploying firm-centric recommendation engines, there has been no empirical evidence regarding their impact in practice. Because consumer choice and recommender design is endogenous, it is unrealistic to expect consumer response behavior to remain the same under these new designs. Consumers may in fact distrust firms that deploy biased recommendations and the net effect, relative to relevance-based recommendations, might in fact be negative. A number of commentators have pointed out that recommendations, even in the absence of biases, might be perceived by consumers as firms' efforts to persuade and manipulate. Once manipulated, it is unclear if the theoretical benefits will be realized. In this paper, we provided the first empirical evidence regarding the impact of profit-based recommendations in the field.

The paper makes two main contributions. First, we showed in one real-world setting that a firm can drive increases in purchase volume and margin per purchase by deploying profit-based recommenders. Further, we found that trust was not impacted even though we expected it to degrade as a result of users being served firm-centric based recommendations. Second, to our best knowledge, this is the first work to report on both consumer purchasing behavior and consumer trust from a real-world deployment of a recommender system. By combining considerations related to the design of a new information system with an evaluation of its impact on consumer usage and trust, this study contributes to both the Information Systems and Marketing literatures.

Our work informs firms that are interested in finding effective ways to deploy firm-centric recommendation systems. An important consideration is the messaging around such recommenders. The firm in our study did not explicitly identify to consumers that the recommendations were biased. This was because the firm wanted to retain the same message for users in all treatments. However, in reality, we believe that it is in the interest of firms to clearly identify such biases to consumers. Evidence from the sponsored search industry suggests that biasing results can be effective and that transparency serves both consumers and firms well. In that market, search engines present "organic search results" to consumers that are based solely on relevance of web pages to the consumer's query. In addition, search engines also present "sponsored search results" that are based on a combination of the relevance of the pages and the search engine's potential revenues from the consumer clicking on the results. Search engines face a similar tradeoff as in our study in the sense that sponsored results represent an opportunity to increase revenues but pose the risk that their excessive use may not be well received by consumers. Search engines have increasingly become transparent about identifying sponsored results to consumers and this has helped the market. Similar transparency might benefit profit-based recommenders. A large research stream has focused on the design and impact of sponsored search markets, focusing on how firms should determine which pages get top placement in the list of sponsored results and how consumers in turn respond to these sponsored lists. A similar stream on profit-based recommenders is just emerging and we believe this stream will continue to grow.

Our findings are not without limitations and these present opportunities for future research. Our results are based on one specific context and it is not clear how applicable the findings are to other business contexts. It is possible that in environments in which purchases are less frequent or item prices are much higher (such as household appliances), consumer response to profit-based recommendations may be different. Thus, an interesting direction for future research might be to test the validity of our findings in other industries, particularly those characterized by higher prices or very different purchasing frequency. Another limitation is that we cannot determine how our results might be affected if users were informed about how the list of recommended items was generated. There is a concern that if the algorithm design were revealed, consumers might trust the system less. Future work can evaluate the impact of transparency and messaging on consumer response. Lastly, the objective of our paper was not to design the optimal profit-based recommender but to evaluate consumer response to a reasonable design that has been proposed in the literature. Future work can evaluate where the optimal design tradeoff between profitability-based and relevance-based recommendations lies.

References

Adomavicius, Gediminas, Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6) 734–749.

Adomavicius, Gediminas, Ramesh Sankaranarayanan, Shahana Sen, Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems 23 103–145.

Adomavicius, Gediminas, Kwon, Young Ok. 2010. Improving aggregate recommendation diversity using ranking-based techniques. Forthcoming IEEE Transactions on Knoledge and Data Engineering.

Aiken, K. D., & Bousch, D. M. (2006). Trustmarks, objective-source ratings, and implied investments in advertising: Investigating online trust and the contextspecific nature of internet signals. Journal of the Academy of Marketing Science, 34, 308–323.

Akoglu, L. and C. Faloutsos. ValuePick: Towards a Value-Oriented Dual-Goal Recommender System. *ICDM Workshop on Optimization Based Methods for Emerging Data Mining Problems* , Sydney, Australia, Dec. 2010

Aksoy, L., and Bloom, P. N. “Impact of Ordered Alternative Lists on Decision Quality: The Role of Perceived Similarity,” paper presented at the Society for Consumer Research Winter Conference, Scottsdale, AZ, February 15-17, 2001.

Ansari, Asim, Carl F. Mela. 2003. E-customization. Journal of Marketing Research 40(2) 131–145.

Ansari, Asim, Skander Essegaier, Rajeev Kohli. 2000. Internet recommendation systems. Journal of Marketing Research 37(3) 363–375.

Balabanovic, Marko, Yoav Shoham. 1997. Fab: Content-based, collaborative recommendation. Communications of the ACM 40 66–72.

Bart, Y., Shankar, V., Sultan, F., & Urban, G. L. (2005). Are the drivers and role of online trust the same for all web sites and consumers? A large-scale exploratory empirical study. Journal of Marketing, 69, 133–152.

Baumol, W. E., A. Ide. 1956. Variety in retailing. Management Sci. 3(1) 93–101.

Beldad, A., Menno de Jong, Michael Steehouder. 2010. How shall I trust the faceless and the intangible? A litterature review on the antecedents of online trust. Computers in Human Behavior Volume 26, Issue 5, 857 – 869.

Bettman, James R, Mary Frances Luce, John W Payne. 1998. Constructive consumer choice processes. Journal of Consumer Research: An Interdisciplinary Quarterly 25(3) 187–217.

Bharati, P., and Chaudhury, A. "An Empirical Investigation of Decision-Making Satisfaction in Web-Based Decision Support Systems," Decision Support Systems (37:2), 2004, pp. 187-197.

Bodapati, Anand V. 2008. Recommendation systems with purchase data. Journal of marketing research 45(1) 77–93.

Brand, M. "A random walks perspective on maximizing satisfaction and profit," in Proc. of SIAM, 2005.

Breese, John S., David Heckerman, Carl Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. Morgan Kaufmann, 43–52.

Briggs, P., Simpson, B., & De Angeli, A. (2004). Personalisation and trust: A reciprocal relationship? In C. M. Karat, J. O. Blom, & J. Karat (Eds.), Designing personalized user experiences in e-commerce (pp. 39–55). Netherlands: Kluwer.

Brin, S., and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Search Engine. Computer Networks and ISDN Systems, 30, 107-117.

Brynjolfsson, E., Smith, M. D. and Hu Y. 2003. Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers. Management Science 49 (11), pp. 1580-1596.

Broniarczyk, S. M., W. D. Hoyer, L. McAlister. 1998. Consumers' perceptions of the assortment offered in a grocery category: The impact of item reduction. J. Marketing Res. 35(2) 166–176.

Casalo, L. V., Flavian, C., & Guinaliu, M. (2007). The influence of satisfaction, perceived reputation and trust on a consumer's commitment to a website. Journal of Marketing Communications, 13(1), 1–17.

Chau, P. Y. K., Hu, P. J. H., Lee, B. L. P., & Au, A. K. K. (2007). Examining customers' trust in online vendors and their dropout decisions: An empirical study. Electronic Commerce Research and Applications, 6, 171–182.

Chen, C. (2006). Identifying significant factors influencing consumer trust in an online travel site. Information Technology and Tourism, 8, 197–214.

Chen, Long-Sheng, Fei-Hao Hsu, Mu-Chen Chen, Yuan-Chia Hsu. 2008. Developing recommender systems with the consideration of product profitability for sellers. Inf. Sci. 178(4) 1032–1048.

Claypool, Mark, Anuja Gokhale, Tim Miranda, Pavel Murnikov, Dmitry Netes, Matthew Sartin. 1999. Combining content-based and collaborative filters in an online newspaper.

Cooke, Alan D. J., Harish Sujan, Mita Sujan, Barton A. Weitz. 2002. Marketing the unfamiliar: The role of context and item-specific information in electronic agent recommendations. Journal of Marketing Research 39(4) 488–497.

Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: A study of consumer perceptions. Electronic Commerce Research and Applications, 2, 203–215.

Das, Aparna, Claire Mathieu, Daniel Ricketts. 2010. Maximizing profit using recommender systems. Working paper.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340.

Dias, M. Benjamin, Dominique Locher, Ming Li, Wael El-Deredy, Paulo J.G. Lisboa. 2008. The value of personalised recommender systems to e-business: a case study. RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems. ACM, New York, NY, USA, 291–294.

Doney, P. M., & Cannon, J. P. (1997). An examination of the nature of trust in buyer–seller relationships. Journal of Marketing, 61, 35–51.

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national culture on the development of trust. Academy of Management Review, 23(3), 601–620.

Dreze, X., S. J. Hoch, M. E. Purk. 1994. Shelf management and space elasticity. J. Retailing 70(4) 301–326.

Flavian, C., Guinaliu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction, and consumer trust on website loyalty. Information & Management, 43, 1–14.

Fleder, Daniel, Kartik Hosanagar. 2009. Blockbuster culture’s next rise or fall: The impact of recommender systems on sales diversity. Manage. Sci. 55(5) 697–712.

Garfinkel, Robert, Gopal, Ram D., Pathak, Bhavik K., Venkatesan, Rajkumar and Yin, Fang, Empirical Analysis of the Business Value of Recommender Systems. Journal of Management Information Systems, 27 (2) 2010, pp. 159-188.

Gefen, D., Karahanna, E., and Straub, D. W. "Trust and TAM in Online Shopping: An Integrated Model," MIS Quarterly (27:1), 2003, pp. 51-90.

Gefen, D. (2000). E-commerce: The roles of familiarity and trust. Omega, 28, 725–737.

Gershoff, A. D., Mukherjee, A., and Mukhopadhyay, A. "Consumer Acceptance of Online Agent Advice: Extremity and Positivity Effects," Journal of Consumer Psychology (13:1-2), 2003, pp. 161-170.

J.L. Herlocker, J.A. Konstan, L.G. Terveen and J.T. Riedl, "Evaluating collaborative filtering recommender systems", ACM T. Inform. Syst., vol. 22, no. 1, pp. 5-53, 2004.

Hess, T. J., Fuller, M. A., and Mathew, J. "Involvement and Decision-Making Performance with a Decision Aid: The Influence of Social Multimedia, Gender, and Playfulness," Journal of Management Information Systems (22:3), 2005, pp. 15-54.

Hoch, S. J., E. T. Bradlow, B.Wansink. 1999. The variety of an assortment. Marketing Sci. 18(4) 527–546.

Hoffman, D. L., Novak, T. P., & Peralta, M. (1999). Building consumer trust online. Communications of the ACM, 42(4), 80–85.

Hosanagar, Kartik, Ramayya Krishnan, Liye Ma. 2008. Recommended for you: The impact of profit incentives on the relevance of online recommendations. Proceedings of the International Conference on Information Systems .

Hosmer, L. T. (1995). Trust: The connecting link between organizational theory and philosophical ethics. Academy of Management Review, 20(2), 379–403.

Iwata, T., K. Saito, T. Yamada. (2008). Recommendation Method for Improving Customer Lifetime Value. IEEE Transactions on Knowledge and Data Engineering, 20 (9), 1254-1263.

Jarvenpaa, S. L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an internet store. Information Technology and Management, 1, 45–71.

Josang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. Decision Support Systems, 43, 618–644.

Kahn, B., D. R. Lehmann. 1991. Modeling choice among assortments. J. Retailing 67(3) 274–299.

Kim, D. J., Song, Y. I., Braynoy, S. B., & Rao, H. R. (2005). A multidimensional trust formation model in B-to-C e-commerce: A conceptual framework and content analyses of academia/practitioner perspectives. Decision Support Systems, 40, 143–165.

Kim, J., & Moon, J. Y. (1998). Designing towards emotional usability in customer interfaces. Trustworthiness of cyber-banking system interfaces. Interacting with Computers, 10, 1–29.

Koehn, D. (2003). The nature of and conditions for online trust. Journal of Business. Ethics, 43, 3–19.

Komiak, S., and Benbasat, I. “The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents,” MIS Quarterly (30:4), 2006.

Koufaris, M., & Hampton-Sosa, W. (2004). The development of initial trust in an online company by new customers. Information & Management, 41, 377–397.

Koufaris, M., Hampton-Sosa, W. Customer trust online: examining the role of the experience with the website. CIS Working paper series, Zicklin School of Business, Baruch College, New York (Oct. 202).

Kuan, H. H., & Bock, G. W. (2007). Trust transference in brick and click retailers: An investigation of the before-online-visit phase. Information & Management, 44, 175–187.

Lee, M. K. O., & Turban, E. (2001). A trust model for consumer internet shopping. International Journal of Electronic Commerce, 6(1), 75–91.

Lee, Wee Sun. 2001. Collaborative learning for recommender systems. In Proc. 18th International Conf. On Machine Learning. Morgan Kaufmann, 314–321.

Liao, C., Palvia, P., & Lin, H. N. (2006). The roles of habit and website quality in ecommerce. International Journal of Information Management, 26, 469–483.

Liang T.P., Lai H. J. and Ku Y.C. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings, Journal of Management Information Systems 23 (3), 2006, pp. 45-70.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organization trust. Academy of Management Review, 20(3), 709–734.

Mcginty, Lorraine, Barry Smyth. 2003. On the role of diversity in conversational recommender systems. Proceedings of the Fifth International Conference on Case-Based Reasoning. Springer, 276–290.

McKnight, D. H., Choudhury, V., Kacmar, C. Developing and validating trust measures for e-commerce: an integrative typology. Information system research, vol. 12, no. 3, Semptember 2002, pp. 334-359.

Metzger, M. J. (2006). Effects of site, vendor, and consumer characteristics on website trust and disclosure. Communication Research, 33(3), 155–179.

Mooney, Raymond J., Loriene Roy. 1999. Content-based book recommending using learning for text categorization. In Proceedings Of The Fifth Acm Conference On Digital Libraries. ACM Press, 195–204.

Nicholas, Ian Soboroff. Charles, Charles K. Nicholas. 1999. Combining content and collaboration in text filtering. In Proceedings of the IJCAI99 Workshop on Machine Learning for Information Filtering. 86–91.

Payne, J., Bettman, J. and Johnson, E. 1993. The adaptive decision maker. Cambridge University Press.

Palmisano, Cosimo, Alexander Tuzhilin, Michele Gorgoglione. 2008. Using context to improve predictive modeling of customers in personalization applications. IEEE Trans. on Knowl. and Data Eng. 20(11) 1535–1549.

Pavlou, P. A. “Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model,” International Journal of Electronic Commerce (7:3), 2003, pp. 101-134.

Pazzani, Michael J., Daniel Billsus. 2007. Content-based recommendation systems. The Adaptive Web: Methods And Strategies Of Web Personalization. Volume 4321 of lecture notes in computer science. Springer-Verlag, 325–341.

Ranganathan, C., Goode, V., & Ramaprasad, A. (2003). Managing the transition to bricks and clicks. Communications of the ACM, 46(12), 308–316.

Ridings, C. M., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. Journal of Strategic Information Systems, 11, 271–295.

Rosenbaum, P. R., Rubin, D. B. The central role of the propensity score in observational studies for causal effects. Biometrika, vol. 70, no. 1, 1983, pp. 41–55.

Schafer, J. Ben, Joseph A. Konstan, , John Riedl, John Riedl. 2001. E-commerce recommendation applications.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. 2007. An integrative model of organizational trust: Past, present, and future. Academy of Management Review, 32 (2), 344-354.

Shannon, C. A Mathematical Theory of Communication. Bell system Technical Journal, vol 27, lug 1948.

Shardanand, Upendra, Pattie Maes. 1995. Social information filtering: Algorithms for automating ”word of mouth”. ACM Press, 210–217.

Short, J., Williams, E., & Christie, B. (1976). The social psychology of telecommunications. London: John Wiley & Sons.

Simonson, I. Determinants of customers’ responses to customized offers: conceptual framework and research propositions. J Market, 69, 1, 2005, 32–45.

Sinha, R., and Swearingen, K. “Comparing Recommendations Made by Online Systems and Friends,” in Proceedings of the 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries, Dublin, Ireland, June 18-20, 2001.

Smeltzer, L. (1997). The meaning and origin of trust in buyer–seller relationships. International Journal of Purchasing and Materials Management, 33(1), 40–48.

Sutanonpaiboon, J., Abuhandieh, A. Factors influencing trust in online consumer-to-consumer (C2C) transactions. Journal of internet commerce, vol 7(2), 2008, 203-219.

Swearingen, K., and Sinha, R. “Beyond Algorithms: An HCI Perspective on Recommender Systems,” paper presented at the ACM SIGIR Workshop on Recommender Systems, New Orleans, LA, September 13, 2001.

Sztompka, P. (1999). Trust: A sociological theory. Cambridge: Cambridge University Press.

Teo, T. S. H., & Liu, J. (2007). Consumer trust in e-commerce in the United States, Singapore, and China. Omega, 35, 22–38

van Herpen, E., R. Pieters. 2002. The variety of an assortment: An extension to the attribute-based approach. Marketing Sci. 21(3) 331–341.

Wang, H., Lee, M. K. O., & Wang, C. (1999). Consumer privacy concerns about internet marketing. Communications of the ACM, 41(3), 63–70.

Wang, W., Benbasat, I. 2005. Trust in and adoption of online recommendation agents. Journal of the association for information systems, vol. 6, no. 3, pp. 72-101.

Xiao B. and Benbasat I. 2007. E-commerce Product Recommendation Agents: Use, Characteristics, and Impact. MIS Quarterly 31 (1), pp. 137-209.

Appendix A

The following eleven questions were included in the survey on trust.

	Q1	My tendency to trust people/things is high
Ability	Q2	This personalized newsletter is like a real expert in assessing comic books because it considers all important attributes
	Q3	The personalized newsletters provided me with relevant recommendations
	Q4	The personalized newsletters recommended comic books that I didn’t know
	Q5	I am willing to let this personalized newsletter assist me in deciding which product to buy
Integrity	Q6	The newsletter is reliable
	Q7	I trust the personalized newsletter
Benevolence	Q8	The company created the personalized newsletter to help me
	Q9	The personalized newsletter is a service provided by the company to customers
Purchases offline	Q10	I bought some of the recommended comic books on channels different from the company’s web site
Perception of price	Q11	The personalized newsletter suggested interesting comic books, but they were too expensive