1-1-2015

Three Essays on Big Data Consumer Analytics in E-Commerce

Dokyun Lee  
University of Pennsylvania, leedokyun@gmail.com

Follow this and additional works at: http://repository.upenn.edu/edissertations

🔗 Part of the Advertising and Promotion Management Commons, Databases and Information Systems Commons, and the Marketing Commons

Recommended Citation
http://repository.upenn.edu/edissertations/1830

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/edissertations/1830
For more information, please contact libraryrepository@pobox.upenn.edu.
Three Essays on Big Data Consumer Analytics in E-Commerce

Abstract
Consumers are increasingly spending more time and money online. Business to consumer e-commerce is growing on average of 20 percent each year and has reached 1.5 trillion dollars globally in 2014. Given the scale and growth of consumer online purchase and usage data, firms' ability to understand and utilize this data is becoming an essential competitive strategy.

But, large-scale data analytics in e-commerce is still at its nascent stage and there is much to be learned in all aspects of e-commerce. Successful analytics on big data often require a combination of both data mining and econometrics: data mining to reduce or structure (from unstructured data such as text, photo, and video) large-scale data and econometric analyses to truly understand and assign causality to interesting patterns. In my dissertation, I study how firms can better utilize big data analytics and specific applications of machine learning techniques for improved e-commerce using theory-driven econometrical and experimental studies. I show that e-commerce managers can now formulate data-driven strategies for many aspect of business including cross-selling via recommenders on sales sites to increasing brand awareness and leads via social media content-engineered-marketing. These results are readily actionable with far-reaching economical consequences.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Operations & Information Management

First Advisor
Kartik Hosanagar

This dissertation is available at ScholarlyCommons: http://repository.upenn.edu/edissertations/1830
THREE ESSAYS ON BIG DATA CONSUMER ANALYTICS IN E-COMMERCE

Dokyun Lee

A DISSERTATION

in

Operation and Information Management

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

Kartik Hosanagar, Professor of OPIM

Graduate Group Chairperson

Eric Bradlow, Professor of Marketing,
Statistics, and Education

Dissertation Committee

Kartik Hosanagar, Professor of Operations and Information Management
Lorin M. Hitt, Class of 1942 Professor of Operations and Information Management
Harikesh S. Nair, Professor of Marketing, Stanford Graduate School of Business
Dedicated to my parents, Hongme and Sangwook,
and my sister, Soyeon.
ACKNOWLEDGEMENT

The five years I spent here has been the best time of my life thus far. I am grateful for the best advisors and friendly colleagues who have helped and stimulated me throughout.

I would like to first thank my advisor, Kartik Hosanagar, who goes above and beyond for his students with a style and grace. I would like to thank Lorin Hitt and Harikesh Nair for being in my committee and for giving me valuable advices and feedbacks on both research and career. I am blessed to have world-leading researchers and educators as my mentors, who excel in advising and guiding as much as they do in researching and teaching. I learned many things beyond researching from them and I hope to learn even more in the decades to come. I am grateful for insightful discussions and general supports from Vibhanshu Abhishek, Christophe Van den Bulte, Raghuram Iyengar, David Bell, Jonah Berger, Dylan Small, Dean Foster, Paul Shaman, Lynn Wu, Xuanming Xu, Noah Gans, Eric Clemons, Jing Peng, Fujie Jin, Eric Bradlow, Karl Ulrich, Jeff Cai, and Arun Gopalakrishnan.

I would like to thank the Baker Retailing Center, the William And Phyllis Mack Institute for Innovation Management, Fishman-Davidson Center for Service and Operations Management, and Wharton Risk Management and Decision Processes Center for their generous and instrumental financial support. I am grateful to Sargent Shriver, Andrea Nurse, Kim Watford, Patricia James, Tara Mullins for all the administrative help and Stan Liu and Jamie Walter for IT support.

Lastly, I would like to thank my parents for providing me with the best opportunities possible for following any goals and dreams I have. They have been my role models all my life and I am forever grateful for their steadfast support and unconditional sacrifice. Finally, most opportunities I was given was possible largely thanks to my lovely sister.
ABSTRACT

THREE ESSAYS ON BIG DATA CONSUMER ANALYTICS IN E-COMMERCE

Dokyun Lee

Kartik Hosanagar

Consumers are increasingly spending more time and money online. Business to consumer e-commerce is growing on average of 20 percent each year and has reached 1.5 trillion dollars globally in 2014. Given the scale and growth of consumer online purchase and usage data, firms’ ability to understand and utilize this data is becoming an essential competitive strategy. But, large-scale data analytics in e-commerce is still at its nascent stage and there is much to be learned in all aspects of e-commerce. Successful analytics on big data often require a combination of both data mining and econometrics: data mining to reduce or structure (from unstructured data such as text, photo, and video) large-scale data and econometric analyses to truly understand and assign causality to interesting patterns. In my dissertation, I study how firms can better utilize big data analytics and specific applications of machine learning techniques for improved e-commerce using theory-driven econometrical and experimental studies. I show that e-commerce managers can now formulate data-driven strategies for many aspect of business including cross-selling via recommenders on sales sites to increasing brand awareness and leads via social media content-engineered-marketing. These results are readily actionable with far-reaching economical consequences.
# TABLE OF CONTENTS

ACKNOWLEDGEMENT .......................................................... iv  

ABSTRACT ................................................................. v  

LIST OF TABLES ........................................................... ix  

LIST OF ILLUSTRATIONS ................................................ xi  

CHAPTER 1:  Introduction .................................................. 1  

CHAPTER 2:  The Effect of Social Media Marketing Content on Consumer Engagement: Evidence from Facebook .................................................. 5  
  2.1 Introduction ............................................................ 5  
  2.2 Data ............................................................... 12  
  2.3 Empirical Strategy .................................................. 29  
  2.4 Results ............................................................ 37  
  2.5 Discussion and Managerial Implications ....................... 50  
  2.6 Conclusions ....................................................... 54  

CHAPTER 3:  People Who Liked This Study Also Liked: The Impact of Recommender Systems on Sales Volume and Diversity .......................................... 56  
  3.1 Introduction ......................................................... 56  
  3.2 Prior Work .......................................................... 59  
  3.3 Problem Formulation .............................................. 64  
  3.4 Data ............................................................... 69  
  3.5 Results ............................................................ 72  
  3.6 Discussion and Conclusion ................................... 85
# LIST OF TABLES

TABLE 1 :  Variable Descriptions and Summary for Content-coded Data  

TABLE 2 :  Examples of Messages and Their Content Tags  

TABLE 3 :  Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation  

TABLE 4 :  User-level Setup Notation.  

TABLE 5 :  EdgeRank Model Estimates  

TABLE 6 :  Persuasive vs Informative  

TABLE 7 :  Aggregate Logistic Regression Results For Comments and Likes  

TABLE 8 :  Predicted versus Actual Engagement Ranking for Three Illustrative messages  

TABLE 9 :  Literature on Impact of Recommender Systems and Claims  

TABLE 10 :  Movie Genres Viewed and Purchased  

TABLE 11 :  Data Summary Statistics  

TABLE 12 :  Individual Item Views Comparison  

TABLE 13 :  Individual Item Purchases Comparison  

TABLE 14 :  Individual Wallet-Size Comparison  

TABLE 15 :  Aggregate View Diversity  

TABLE 16 :  Aggregate Sales Diversity  

TABLE 17 :  Individual View Diversity  

TABLE 18 :  Individual Purchase Diversity  

TABLE 19 :  Hypotheses Tested  

TABLE 20 :  Permutation Test Results for Co-purchase Network Comparisons  

TABLE 21 :  Hypotheses under Robustness Checks  

TABLE 22 :  Product Categories Occurring In the Dataset  
TABLE 23 : Variable Descriptions and Summary for Content-coded Data. . . . 94
TABLE 24 : Utilitarian VS. Hedonic Product Cluster Means . . . . . . . . . . 99
TABLE 25 : Search VS. Experience Product Cluster Means . . . . . . . . . . . . 103
TABLE 26 : Logistic Regression Results Table . . . . . . . . . . . . . . . . . . 107
TABLE 27 : Multiple Specifications for Review Related Variables . . . . . . . 110
TABLE 28 : Hypotheses and Results . . . . . . . . . . . . . . . . . . . . . . . . 110
TABLE 29 : Other Takeaways . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
TABLE 30 : A Few Examples of Message Attributes Used in Natural Language Processing Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
TABLE 31 : Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 121
TABLE 32 : Survey Instrument . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
LIST OF ILLUSTRATIONS

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIGURE 1</td>
<td>(Left) Example of a firm’s Facebook Page (Walmart). (Right) Example of a</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>firm’s message and subsequent user engagement with that message (Tennis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Warehouse). Example is not necessarily from our data.</td>
<td></td>
</tr>
<tr>
<td>FIGURE 2</td>
<td>Co-occurrence of Attribute Characteristics Across messages</td>
<td>20</td>
</tr>
<tr>
<td>FIGURE 3</td>
<td>Bubble Chart of Broader Industry Category vs Message Content</td>
<td>21</td>
</tr>
<tr>
<td>FIGURE 4</td>
<td>Box Plots of Log(engagement+1) vs Time since message Release</td>
<td>22</td>
</tr>
<tr>
<td>FIGURE 5</td>
<td>Average Likes and Comments by Message Type</td>
<td>23</td>
</tr>
<tr>
<td>FIGURE 6</td>
<td>Average Likes and Comments by Message Type by Industry</td>
<td>23</td>
</tr>
<tr>
<td>FIGURE 7</td>
<td>Average Likes and Comments by Message Content</td>
<td>24</td>
</tr>
<tr>
<td>FIGURE 8</td>
<td>Cronbach’s Alphas for 5,000 Messages</td>
<td>25</td>
</tr>
<tr>
<td>FIGURE 9</td>
<td>Impression-Engagement Funnel</td>
<td>31</td>
</tr>
<tr>
<td>FIGURE 10</td>
<td>Page-level Fixed effect Estimates from Generalized Additive Model Across</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>14 Demographic Bins</td>
<td></td>
</tr>
<tr>
<td>FIGURE 11</td>
<td>Time Since message Release ($\tau$) Coefficients Box plot Across Demographics</td>
<td>40</td>
</tr>
<tr>
<td>FIGURE 12</td>
<td>Message Characteristic Coefficients for Comments and Likes</td>
<td>46</td>
</tr>
<tr>
<td>FIGURE 13</td>
<td>Logistic Regression by Industry</td>
<td>47</td>
</tr>
<tr>
<td>FIGURE 14</td>
<td>Proportion of Content Posted Split into Hour-bin</td>
<td>49</td>
</tr>
<tr>
<td>FIGURE 15</td>
<td>Message Characteristic Coefficients for Shares and Click-throughs</td>
<td>52</td>
</tr>
<tr>
<td>FIGURE 16</td>
<td>Lorenz Curve</td>
<td>65</td>
</tr>
<tr>
<td>FIGURE 17</td>
<td>Recommender Example</td>
<td>70</td>
</tr>
<tr>
<td>FIGURE 18</td>
<td>Average Individual Statistics</td>
<td>74</td>
</tr>
<tr>
<td>FIGURE 19</td>
<td>Lorenz Curves for Movie Genres Purchased</td>
<td>75</td>
</tr>
</tbody>
</table>
FIGURE 20: Co-Purchase Network Graphs of Genre Purchases under Control and Purchase-Based Collaborative Filtering. 83

FIGURE 21: Genre Purchase Share Comparison on Purchase-based CF vs. Control 84

FIGURE 22: Recommendation Panel 93

FIGURE 23: Survey Form Used in Amazon Mechanical Turk 113

FIGURE 24: Diagram of NLP Training and Tagging Procedure 119
CHAPTER 1 : Introduction

Consumers are increasingly spending more time and money online. A 2013 study (eMarketer, 2013b) reports that for the first time, the average adult in the US will spend more time online than watching TV at just above five hours per day. The spread of mobile devices like smartphones and tablet PCs are also fueling this dramatic increase in consumer online activities. Consequently, e-commerce is growing faster than ever. Business to consumer e-commerce is growing on average of 20% each year and has reached $1.5 trillion globally in 2014 (eMarketer, 2014).

This growth in online activity has given arise to a new phenomenon called “big data”. “Big data” is a catch-phrase used to describe massive data recorded online (e.g., e-commerce, search engine) and offline by a myriad of sensors (e.g., surveillance, traffic monitor). The term is used to describe four different aspects of challenges arising from exploding data: 1) Volume, which refers to the sheer amount of volume recorded\(^1\); 2) Variety, arising from unstructured data like text and photos that come from numerous sources such as social media sites; 3) Velocity, which refers to the speed at which data gets recorded; and 4) Veracity, which refers to the uncertainty and missing data. The “big data” phenomenon has created problems and challenges to virtually everyone including marketers, business managers, academics, and policy makers: How can big data be utilized for improved marketing, business managing, and policy making?

Given the scale and growth of consumer online purchase and usage data, firms’ ability to understand and utilize this big data is becoming an essential competitive strategy. Several academic and industry reports (Kiron et al., 2011; Rogers and Sexton, 2012; Monetate, 2014a,b) show that while 63% of organizations see big data analytics as a competitive advantage, 80% of marketers say they don’t know how to translate data into action and that 95% of data within organizations remain unused. Even more perplexing, one survey

\(^1\)In 2010, Eric Schmidt, the CEO of Google, famously mentioned that every two days, the world creates as much information as it did from the dawn of civilization up until 2003, or five exabytes of data (1 exabyte = 1 billion gigabytes).
(Allen et al., 2005) shows that while 80% of CEOs believe they deliver superior customer experience, only 8% of customers agree. Big data analytics in e-commerce is still at its nascent stage and there is much to be learned in all aspects of e-commerce. Particularly lacking is the area of social media marketing (specifically, content engineering for better engagement) and the impact of recommender systems, in which there are little to no large-scale level analyses or much disagreement on what strategies actually work.

Big data analytics is challenging because successful analytics on big data often require a combination of both data mining and econometrics: Data mining to reduce or structure (from unstructured data such as text, photo, and video) large-scale data and econometric analyses to truly understand and assign causality to interesting patterns. In my dissertation, I study how firms can better utilize big data analytics and specific applications of machine learning techniques for improved e-commerce using theory-driven econometrical and experimental studies. Specifically, in the first essay, I investigate how firms can actively content engineer their social media page postings (e.g., Facebook Pages and Twitter) to better engage connected consumers. In the second essay, I investigate how different recommender algorithms on e-commerce sites (e.g., Amazon.com’s “Consumers who purchased this also purchased”) influence sales volume and diversity. In the third essay, I plan to study how product attributes and reviews moderate the performance of recommender systems. Based on completed results, it can be observed that big data analytics that combine data mining and econometrical studies can provide readily actionable strategies to improve many aspects of e-commerce with far-reaching economical consequences. A detailed description of three essays is given below.

**Essay 1- The Effect of Social Media Marketing Content on Consumer Engagement: Evidence from Facebook**  
We investigate the effect of social media content on customer engagement using a large-scale field study on Facebook. We content-code more than 100,000 unique messages across 800 companies engaging with users on Facebook using a combination of Amazon Mechanical Turk and state-of-the-art Natural Language Process-
ing algorithms. We use this large-scale database of advertising attributes to test the effect of ad content on subsequent user engagement defined as Likes and comments – with the messages. We develop methods to account for potential selection biases that arise from Facebook’s filtering algorithm, EdgeRank, that assigns posts non-randomly to users. We find that inclusion of persuasive content – like emotional and philanthropic content – increases engagement with a message. We find that informative content – like mentions of prices, availability and product features – reduce engagement when included in messages in isolation, but increase engagement when provided in combination with persuasive attributes. Persuasive content thus seems to be the key to effective engagement. Our results inform advertising design in social media, and the methodology we develop to content-code large-scale textual data provides a framework for future studies on unstructured natural language data such as advertising content or product reviews.

Essay 2- “People Who Liked This Study Also Liked”: An Empirical Investigation of the Impact of Recommender Systems on Sales Volume and Diversity

We investigate the impact of collaborative filtering recommender algorithms (e.g., Amazon.com’s “Customers who bought this item also bought”), commonly used in e-commerce, on sales volume and diversity. We use data from a randomized field experiment on movie sales run by a top retailer in North America. For sales volume, we show that different algorithms have differential impacts. Purchase-based collaborative filtering (“Customers who bought this item also bought”) causes a 25% lift in views and a 35% lift in the number of items purchased over the control group (no recommender). In contrast, View-based collaborative filtering (“Customers who viewed this item also viewed”) shows only a 3% lift in views and a 9% lift in the number of items purchased, albeit not statistically significant. For sales diversity, we find that collaborative filtering algorithms cause individuals to discover and purchase a greater variety of products but push users to the same set of titles, leading to concentration bias at the aggregate level. We show that this differential impact on individual versus aggregate diversity is caused by users exploring into only a few 'pathway' popular genres. That is, the recommenders were more effective in aiding discovery for a
few popular genres rather than uniformly aiding discovery in all genres. For managers, our results inform personalization and recommender strategy in e-commerce. From an academic standpoint, we provide the first empirical evidence from a randomized field experiment to help reconcile opposing views on the impact of recommenders on sales diversity.

Essay 3- When do Recommender Systems Work the Best? The Moderating Effects of Product Attributes and Consumer Reviews on Recommender Performance

We investigate the moderating effect of product attributes and consumer reviews on the efficacy of a collaborative filtering recommender system on an e-commerce site. We run a randomized field experiment on a top North American retailer’s website with 184,375 users split into a recommender-treated group and a control group with 37,215 unique products in the dataset. By augmenting the dataset with Amazon Mechanical Turk tagged product attributes and consumer reviews from the website, we study their moderating influence on recommenders in generating conversion.

We first confirm that the use of recommenders increases the baseline conversion rate by 5.9%. We find that recommenders act as substitutes for high average review ratings and review volumes with the effect of using recommenders increasing the conversion as much as about two additional average star ratings. Additionally, we find that positive impact on conversion from recommenders are greater for hedonic products compared to utilitarian products while search-experience quality did not have any impact. Lastly, we find that the higher the price, the lower the positive impact of recommenders, while providing more product descriptions increased the recommender effectiveness.

For managers, we 1) identify the products with which to use recommenders and 2) show how other product information sources on e-commerce sites interact with recommenders. From an academic standpoint, we provide insight into the underlying mechanism behind how recommenders cause consumers to purchase.
CHAPTER 2 : The Effect of Social Media Marketing Content on Consumer Engagement: Evidence from Facebook

2.1. Introduction

Social networks are increasingly taking up a greater share of consumers’ time spent online. As a result, social media — which includes advertising on social networks and/or marketing communication with social characteristics — is becoming a larger component of firms’ marketing budgets. Surveying 4,943 marketing decision makers at U.S. companies, the 2013 Chief Marketing Officer survey (www.cmosurvey.org) reports that expected spending on social media marketing will grow from 8.4% of firms’ total marketing budgets in 2013 to about 22% in the next five years. As firms increase their social media activity, the role of content engineering has become increasingly important. Content engineering seeks to develop content that better engages targeted users and drives the desired goals of the marketer from the campaigns they implement. This raises the question: what content works best? The most important body of academic work on this topic is the applied psychology and consumer behavior literature which has discussed ways in which the content of marketing communication engages consumers and captures attention. However, most of this work has tested and refined theories about content primarily in laboratory settings. Surprisingly, relatively little has been explored systematically about the empirical consequences of advertising and promotional content in real-world, field settings outside the laboratory. Despite its obvious relevance to practice, Marketing and advertising content is also relatively under emphasized in economic theory. The canonical economic model of advertising as a signal (c.f. Nelson (1970); Kihlstrom and Riordan (1984); Milgrom and Roberts (1986)) does not postulate any direct role for ad content because advertising intensity conveys all relevant information about product quality in equilibrium to market participants. Models of informative advertising (c.f. Butters (1977); Grossman and Shapiro (1984)) allow for advertising to inform agents only about price and product existence — yet, casual observation and several studies in lab settings (c.f. Armstrong (2010); Berger (2012)) suggest that adver-
Advertisements contain much more information and content beyond prices. In this paper, we explore the role of content in driving consumer engagement in social media in a large-scale field setting. We document the kinds of content used by firms in practice. We show that a variety of emotional, philanthropic, and informative advertising content attributes affect engagement and that the role of content varies significantly across firms and industries. The richness of our engagement data and the ability to content code social media messages in a cost-efficient manner enables us to study the problem at a larger scale than much of the previous literature on the topic.

Our analysis is of direct relevance to industry in better understanding and improving firms’ social media marketing strategies. Many industry surveys (Ascend2, 2013; Gerber, 2014; eMarketer, 2013a; SmartBrief, 2010; Ragan and Solutions, 2012) report that achieving engagement on large audience platforms like Facebook is the top most important social media marketing goals for consumer-facing firms.\(^1\)Social media marketing agencies’s financial arrangements are increasingly contracted on the basis of the engagement these agencies promise to drive for their clients. In the early days of the industry, it was thought that engagement was primarily driven by the volume of users socially connected to the brand by increasing the reach of posts released by the firms. Accordingly, firms aggressively acquired fans and followers on platforms like Facebook by investing heavily in ads on the network. However, early audits of the data (e.g., Creamer 2012) suggested that only about 1% of an average firm’s Facebook fans show any engagement with the brand by Liking, sharing, or commenting on messages by the brand on the platform. As a result, industry attention shifted from acquisition of social media followers per se, to the design of content that achieves both better reach and engagement amongst social media followers, especially since the design of websites like Facebook also uses current engagement level to determine firms’ future reach. In a widely reported example that reflects this trend (WSJ, 2012), General Motors curtailed its annual spending of $10M on Facebook’s paid ads (a vehicle for acquir-

---

\(^1\)With the percentage of marketers who say so varying from 60% to more than 90% across different surveys.
ing new fans for the brand), choosing instead to focus on creating content for its branded Facebook Page, on which it spent $30M. While attention in industry has shifted towards content in this manner, industry still struggles with understanding what kinds of content work better for which firms and in what ways. For example, are messages seeking to inform consumers about product or price attributes more effective than persuasive messages with humor or emotion? Do messages explicitly soliciting user response (e.g., “Like this post if . . .”) draw more engagement or in fact turn users away? Does the same strategy apply across different industries? Our paper systematically explores these kinds of questions and contributes to the formulation of better content engineering policies in practice.\(^2\)

Our empirical investigation is implemented on Facebook, which is the largest social media platform in the world. As alluded to above, many top brands now maintain a Facebook page from which they serve posts and messages to connected users. This is a form of free social media marketing that has increasingly become a popular and important channel for marketing. Our data comprises information on about 100,000 such messages posted by a panel of about 800 firms over a 11-month period between September 2011 and July 2012. For each message, our data also contains time-series information on two kinds of engagement measures — Likes and comments — observed on Facebook. In addition, we have cross-sectional data on shares and click-throughs. We supplement these engagement data with message attribute information that we collect using a large-scale survey we implement on Amazon Mechanical Turk (henceforth “AMT”), combined with a Natural Language Processing algorithm (henceforth “NLP”) we build to tag messages. We incorporate new methods and procedures to improve the accuracy of content tagging on AMT and our NLP algorithm. As a result, our algorithm achieves great accuracy, recall, and precision under 10-fold cross validation for almost all tagged content profiles.\(^3\)

\(^2\)As of December 2013, industry-leading social media analytics firms such as Wildfire (now part of Google) do not offer detailed content engineering analytics connecting a wide variety of social media content with real engagement data. Rather, to the best of our knowledge, they provide simpler analytics such as optimizing the time-of-the-day or day-of-the-week to post and whether to include pictures or videos.

\(^3\)The performance of NLP algorithms are typically assessed on the basis of accuracy (the total % correctly classified), precision (out of predicted positives, how many are actually positive), and recall (out of actual positives, how many are predicted as positives). An important tradeoff in such algorithms is that an increase
develop will be useful in future studies analyzing other kinds of advertising content and product reviews.

Our data has several advantages that facilitate a detailed study of content. First, Facebook messages have rich content attributes (unlike say, Twitter tweets, which are restricted to 140 characters) and rich data on user engagement. Second, Facebook requires real names and, therefore, data on user activity on Facebook is often more reliable compared to other social media sites. Third, engagement is measured on a daily basis (panel data) by actual message-level engagement such as Likes and comments that are precisely tracked within a closed system. These aspects make Facebook an almost ideal setting to study the role of content for this type of marketing communication.

Our strategy for coding content is motivated by the psychology, marketing and economic literatures on advertising (see Cialdini (2001); Bagwell (2007); Berger (2012); Chandy et al. (2001); Vakratsas and Ambler (1999) for some representative overviews). In the economics literature, it is common to classify advertising as informative (shifting beliefs about product existence or prices) or persuasive (shifting preferences directly). The basis of informative content is limited to prices and/or existence, and persuasive content is usually treated as a “catch-all” without finer classification. Rather than this coarse distinction, our classification follows the seminal classification work of Resnik and Stern (1977), who operationalize informative advertising based on the number of characteristics of informational cues (see Abernethy and Franke, 1996 for an overview of studies in this stream). Some criteria for classifying content as informative include details about products, promotions, availability, price, and product related aspects that could be used in optimizing the purchase decision. Following this stream, any product oriented facts, and brand and product mentions are categorized as informative content. Following suggestions in the persuasion literature (Cialdini, 2001; Nan and Faber, 2004; Armstrong, 2010; Berger, 2012), we classify “persuasive” content as those that broadly seek to influence by appealing to ethos, pathos, and in precision often causes decrease in recall or vice versa. This tradeoff is similar to the standard bias-variance tradeoff in estimation.
logos strategies. For instance, the use of a celebrity to endorse a product or attempts to gain trust or good-will (e.g., via small talk, banter) can be construed as the use of ethos — appeals through credibility or character — and a form of persuasive advertising. Messages with philanthropic content that induce empathy can be thought of as an attempt at persuasion via pathos — an appeal to a person’s emotions. Lastly, messages with unusual or remarkable facts that influence consumers to adopt a product or capture their attention can be categorized as persuasion via logos — an appeal through logic. We categorize content that attempt to persuade and promote relationship building in this manner as persuasive content. Though we believe we consider a larger range of content attributes than the existing literature, it is practically impossible to detail the full range of possible content profiles produced on a domain as large as Facebook (or in a data as large as ours). We choose content profiles that reflect issues flagged in the existing academic literature and those that are widely used by companies on Facebook. We discuss this in more detail in Section 2.2.

Estimation of the effect of content on subsequent engagement is complicated by the non-random allocation of messages to users implemented by Facebook via its EdgeRank algorithm. EdgeRank tends to serve to users messages that are newer and are expected to appeal better to his/her tastes. We account for the selection induced by EdgeRank by developing a semi-parametric correction for the filtering it induces. One caveat to the correction is that it is built on prior (but imperfect) knowledge of how EdgeRank is implemented. In the absence of additional experimental/exogenous variation, we are unable to address all possible issues with potential nonrandom assignment perfectly. We view our work as a large-scale, and relatively exhaustive exploratory study of content variables in social media that could be the basis of further rigorous testing and causal assessment, albeit at a more limited scale. A fully randomized large-scale experiment that provides a cross-firm and cross-industry assessment like provided here may be impossible or cost-prohibitive to implement, and hence, we think a large-scale cross-industry study based on field data of this sort is valuable.

Our main finding from the empirical analysis is that persuasive content drives social media
engagement significantly. Additionally, informative content tends to drive engagement positively only when combined with such content. Persuasive content thus seem to be the key to effective content engineering in this setting. This finding is of substantive interest because most firms post messages with one content type or other, rather than in combination. Our results suggest therefore that there may be substantial gains to content engineering by combining characteristics. The empirical results also unpack the persuasive effect into component attribute effects and also estimate the heterogeneity in these effects across firms and industries, enabling fine tuning these strategies across firms and industries.

Our paper adds to a growing literature on social media. Studies have examined the the diffusion of user-generated content (Susarla et al., 2012) and their impact on firm performance (Rui et al., 2013; Dellarocas, 2006). A few recent papers have also examined the social media strategies of firms, focusing primarily on online blogs and forums. These include studies of the impacts of negative blog messages by employees on blog readership (Aggarwal et al., 2012), blog sentiment and quality on readership (Singh et al., 2014), social product features on consumer willingness to pay (Oestreicher-Singer and Zalmanson, 2013), and the role of active contributors on forum participation (Jabr et al., 2014). We add to this literature by examining the impact of firms’ content strategies on user engagement.

An emerging theoretical literature in advertising has started to investigate the effects of content. This includes new models that allow ad content to matter in equilibrium by augmenting the canonical signaling model in a variety of ways (e.g. Anand and Shachar (2009)) by allowing ads to be noisy and targeted; Anderson and Renault (2006) by allowing ad content to resolve consumers’ uncertainty about their match-value with a product; and Mayzlin and Shin (2011) and Gardete (2013) by allowing ad content to induce consumers to search for more information about a product). Our paper is most closely related to a small empirical literature that has investigated the effects of ad content in field settings. These include Bertrand et al. (2010) (effect of direct-mail ad content on loan demand); Anand and Shachar (2011); Liaukonyte et al. (2013) (effect of TV ad content on viewership
and online sales); Tucker (2012b) (effect of ad persuasion on YouTube video sharing) and Tucker (2012a) (effect of “social” Facebook ads on philanthropic participation). Also related are recent studies exploring the effect of content more generally (and not specifically ad content) including Berger and Milkman (2012) (effect of emotional content in New York Times articles on article sharing) and Gentzkow and Shapiro (2010) (effect of newspaper’s political content on readership). Relative to these literatures, our study makes two main contributions. First, from a managerial standpoint, we show that while persuasive ad content — especially emotional and philanthropic content — positively impacts consumer engagement in social media, informative content has a negative effect unless it is combined with persuasive content attributes. This can help drive content engineering policies in firms. We also show how the effects differ by industry type. Second, none of the prior studies on ad content have been conducted at the scale of this study, which spans a large number of industries. The rigorous content-tagging methodology we develop, which combines surveys implemented on AMT with NLP-based algorithms, provides a framework to conduct large-scale studies that analyze the content of marketing communication.

Finally, the reader should note we do not address the separate but important question of how engagement affects product demand and firm’s profits so as to complete the link between ad-attributes and those outcome measures. First, the data required for the analysis of this question at a scale comparable to this study are still not widely available to researchers. Second, as mentioned, firms and advertisers care about engagement per se and are willing to invest in advertising for generating engagement, rather than caring only about sales. This is consistent with our view that advertising is a dynamic problem and a dominant role of advertising is to build long-term brand-capital for the firm. Even though the current period effects of advertising on demand may be small, the long-run effect of advertising may be large, generated by intermediary activities like increased consumer engagement, increased awareness and inclusion in the consumer consideration set. Thus, studying the formation and evolution of these intermediary activities — like engagement — is worthwhile in order to better understand the true mechanisms by which advertising affects outcomes in market
settings. We note other papers such as Kumar et al. (2013); Goh et al. (2013); Rishika et al. (2013); Li and Wu (2013); Miller and Tucker (2013); Sunghun et al. (2014); Luo and Zhang (2013); Luo et al. (2013) as well as industry reports (comScore, 2013; Chadwick-Martin-Bailey, 2010; 90octane, 2012; HubSpot, 2013) have linked the social media engagement measures we consider to customer acquisition, sales, and profitability metrics.

2.2. Data

Our dataset is derived from the “pages” feature offered by Facebook. The feature was introduced on Facebook in November 2007. Facebook Pages enable companies to create profile pages and to post status updates, advertise new promotions, ask questions and push content directly to consumers. The left panel of Figure 1 shows an example of Walmart’s Facebook Page, which is typical of the type of pages large companies host on the social network. In what follows, we use the terms pages, brands, and firms interchangeably. Our data comprises posts served from firms’ pages onto the Facebook profiles of the users that are linked to the firm on the platform. To fix ideas, consider a typical message (see the right panel of Figure 1): “Pretty cool seeing Andy giving Monfils some love... Check out what the pros are wearing here: http://bit.ly/nyiPew.” In this status update, a tennis equipment retailer starts with small talk, shares details about a celebrity (Andy Murray and Gael Monfils) and ends with link to a product page. Each such message is a unit of analysis in our data.

2.2.1. Data Description

Raw Data and Selection Criteria

To collect the data, we partnered with an anonymous firm, henceforth referred to as Company X that provides analytics services to Facebook Page owners by leveraging data from Facebook’s Insights. Insights is a tool provided by Facebook that allows page owners to monitor the performance of their Facebook messages. Company X augments data from

\textsuperscript{4}Retailer picked randomly from an online search; not necessarily from our data.
Facebook Insights across a large number of client firms with additional records of daily message characteristics, to produce a raw dataset comprising a message-day-level panel of messages posted by companies via their Facebook pages. The data also includes two consumer engagement metrics: the number of Likes and comments for each message each day. These metrics are commonly used in industry as measures of engagement. They are also more granular than other metrics used in extant research such as the number of fans who have Liked the page. Also available in the data are the number of impressions of each message per day (i.e., the total number of users the message is exposed to - we have both the unique user impression and the total impression). In addition, page-day level information such as the aggregate demographics of users (fans) who Liked the page on Facebook or have ever seen messages by the page are collected by Company X on a daily level. This comprises the population of users a message from a firm can potentially be served to. We leverage this information in the methodology we develop later for accounting for non-random assignment of messages to users by Facebook. Once a firm serves a message, the message’s impressions, Likes, and comments are recorded daily for an average of about 30 days (maximum:
The raw data contains about a million unique messages by about 2,600 unique companies. The reader should note that as of this writing, our data is the most complete observational data available outside of Facebook — the data includes details such as demographics of page fans and engaged fans, which cannot be scraped by outsiders (but are essential for correcting for EdgeRank) but are available only to the page owners via Facebook’s Application Programming Interface. Our data also includes daily snapshots of message-level engagement that Facebook does provide to page owners (Page owners must take snapshots themselves if they want this data). These daily snapshots generate the within-message variation that enables the panel analysis in our paper. Finally, page-owners do not have access to data on performance of any messages by other pages, unlike our dataset which spans a large number of companies across sectors.

We clean the data to reflect the following criteria:

- Only pages located in the US, and,
- Only messages written in English, and,
- Only messages with complete demographics data.

After cleaning, the data span 106,316 unique messages posted by 782 companies (including many large brands) between September 2011 and July 2012. This results in about 1.3 million rows of message-level daily snapshots recording about 450 million page fans’ responses. Removing periods after which no significant activity is observed for a message reduces this to 665,916 rows of message-level snapshots (where activity is defined as either impressions, Likes, or comments). The companies in our dataset are categorized into 6 broader industry categories following Facebook’s page classification criteria: Celebrities & Public Figure (e.g., Roger Federer), Entertainment (e.g., Star Trek), Consumer Products & Brands (e.g., Tesla

---

5A vast majority of messages do not get any impression or engagement after 7 days. After 15 days, virtually all engagements and impressions (more than 99.9%) are accounted for. Unfortunately, reliable tabulation of “shares” is not available in the data.
Motors), Organizations & Company (e.g., WHO), Websites (e.g., TED), Local Places & Businesses (e.g., MoMA).

**Content-coded Data**

We use a two-step method to label content. First, we contract with workers through AMT and tag 5,000 messages for a variety of content profiles. Subsequently, we build an NLP algorithm by combining several statistical classifiers and rule-based algorithms to extend the content-coding to the full set of 100,000 messages. This algorithm uses the 5,000 AMT-tagged messages as the training data-set. We describe these methods in more detail later in the paper.

The content in Facebook messages can be categorized as informative, persuasive, or both. Some messages inform consumers about deals and discounts about products, while other messages seek to connect with consumers on a personal level to promote brand personality, form relationships and are social in nature. We call the first type informative content, and the second persuasive content. Some messages do both at the same time by including casual banter and product information simultaneously (e.g., “Are you a tea person or a coffee person? Get your favorite beverage from our website: http://www.specific-link-here.com”).

Table 1 outlines the finer classification of the attributes we code up, including precise definitions, summary statistics, and the source for coding the attribute. In Table 1, the 8 variables: BRANDMENTION, DEAL, PRICECOMPARE, PRICE, TARGET, PRODAVAIL, PRODLOCATION, and PRODMENTION are informative. These variables enable us to assess the effect of search attributes, brand, price, and product availability information on engagement. The 8 variables: REMFACT, EMOTION, EMOTICON, HOLIDAYMENTION, HUMOR, PHILANTHROPIC, FRIENDLIKELY, and SMALLTALK are classified as persuasive. These definitions include emotional content, humor, banter, and philanthropic content.
Informative content variables are identified using the seminal work by Resnik and Stern (1977), which provides an operational guideline to identify informative content. Their work provides fourteen evaluative content criteria to identify informative content that includes content such as product price and availability. Our persuasive content are identified mostly from existing consumer behavior research. For example, emotional and humorous content have been identified as drivers of word-of-mouth and of viral marketing tactics (Porter and Golan, 2006; Berger, 2012, 2011; Berger and Milkman, 2012). Philanthropic content has been studied in the context of advertising effectiveness (Tucker, 2012a). Similarly, Berger and Schwartz (2011) documented that the interestingness of content such as mentions of remarkable facts is effective in generating word-of-mouth. For a survey of papers motivating our choice of persuasive variables, see Berger (2012). While not fully exhaustive, we have attempted to cover most variables that are 1) highlighted by prior academic research to be relevant, 2) commonly discussed and used in the industry.

Besides these main variables of interest, controls and content-related patterns noted as important in industry reports are profiled. We include these content categories to investigate more formally considerations laid out in industry white papers, trade-press articles, and blogs about the efficacy of message attributes in social media engagement. It includes content that explicitly solicits readers to comment or includes blanks for users to fill out (thus providing an explicit option to facilitate engagement). Additionally, characteristics like whether the message contained photos, website links, and the types of the page-owner (e.g., business organization versus celebrity) are also coded. Other message-specific characteristics and controls include metrics such as message length in characters and SMOG (“Simple Measure of Gobbledygook”), an automatically computed reading complexity index that is used widely. Higher values of SMOG implies a message is harder to read. Table 2 shows sample messages taken from Walmart’s page in December 2012 and shows how we would have tagged them. The reader should note that some elements of content tagging and classification are necessarily subjective and based on human judgement. We discuss our methods (which involve obtaining agreement across 9 tagging individuals) in section 2.2.2.
All things considered, we believe this is one of the most comprehensive attempts at tagging marketing communication related content in the empirical literature.

**Data Descriptive Graphics**

This section presents descriptive statistics of the main stylized patterns in the data. The first thing we would like to report is what kinds of content are used by firms (this may be useful for instance, for a researcher interested in studying the role of specific content profiles in Facebook, who would like to know what content variables are used a lot by firms). Table 1 reports on the mean proportion of messages that have each content characteristic. One can see that messages with videos, product or holiday mentions or emoticons are relatively uncommon, while those with smalltalk and with information about where to obtain the product (location/distribution attributes) are very common. Figure 2 reports on the co-occurrence of the various attributes across messages. The patterns are intuitive. For instance, emotional and philanthropic content co-occur often, so does emotional and friend-like content, as well as content that describes product deals and availability. To better describe the correlation matrix graphically and to cluster highly correlated variables together, we ran cluster analysis to determine the optimal number of clusters using the average silhouette width (Rousseeuw, 1987), which suggested that there are two clusters in the data. Figure 2 shows via a solid line how content types are clustered across messages. We see that persuasive content types and informative content types are split into two separate clusters, suggesting that firms typically tend to use one or the other in their messages. Later in the paper, we show evidence suggesting that this strategy may not be optimal. Figure 3 shows the percentage of messages featuring a content attribute split by industry category. We represent the relative percentages in each cell by the size of the bubbles in the chart. The largest bubble is SMALLTALK for the celebrities category (60.4%) while the smallest is PRICECOMPARE for the celebrities category (0%). This means that 6 in 10 messages by celebrity pages in the data have some sort of small talk (banter) and/or content that does

---

*Clustered with hierarchical clustering.*
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAU (τ)</td>
<td>Time since the post release (Day)</td>
<td>Facebook</td>
<td>6.253</td>
<td>3.657</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>LIKES</td>
<td>Number of “Likes” post has obtained</td>
<td>Facebook</td>
<td>48.373</td>
<td>19.0</td>
<td>0</td>
<td>32453</td>
</tr>
<tr>
<td>COMMENTS</td>
<td>Number of “Comments” post has obtained</td>
<td>Facebook</td>
<td>7.362</td>
<td>2.991</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>IMPRESSIONS</td>
<td>Number of times message was shown to users (unique)</td>
<td>Facebook</td>
<td>9969.2</td>
<td>129874</td>
<td>4.5×10⁷</td>
<td></td>
</tr>
<tr>
<td>SMOG</td>
<td>SMOG readability index (higher means harder to read)</td>
<td>Computed</td>
<td>157.41</td>
<td>134.54</td>
<td>1</td>
<td>6510</td>
</tr>
<tr>
<td>HTTP</td>
<td>Message contains a link</td>
<td>Computed</td>
<td>0.353</td>
<td>0.478</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>QUESTION</td>
<td>Message contains questions</td>
<td>Computed</td>
<td>0.358</td>
<td>0.479</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BLANK</td>
<td>Message contains blanks (e.g. “My favorite artist is ___”)</td>
<td>Computed</td>
<td>0.010</td>
<td>0.099</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ASKLIKE</td>
<td>Explicit solicitation for “Likes” (e.g. “Like if . . .”)</td>
<td>Computed</td>
<td>0.006</td>
<td>0.080</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ASKCOMMENT</td>
<td>Explicit solicitation for “Comments”</td>
<td>Computed</td>
<td>0.001</td>
<td>0.029</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persuasive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REMFACT</td>
<td>Remarkable fact mentioned</td>
<td>AMT</td>
<td>0.527</td>
<td>0.499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>EMOTION</td>
<td>Any type of emotion present</td>
<td>AMT</td>
<td>0.524</td>
<td>0.499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>EMOTICON</td>
<td>Contains emoticon or net slang (approximately 1000 scraped from</td>
<td>Computed</td>
<td>0.012</td>
<td>0.108</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HOLIDAYMENTION</td>
<td>Mentions US Holidays</td>
<td>Computed</td>
<td>0.006</td>
<td>0.076</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HUMOR</td>
<td>Humor used</td>
<td>AMT</td>
<td>0.375</td>
<td>0.484</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PHILANTHROPIC</td>
<td>Philanthropic or activist message</td>
<td>AMT</td>
<td>0.498</td>
<td>0.509</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FRIENDLIKELY</td>
<td>Answer to question: “Are your friends on social media likely to post message such as the shown”?</td>
<td>AMT</td>
<td>0.533</td>
<td>0.499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SMALLTALK</td>
<td>Contains small talk or banter (defined to be content other than</td>
<td>AMT</td>
<td>0.852</td>
<td>0.355</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRANDMENTION</td>
<td>Mentions a specific brand or organization name</td>
<td>AMT+Comp</td>
<td>0.264</td>
<td>0.441</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>DEAL</td>
<td>Contains deals: any type of discounts and freebies</td>
<td>AMT</td>
<td>0.620</td>
<td>0.485</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRICECOMPARE</td>
<td>Compares price or makes price match guarantee</td>
<td>AMT</td>
<td>0.442</td>
<td>0.497</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRICE</td>
<td>Contains product price</td>
<td>AMT+Comp</td>
<td>0.051</td>
<td>0.220</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TARGET</td>
<td>Message is targeted towards an audience segment (e.g. demographics, certain qualifications such as “Moms”)</td>
<td>AMT</td>
<td>0.530</td>
<td>0.499</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRODVAAIL</td>
<td>Contains information on product availability (e.g. stock and release data)</td>
<td>AMT</td>
<td>0.557</td>
<td>0.497</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRODLOCATION</td>
<td>Contains information on where to obtain product (e.g. link or physical location)</td>
<td>AMT</td>
<td>0.690</td>
<td>0.463</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRODMENTION</td>
<td>Specific product has been mentioned</td>
<td>AMT+Comp</td>
<td>0.146</td>
<td>0.353</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MSGTYPE</td>
<td>Categorical message type assigned by the Facebook</td>
<td>Facebook</td>
<td>0.099</td>
<td>0.299</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>APP</td>
<td>Application related messages</td>
<td>Facebook</td>
<td>0.389</td>
<td>0.487</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Link</td>
<td></td>
<td>Facebook</td>
<td>0.366</td>
<td>0.481</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Photo</td>
<td></td>
<td>Facebook</td>
<td>0.140</td>
<td>0.347</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Status Update</td>
<td></td>
<td>Facebook</td>
<td>0.005</td>
<td>0.073</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Video</td>
<td></td>
<td>Facebook</td>
<td>0.088</td>
<td>0.283</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>


Table 1: Variable Descriptions and Summary for Content-coded Data: To interpret the “Source” column, note that “Facebook” means the values are obtained from Facebook, “AMT” means the values are obtained from Amazon Mechanical Turk and “Computed” means it has been either calculated or identified using online database resources and rule-based methods in which specific phrases or content (e.g. brands) are matched. Finally, “AMT+Computed” means primary data has been obtained from Amazon Mechanical Turk and it has been further augmented with online resources and rule-based methods.
Sample Messages

*Cheers! Let Welch's help ring in the New Year.*

BRANDMENTION, SMALLTALK, HOLIDAYMENTION, EMOTION

*Maria’s mission is helping veterans and their families find employment. Like this and watch Maria’s story.*

http://walmarturl.com/VsWFh

PHILANTHROPIC, SMALLTALK, ASKLIKE, HTTP

*On a scale from 1--10 how great was your Christmas?*

SMALLTALK, QUESTION, HOLIDAYMENTION

*Score an iPad 3 for an iPad2 price! Now at your local store, $50 off the iPad 3. Plus, get a $30 iTunes Gift Card. Offer good through 12/31 or while supplies last.*

PRODMENTION, DEAL, PRODLOCATION, PRODAVAIL, PRICE

*They’re baaaaaack! Now get to snacking again. Find Pringles Stix in your local Walmart.*

EMOTION, PRODMENTION, BRANDMENTION, PRODLOCATION

Table 2: Examples of Messages and Their Content Tags: The messages are taken from 2012 December messages on Walmart’s Facebook page.

not relate to products or brands; and that there are no messages by celebrity owned pages that feature price comparisons. “Remarkable facts” (our definition) are posted more by firms in the entertainment category and less by local places and businesses. Consistent with intuition, consumer product pages and local places/businesses post the most about products (PRODMENTION), product availability (PRODAVAIL), product location (PRODLOC), and deals (DEAL). Emotional (EMOTION) and philanthropic (PHILAN) content have high representation in pages classified as celebrity, organization, and websites. Similarly, the AMT workers identify a larger portion of messages posted by celebrity, organization and website-based pages to be similar to messages by friends.

We now discuss the engagement data. Figure 4 shows box plots of the log of impressions, Likes, and comments versus the time (in days) since a message is released ($\tau$). Both comments and Likes taper off to zero after two and six days respectively. The rate of decay of impressions is slower. Virtually all engagements and impressions (more than 99.9%) are accounted for within 15 days of release of a message.
Figure 2: Co-occurrence of Attribute Characteristics Across messages. Shades in upper triangle represent correlations. Numbers in lower triangle represent the same correlations in numerical form in 100-s of units (range $-100, +100$). For e.g., the correlation in occurrence of smalltalk and humor across messages is 0.26 (cell [3,2]). The dark line shows the separation into 2 clusters. Persuasive content and informative content attributes tend to form two separate clusters.

Figure 5 shows the average number of Likes and comments by message type (photo, link, etc.) over the lifetime of a message. Messages with photos have the highest average Likes (94.7) and comments (7.0) over their lifetime. Status updates obtain more comments (5.5) on average than videos (4.6) but obtain less Likes than videos. Links obtain the lowest Likes on average (19.8) as well as the lowest comments (2.2). Figure 6 shows the same bar plots split across 6 industry categories. A consistent pattern is that messages with photos always obtain highest Likes across industries. The figure also documents interesting heterogeneity.
Figure 3: Bubble Chart of Broader Industry Category vs Message Content: Each bubble represents the percentage of messages within a row-industry that has the column-attribute. Computed for the 5000 tagged messages. Larger and lighter bubbles imply higher percentage of messages in that cell. Percentages do not add up to 100 along rows or columns as any given message can have multiple attributes included in it. The largest bubble (60.4%) corresponds to SMALLTALK for the celebrity page category and the smallest bubble (0%) corresponds to PRICECOMPARE for the celebrity category.

Figure 7 presents the average number of Likes and comments by content attribute. Emotional messages obtain the most number of Likes followed by messages identified as “likely to be posted by friends” (variable: FRIENDLIKELY). Emotional content also obtain the highest number of comments on average followed by SMALLTALK and FRIENDLIKELY. The reader should note these graphs do not account for the market-size (i.e., the number of impressions a message reached). Later, we present an econometric model that incorporates market-size as well as selection by Facebook’s filtering algorithm to assess user engagement more formally.
Figure 4: Box Plots of Log(engagement+1) vs Time since message Release: Three graphs show the box plots of (log) impressions, comments and Likes vs. $\tau$ respectively. Both comments and Likes taper to zero after two and six days respectively. Impressions take longer. After 15 days, virtually all engagements and impressions (more than 99.9%) are accounted for. There are many outliers.
Figure 5: Average *Likes* and Comments by Message Type: This figure shows the average number of *Likes* and comments obtained by messages over their lifetime on Facebook, split by message type.

Figure 6: Average *Likes* and Comments by Message Type by Industry: This figure shows the average number of *Likes* and comments obtained by messages over their lifetime split by message type for each industry.
Figure 7: Average Likes and Comments by Message Content: This figure shows the average number of Likes and comments obtained by messages over their lifetime split by message content.

2.2.2. Amazon Mechanical Turk (AMT)

We now describe our methodology for content-coding messages using AMT. AMT is a crowdsourcing marketplace for simple tasks such as data collection, surveys, and text analysis. It has now been successfully leveraged in several academic papers for online data collection and classification. To content-code our messages, we create a survey instrument comprising of a set of binary yes/no questions we pose to workers (or “Turkers”) on AMT. To ensure high quality responses from the Turkers, we follow several best practices identified in literature (e.g., we obtain tags from at least 9 different Turkers choosing only those who are from the U.S., have more than 100 completed tasks, and an approval rate more than 97%. We also include an attention-verification question.) Please see the appendix for the final survey instrument and the complete list of strategies implemented to ensure output quality.

Figure 8 presents the histogram of Cronbach’s Alphas, a commonly used inter-rater reliability measure, obtained for the 5,000 messages. The average Cronbach’s Alpha for our 5,000

---

*Recall, there are at least 9 Turker inputs per message. We calculate a Cronbach’s Alpha for each message by computing the reliability across the 9 Turkers, across all the content classification tasks associated with the message.*

24
tagged messages is 0.82 (median 0.84), well above typically acceptable thresholds of 0.7. About 87.5% of the messages obtained an alpha higher than 0.7, and 95.4% higher than 0.6. For robustness, we replicated the study with only those messages with alphas above 0.7 (4,378 messages) and found that our results are qualitatively similar.

At the end of the AMT step, approximately 2,500 distinct Turkers contributed to content-coding 5,000 messages. This constitutes the training dataset for the NLP algorithm used in the next step.

![Figure 8: Cronbach’s Alphas for 5,000 Messages](image)

**Figure 8: Cronbach’s Alphas for 5,000 Messages:** This bar graph shows the inter-rater reliability measure of Cronbach’s Alpha among at least 9 distinct Turkers’ inputs for each 5,000 messages. The mean is 0.82 and the median is 0.84. We replicated the study with only those above 0.7 and found the result to be robust.

### 2.2.3. Natural Language Processing (NLP) for Attribute Tagging

We use NLP techniques to label message content from Facebook messages using the AMT-labeled messages as the training data. Typical steps for such labeling tasks include: 1) breaking the sentence into understandable building blocks (e.g., words or lemmas) and identifying sentence-attributes similar to what humans do when reading; 2) obtaining a
set of training sentences with labels tagged from a trusted source identifying whether the sentences do or do not have a given content profile (in our case, this source comprise the 5000 AMT-tagged messages); 3) using statistical tools to infer which sentence-attributes are correlated with content outcomes, thereby learning to identify content in sentences. When presented with a new set of sentences, the algorithm breaks the sentence down to building blocks, identifies sentence-level attributes, and assigns labels using the statistical models that were fine-tuned in the training process. We summarize our method here briefly. A detailed description of the algorithms employed is presented in the Appendix.

The use of NLP techniques has been gaining traction in business research due to readily available text data online (e.g., Netzer et al. (2012); Ghose et al. (2012); Geva and Zahavi (2013)), and there are many different techniques. Our NLP methods closely mirror cutting edge multi-step methods used in the financial services industry to automatically extract financial information from textual sources (e.g., Hassan et al. (2011)) and are similar in flavor to winning algorithms from the recent Netflix Prize competition.\(^8\) The method we use combines five statistical classifiers with rule-based methods via heterogeneous “ensemble learning”. Statistical classifiers are binary classification machine learning models that take attributes as input and output predicted classification probabilities.\(^9\) Rule-based methods usually use large data sources (a.k.a dictionaries) or use specific if-then rules inputted by human experts, to scan through particular words or occurrences of linguistic entities in the messages to generate a classification. For example, in identifying brand and product mentions, we augment our AMT-tagged answers with several large lists of brands and products from online sources and a company list database from Thomson Reuters. Further, to increase the range of our brand name and product database, we also ran a separate

\(^8\)See [http://www.netflixprize.com](http://www.netflixprize.com).

\(^9\) We use a variety of different classifiers in this step including logistic regression with L1 regularization (which penalizes the number of attributes and is commonly used for attribute selection for problems with many attributes; see (Hastie et al., 2009)), Naive Bayes (a probabilistic classifier that applies Bayes theorem based on presence or absence of features), and support vector machines (a gold-standard algorithm in machine learning that works well for high dimensional problems) with L1 and L2 regularization and various kernels including linear, radial basis function, and polynomial kernels. We also utilize class-weighted classifiers and resampling method to account for imbalance in positive and negative labels.
AMT study with 20,000 messages in which we asked AMT Turkers to identify any brand or product name included in the message. We added all the brand and product names we harvested this way to our look-up database. We then utilize rule-based methods to identify brand and product mentions by looking up these lists. Similarly, in identifying emoticons in the messages, we use large dictionaries of text-based emoticons freely available on the internet.

Finally, we utilize ensemble learning methods that combine classifications from the many classifiers and rule-based algorithms we use. Combining classifiers is very powerful in the NLP domain since a single statistical classifier cannot successfully overcome the classic precision-recall tradeoff inherent in the classification problem. The final combined classifier has higher precision and recall than any of the constituent classifiers.

**Assessment**  We assess the performance of the overall NLP algorithm on three measures, viz., accuracy, precision, and recall (as defined in Footnote 3) using 10-fold cross-validation. 10-fold cross-validation is computationally intensive and makes it harder to achieve higher accuracy, precision and recall, but we find using the criterion critical to obtaining the external validity required for large scale classification. Table 3 shows these metrics for different content profiles. The performance is extremely good and comparable to performance achieved by the leading financial information text mining systems (Hassan et al., 2011). We also report the improvement of the final ensemble learning method relative to using only a support vector machine classifier. As shown, the gains from combining classifiers are very substantial. We obtain similar results for negative class labels.

As a final point of assessment, note that several papers in the management sciences using NLP methods implement *unsupervised* learning which does not require human-tagged data. These techniques use existing databases such as WordNet (lexical database for English) or tagged text corpus (e.g., tagged Brown Corpus) to learn content by patterns and correlations. *Supervised* NLP instead utilizes human-taggers to obtain a robust set of data that can be used to train the algorithm by examples. While unsupervised NLP is inexpensive, its
Table 3: Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation: This table presents metrics for performance of the classification algorithms used. The left 3 columns show the metrics for the final algorithm which combines classifiers via ensemble learning methods while the right 3 columns shows the metrics for a support vector machine algorithm. Notice that the support vector machine classifier tends to have low recall and high precision. Naive Bayes tends to have high recall but low precision. Classifiers on their own cannot successfully overcome the standard precision-recall tradeoff (if one is higher, the other is lower). But combining many different classifiers with ensemble learning can increase both precision and recall. We obtain similar results for negative class labels.

![Table](#)

<table>
<thead>
<tr>
<th>Category</th>
<th>With Ensemble Learning</th>
<th>Without Ensemble Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(The Best Performing Algorithm)</td>
<td>(Support Vector Machine version 1 + Rule-based)</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>Precision</td>
</tr>
<tr>
<td>REMFACT</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>EMOTION</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>HUMOR</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>PHILANTHROPIC</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>FRIENDLIKELY</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>SMALLTALK</td>
<td>0.85</td>
<td>0.88</td>
</tr>
<tr>
<td>DEAL</td>
<td>0.94</td>
<td>0.99</td>
</tr>
<tr>
<td>PRICECOMPARE</td>
<td>0.99</td>
<td>0.99</td>
</tr>
<tr>
<td>TARGETING</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>PRODAVAILABILITY</td>
<td>0.96</td>
<td>0.99</td>
</tr>
<tr>
<td>PRODLOCATION</td>
<td>0.97</td>
<td>0.99</td>
</tr>
</tbody>
</table>

performance is significantly poor compared to that of supervised NLP algorithms like the ones implemented here. Finally, To the best of our knowledge, the NLP method used in this paper that uses ensemble learning to combine several statistical classifiers and rule-based methods, has not been used in business research journals.\(^{10}\)

Further, several current implementations of NLP do not utilize the strict bar of utilizing the 10-fold cross-validation criterion. We believe one of the contributions of this paper is to demonstrate how to utilize AMT in combination with ensemble learning techniques, to implement supervised NLP in business research to produce robust and cost-efficient NLP algorithms that perform well at the scale required for empirical work. We believe the method will be useful in future studies on unstructured natural language data such as advertising content or product reviews. For

\(^{10}\)Although there exist business research papers combining statistical classifiers and rule-based algorithms, to our knowledge, none utilize ensemble learning methods which we find are critical in increasing accuracy, precision, and recall.
interested readers, a detailed step-by-step description of our NLP algorithm’s training and classification procedures is presented in the Appendix.

2.3. Empirical Strategy

Our empirical goal is to investigate the effect of message ad content on subsequent customer engagement. Engagement — the $y$-variable — is observed in the data; and content — the $x$-variables — has been tagged as above and is also observed. If messages are randomly allocated to users, the issue of assessing the effect of message-content on engagement is straightforward; one simply projects $y$ on $x$. Unfortunately, a complication arises because Facebook’s policy of delivery of messages to users is non-random: users more likely to find a message appealing are more likely to see the message in their newsfeed, a filtering implemented via Facebook’s “EdgeRank” algorithm. The filtering implies a selection problem in estimation of the effect of message-characteristics on engagement — if we see that messages with photos are more likely to be commented on by users, we do not know if this is the effect of including a photo in a message, or whether Facebook is more likely to show messages with photos to users who are more likely to comment on them. To our knowledge, the issue has been ignored in the literature on social media analysis so far.\(^{11}\)

We address the selection issue via a two-step procedure, first by building a semi-parametric model of “EdgeRank” that delivers an estimate of the expected number of impressions a message is likely to receive, and then, by incorporating this model to run a selectivity-corrected projection of Likes and comments on message characteristics in the second-stage. For the first-stage, we exploit the fact that we observe the aggregated decisions of Facebook to serve impressions to users, and that “EdgeRank” is based on three variables as revealed by Facebook: Type, Tie, and Time.\(^{12}\)

- **Type** ($z$) refers to the type of message. Facebook categorizes message-type into 5

\(^{11}\)We discuss later in this section why other sources of confounds (like direct targeting by firms) are second-order in this setting, compared to the selection induced by EdgeRank-based filtering.

\(^{12}\)As disclosed first at the 2010 “f8” conference. See [http://whatisEdgeRank.com](http://whatisEdgeRank.com) for a brief description of EdgeRank. For the duration of our data collection, this EdgeRank specification holds true.
classes: status update, photo, video, app, or link.

- *Tie* \((h_{ijt})\) refers to the affinity score between page \(j\) (company) and the Facebook user \(i\) (viewer of the message) at time \(t\) which is based on the strength and frequency of the interaction history between the user and the page.

- *Time* \((\tau)\) refers to the time since the message.

Our dataset contains direct observations on the variables Type and Time. We do not have individual-level data on a user’s history with pages to model tie strengths. However, we exploit the fact that we have access to demographics data on the set of users who could potentially have been shown a message, versus who were actually shown the message. The difference reflects the selection by EdgeRank, which we utilize as a proxy measure of Tie-strength based targeting. Since we do not know the exact functional form of EdgeRank’s targeting rule, we work with a semi-parametric specification, utilizing flexible splines to capture the effect of EdgeRank. At the end of this step, we thus develop a flexible approximation to EdgeRank’s targeting. In the second step, we can then measure the effect of ad content on *Likes* and comments, by controlling for the non-random targeting using our first-stage model. Figure 9 shows the empirical strategy visually. The advantages of directly modeling EdgeRank this way, are that 1) we are also able to predict which message would eventually reach users in addition to handing selection, which has auxiliary managerial value for advertisers seeking higher reach and 2) by separating Facebook’s impression mechanism from the effect of content on consumer engagement, we increase external validity of our results to realms outside of Facebook Pages.

2.3.1. First-stage: Approximating EdgeRank’s Assignment

We represent message \(k\)’s type in a vector \(z_k\), the time since message \(k\) was released in \(\tau_k\), and the history of user \(i\)’s past engagement with company \(j\) on Facebook in a vector \(h_{ijt}\). Table 4 summarizes the notation.
FB’s EdgeRank determines which fans to show a particular post to using 3Ts (Time, Type, Tie)

Notation Description

$i$ User

$j$ Firm

$k$ message

$t$ Time (day)

$z_k$ message $k$’s media type (5 options: photo, video, status update, app, link)

$\tau_k$ Time since message $k$ was released

$h_{ijt}$ History of user $i$’s past engagement with firm $j$

$g(\cdot)$ EdgeRank score approximating function

$n_{kjt}^{(d)}$ Impressions of message $k$ by page $j$ at time $t$ by users in demographics bin $d$

$N_{jt}^{(d)}$ Number of users of demographics bin $d$ who Liked page $j$ as of time $t$

$\theta^{(0)}(d)$ Intercept term for each demographics $d$

$\theta^{(1)}(d)$ Parameters in EdgeRank approximation for demographics bin $d$

Table 4: User-level Setup Notation.

To understand our procedure, let $n_{kjt}^{(d)}$ denote the number of users of demographic type $d = 1, \ldots, D$ who were shown message $k$ by firm $j$ at time $t$. We refer to $n_{kjt}^{(d)}$ as impressions. We observe $n_{kjt}$ directly, and $n_{kjt}^{(d)}$ is indirectly reported in the data and can be reverse-engineered from Company X’s reports. Let $N_{jt}^{(d)}$ denote the total number of users of demographic type $d$ for firm $j$ on day $t$ to whom the message can potentially be delivered. $N_{jt}^{(d)}$ is directly observed in the data, and comprises all users of demographics $d$ who have Liked the firm on Facebook. To be clear, note that Liking a message is different from Liking...
a page — Liking a page provides the firm that maintains that page an opportunity to serve its messages to that user via Facebook’s Newsfeed. \( N^{(d)}_{jt} \) is a count of all such users.

Now, note that by EdgeRank’s assignment rule, the aggregated impressions for demographic type \( d \), \( n^{(d)}_{kjt} \), is an (unknown) function of liked-fans \( N^{(d)}_{jt} \), the tie strength between users within demographic bucket \( d \) and the posting firm, \( h^{(d)}_{ijt} \), the type of message \( z_k \), and time since message release \( \tau_k \),

\[
E(n^{(d)}_{kjt}) = g(N^{(d)}_{jt}, h^{(d)}_{ijt}, z_k, \tau_k) \quad (2.1)
\]

We do not observe individual-level data on each user’s interaction with every message which could be the basis of estimating Equation (2.1). Instead, we can construct the aggregated number of impressions and liked-fans within a set of demographic buckets in the data. To use this variation as a source of approximating EdgeRank, we approximate the RHS of Equation (2.1) as,

\[
E(n^{(d)}_{kjt}) \approx g_d(N^{(d)}_{jt}, \theta^{(d)}_{1j}, z_k, \tau_k) \quad (2.2)
\]

where, we use a firm-demographic bin specific fixed effect, \( \theta^{(d)}_{1j} \), to capture the effect of user history. This approximation would literally be true if all individuals within demographic bucket \( d \) had the same history with firm \( j \). In practice, this is not the case, and this may induce approximation errors into the procedure, because additional history-heterogeneity within demographic buckets is not modeled (or is assumed into the error term). This is a caveat to our analysis. Access to individual-level data could be the basis of improving this procedure and relaxing this assumption. We view Equation (2.2) as a flexible approximation that allows us to leverage the observed variation in firm-level impressions across demographics, while enabling us to include firm and demographic-level fixed effects into a procedure that best approximates EdgeRank based on what we as researchers (and firms) know about Facebook’s filtering algorithm. We will also estimate the right-hand function \( g_d(.) \) separately for each demographic bucket, in effect allowing for slope heterogeneity in demographics in addition to intercept heterogeneity across demographics.
The next step relates to approximating the function \( g_d(.) \). Since we do not know the exact functional form of the above selection equation, we approximate the function semiparametrically via a Generalized Additive Model (GAM) (c.f., Hastie and Tibshirani (1990)). The GAM is a generalized linear model with additive predictors consisting of smoothed (e.g. interpolation and curve fitting) covariates. The GAM fits the following flexible relationship between a set of covariates \( X \) and dependent variable \( Y \),

\[
\mu(\mathbb{E}(Y|X_1, X_2, \ldots, X_p)) = \alpha + s_1(X_1) + s_2(X_2) + \cdots + s_p(X_p) \tag{2.3}
\]

where \( \mu \) is a link function (e.g. gaussian, poisson, gamma), and \( s_1, s_2, \ldots s_p \) are nonparametric smoothing functions such as cubic splines or kernel smoothers. We model the EdgeRank selection equation for each demographic \( d \) as the following,

\[
h_d \left[ \log(n_{kjt}^{(d)} + 1) \right] = \theta_0^{(d)} + \theta_1^{(d)} j + \theta_2^{(d)} N_{jt}^{(d)} + s_1(N_{jt}^{(d)}; \theta_3^{(d)}) + \sum_{r=2}^{5} \theta_4^{(d)} I(z_k = r) + \sum_{r=2}^{16} \theta_5^{(d)} I(\tau_k = r) + \epsilon_{kjt}^{(d)} \tag{2.4}
\]

where, \( h_d \equiv g_d^{-1}(.) \) is the identity (Gaussian) link function, \( \theta_0^{(d)} \) is an intercept term unique to each demographic, \( d \), and \( \theta_1^{(d)} j \) is a firm-demographic fixed effect that captures the tie strength between the firm \( j \) and demographics \( d \).\(^{13} \) \( N_{jt}^{(d)} \) is the number of fans of demographic \( d \) for firm \( j \) at time \( t \) and denotes the potential audience for a message. \( s_1 \) is a cubic spline smoothing function, essentially a piecewise-defined function consisting of many cubic polynomials joined together at regular intervals of the domain such that the fitted curve, the first and second derivatives are continuous. We represent the interpolating function \( s_1(.) \) as a linear combination of a set of basis functions \( b(.) \) and write: \( s_1(N_{jt}^{(d)}; \theta_3^{(d)}) = \sum_{r=3}^{q} b_r \left( N_{jt}^{(d)} \right) \theta_{3r}^{(d)} \), where the \( b_r(.) \) are a set of basis functions of dimension \( q \) to be chosen

\(^{13}\) We also tried Poisson and Negative Binomial link functions (since \( n_{kjt}^{(d)} \) is a count variable), as well as the identity link function without logging the \( y \)-variable. Across these specifications, we found the identity link function with log \( (y) \) resulted in the best fit, possibly due to many outliers. We also considered specifications with numerous interaction of the covariates included, but found they were either not significant or provided trivial gains in the \( R^2 \).
and $\theta^{(d)}_{3\alpha}$ are a set of parameters to be estimated. We follow a standard method of generating basis functions, $b_r(\cdot)$, for the cubic spline interpolation as defined in Wood (2006). Fitting the spline also requires choosing a smoothing parameter, which we tune via generalized cross-validation. We fit all models via the R package `mgcv` described in Wood (2006).

Finally, we include dummy variables for message-type ($z_k$) and for each day since release of the message ($\tau_k$; up to 16 days), to capture the effect of message-type and time-since-release semiparametrically. These are allowed to be $d$–specific. We collect the set of parameters to be estimated for each demographic bucket in a vector, $\hat{\theta}^{(d)}$, which we estimate by GAM estimation. The estimated parameter vector, denoted $\hat{\theta}^{(d)}$, $d = 1, \ldots, D$, serves as an input to the second stage of the estimation procedure.

### 2.3.2. Second-stage: Modeling Engagement Given Message Assignment

We operationalize engagement via two actions, Likes and comments on the message. The selection problem is that users can choose to Like or comment on a message only if they are served impressions, which generates non-random censoring because impression assignment is endogenous to the action. We address the censoring by including a correction for the fact that a user is shown a message non-randomly, estimated semiparametrically as above. Suppose $\hat{\Psi}^{(d)}_{kjt}$ denotes the fitted estimate from the first-stage of the expected number of impressions of message $k$ for firm $j$ amongst users of type $d$ at time $t$,

$$\hat{\Psi}^{(d)}_{kjt} = g_{d} \left( N^{(d)}_{jt}, z_{k}, \tau_{k}; \hat{\theta}^{(d)} \right)$$

We model the probability that users of type-$d$ will Like a message given the full set of message characteristics, $M_{kt}$, as logistic with parameters $\Omega = (\delta_{d}, \psi)_{d=1,\ldots,D}$,

$$\pi_{d}(M_{kt}; \Omega) = \frac{1}{1 + e^{-(\delta_{d} + M_{kt}\psi)}}$$
The parameter vector, \( \Omega \), is the object of inference in the second stage.\(^{14} \)

We will estimate \( \Omega \) by fitting the model to explain \( Q_{kjt} \), the observed number of Likes of the message in each period in the data. To see the intuition for how our correction works in the estimation, note that we can aggregate Equation (2.6) across users, so that the expected number of Likes is,

\[
\mathbb{E}(Q_{kjt}; \Omega) \approx \sum_{d=1}^{D} \hat{\Psi}_{kjt}^{(d)} \times \left[ \frac{1}{1 + e^{-(\delta_d + M_{kt} \hat{\psi})}} \right]
\]

(2.7)

with \( \hat{\Psi}_{kjt}^{(d)} \) are treated as known from the first-stage (Equation 2.5). The right-hand side is a weighted sum of logit probabilities of Liking a message. Intuitively, the decision to Like a message is observed by the researcher only for a subset of users who were endogenously assigned an impression by FB. The selection functions \( \hat{\Psi}_{kjt}^{(d)} \) serve as weights that reweigh the probability of Liking to account for the fact that those users were endogenously sampled, thereby correcting for the non-random nature of message assignment when estimating the outcome equation.

We could use the expectation in Equation (2.7) as the basis of an estimation equation. Instead, for efficiency, we estimate the parameter vector \( \Omega \) by maximum likelihood. To set up the likelihood, note the expected number of impressions of message \( k \) for firm \( j \) at time \( t \) across all demographic buckets is simply the sum,

\[
\hat{\Psi}_{kjt} = \sum_{d=1}^{D} g_d \left( \mathcal{N}_{jt}^{(d)}, \mathcal{Z}_k, \mathcal{T}_k; \hat{\theta}^{(d)} \right)
\]

(2.8)

We can obtain an estimate of the implied probability that an impression picked at random from the pool is of type-\( d \),

\[
\hat{\varrho}_{dkt} = \frac{\hat{\Psi}_{kjt}^{(d)}}{\hat{\Psi}_{kjt}}
\]

(2.9)

\(^{14}\)Allowing \( \psi \) to be \( d \)-specific as well in Equation (2.6) is conceptually straightforward. Unfortunately, this results in parameter proliferation and trouble with convergence; hence we settled for a more limited specification with \( d \)-specific intercepts.
Thus, the probability $\pi(M_{kt}; \Omega)$ that an impression picked at random from the pool will
Like the message given a guess of $\Omega$, is,

$$
\pi(M_{kt}; \Omega) = \sum_{d=1}^{D} p_{kt}(d) \times p_{kt}(\text{Like}|d) = \sum_{d=1}^{D} \hat{\varrho}_{dkt} \times \pi_{d}(M_{kt}; \Omega) \quad (2.10)
$$

Intuitively, with probability $p_{kt}(d) = \hat{\varrho}_{dkt}$ an impression is of type-$d$, and with probability
$p(Like|d) = \pi_{d}(M_{kt}; \Omega)$, an impression will Like the message conditional on being type-$d$; hence the unconditional probability a random impression will Like the message is the
sum-product of these marginals and conditionals across all $D$ types.

The number of Likes is a count variable for which we specify a Binomial likelihood. Accordingly, the probability that $Q_{kjt}$ out of the $\hat{\Psi}_{kjt}$ assigned impressions are observed to
Like the message, and that $\hat{\Psi}_{kjt} - Q_{kjt}$ of the remaining impressions are observed not to, is binomial with probability, $\pi(M_{kt}; \Omega)$,

$$
Q_{kjt} \sim \text{Binomial}(\hat{\Psi}_{kjt}, \pi(M_{kt}; \Omega)) \quad (2.11)
$$

Maximizing the implied binomial likelihood across all the data, treating $\hat{\Psi}_{kjt}$ as given, then
delivers estimates of $\Omega$. The intuition for the selection correction here is the same as that
encapsulated in Equation (2.7). We can repeat the same procedure using the number of
comments on the message as the dependent variable so as the recover the effect of message-
characteristics on commenting as well. This two-step procedure thus delivers estimates of
the effects of message-characteristics on the two outcomes of interest. Standard errors are
obtained by bootstrapping both steps 1 and 2 over the entire dataset.

**Discussion of Identification**  Identification in the model derives from two sources. First,
we exploit the observed discrepancy in demographic distributions between the set of indi-
viduals to whom a message could have been served, versus those who were actually served.
The discrepancy reflects the filtering by EdgeRank. Our first stage essentially projects this
discrepancy onto message-type, time-since-release, page and demographic characteristics in
a flexible way. This essentially serves as a “quasi” control function that corrects for the selectivity in the second stage (Blundell and Powell, 2003), where we measure the effect of message characteristics on outcomes. The second source of identification arises from exploiting the implied restriction that the rich set of AMT-content-coded attributes affect actual engagement, but are not directly used by EdgeRank to assign messages to users. This serves as an implicit exclusion that helps address selection. The only message-characteristic used by EdgeRank for assignment is \( z_k \), which is controlled for. Thus, any systematic correlation in outcomes with AMT-content-coded characteristics, holding \( z_k \) fixed, do not reflect selection-related considerations. One caveat is the control for selection does depend on assumptions we made about EdgeRank based on what is known publicly. We used a flexible first-stage specification so as to be as robust as possible to these assumptions. Notwithstanding these aspects, to the best of our knowledge, the full details of EdgeRank are not known to any firm or researcher. In our view, a “perfect” solution to the selection problem is unlikely to be achieved without full knowledge of Facebook’s targeting rule.

2.4. Results

2.4.1. First-Stage

The first-stage model, as specified in Equation 2.4, approximates EdgeRank’s message assignment algorithm. We run the model separately for each of the 14 age-gender bins used by Facebook. These correspond to two gender and seven age bins. For a given bin, the model relates the number of users of demographic type \( d \) who were shown message \( k \) by firm \( j \) at time \( t \) to the message type \( (z_k) \), days since message \( (\tau) \), and tie between the firm and the user. Table 5 presents the results. The intercepts \( (\theta_0^{(d)}) \) indicate that messages by companies in our dataset are shown most often to Females ages 35–44, Females 45–54, and Males 25–34. The lowest number of impressions are for the 65+ age group. In our model, tie between a user and a firm is proxied by a fixed-effect for each firm-demographic pair. This implies \( 800 \times 14 \) fixed effects corresponding to 800 firms and 14 demographic bins. Due to space constraints, we do not present all the estimated coefficients. Table 5 presents
the coefficients for two randomly chosen firms. The first is a new-born clothing brand and the second is a protein bar brand. For ease of visualization, these fixed effects are shown graphically in Figure 10 (only the statistically significant coefficients are plotted). For messages by the new-born clothing brand, the most impressions are among females in the age-groups of 25–34, 18–24, and 35–44. Among males, ages 25–34 receive the most number of impressions. For messages by the protein bar brand, impressions are more evenly distributed across the different demographic bins, with the Male 18–24 group receiving the most impressions. These estimated coefficients are consistent with our expectations for the two brands.

The estimates for message type are roughly the same in all demographic bins. For all demographics, the photo type has the highest coefficient (around 0.25) suggesting that photos are preferred to all other media types by EdgeRank. This is likely because users have historically engaged better with photos causing Facebook to show photos more often. The next most preferred message type is the status update with coefficients averaging around 0.12 followed by videos and links. The baseline message type, apps, is the message type that is least preferred by EdgeRank. The rank ordering of coefficients for message type do not strictly follow the rank ordering of number of messages released by firms, which is shown in Table 1. Whereas links are posted more often, photos get more impressions relative to messages of other types, clearly highlighting the role of EdgeRank. Days since message ($\tau$) are not presented in Table 5 due to space constraints. However, Figure 11 presents a box plot of the coefficients for $\tau$ across all 14 demographic bins. All coefficients are negative and significant and also more negative for higher values of $\tau$, implying that EdgeRank prefers to show more recent messages. Finally, the coefficients for number of fans, $n_{jt}^{(d)}$, are positive and significant but they have relatively low magnitude. This is because our model includes a smoothed term of the number of fans, $s(n_{jt}^{(d)})$, which soaks up both the magnitude and nonlinearity. The smoothed fan-numbers are all significant.

The generalized additive model of EdgeRank recovers coefficients that make intuitive sense
Table 5: EdgeRank Model Estimates: This table presents the coefficients obtained from 14 generalized additive models for EdgeRank, calculated for each demographic bin. There are 14 demographic (gender-age) bins provided by Facebook. F13–17 means all females in the age between 13 and 17. Time since message ($\tau$), and page-level fixed effects are not included in the table and presented graphically separately.

and are consistent with claims made in several industry reports (e.g. that photos have the highest EdgeRank weight). Further, the model fit appears to be good especially given that we have used generalized cross-validation to guard against overfitting.

### 2.4.2. Second-Stage

In the second-stage, we measure the effect of content characteristics on engagement using our selectivity-corrected model from the first-stage. All results in this section are based on
Figure 10: Page-level Fixed effect Estimates from Generalized Additive Model Across 14 Demographic Bins: This bar graph shows two randomly chosen page-level fixed effect estimates from the EdgeRank models. Only the statistically significant estimates are shown. New born clothing brands are positively significant for 18–24 female, 25–34 female, 35–44 female, and 25–34 male. Protein bar brands have the highest fixed effect among 18–24 male demographics.

Figure 11: Time Since message Release (τ) Coefficients Box plot Across Demographics: This box plot shows the coefficients on τ across all the demographics bin. τ = 1 is the base case and every coefficients are significant at the highest level of p < 0.001.

an analysis of the entire set of over 100,000 messages (i.e. the 5,000 AMT-tagged messages as well as the messages tagged using NLP). To present the results in a simple way, we first cre-
ate two composite summary variables corresponding to persuasive content and informative content. Persuasive (informative) composite variables are created by adding up the content variables categorized as persuasive (informative) in Table 1. To be clear, the persuasive variable is obtained by adding values of REMFACT, EMOTION, EMOTICON, HOLIDAY-MENTION, HUMOR, PHILANTHROPIC, FRIENDLIKELY, and SMALLTALK resulting in a composite variable ranging from 0 to 8. The informative composite variable is obtained by adding values of BRANDMENTION, DEAL, PRICECOMPARE, PRICE, TARGET, PRODAVAIL, PRODLOCATION, and PRODMENTION resulting in a composite variable ranging from 0 to 8. Table 6 shows the result of logistic regression on engagement with these composite variables and interaction of those two variables as the x-s.

We find that inclusion of more persuasive content has a positive and statistically significant effect on both types of engagement; further, inclusion of more informative content reduces engagement. Interestingly, the interaction between persuasive and informative content is positive, implying that informative content increases engagement in the presence of persuasive content in the message. This results suggests broad guidelines for marketers: persuasive content in isolation is preferred to purely informative ones. Further, mixing persuasive and informative content should be made a basis of content engineering for improving engagement with consumers on this medium.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comment</th>
<th>Like</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>−6.913†</td>
<td>−4.671†</td>
</tr>
<tr>
<td>Persuasive</td>
<td>0.053†</td>
<td>0.061†</td>
</tr>
<tr>
<td>Informative</td>
<td>−0.143†</td>
<td>−0.068†</td>
</tr>
<tr>
<td>Persuasive × Informative</td>
<td>0.012†</td>
<td>0.003†</td>
</tr>
<tr>
<td>McFadden R-sq.</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>Nagelkerke R-sq.</td>
<td>0.015</td>
<td>0.009</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>−4208220.431</td>
<td>−33678695.014</td>
</tr>
<tr>
<td>Deviance</td>
<td>8012471.987</td>
<td>66409947.187</td>
</tr>
<tr>
<td>AIC</td>
<td>8416448.861</td>
<td>67357398.028</td>
</tr>
<tr>
<td>N</td>
<td>665916</td>
<td>665916</td>
</tr>
</tbody>
</table>

Significance † 0.001 †† 0.01 †‡ 0.05 †† 0.1

Table 6: Persuasive vs Informative: Logistic regression for \{Comment, Like\} with composite summary variables for persuasive and informative content.
Table 7 presents the results of aggregate logistic regression with the full list of content variables. We present results for both engagement metrics (Likes/comments) as well as for models with and without the EdgeRank correction. We exclude the 16 estimated $\tau$ coefficients from the table since they are all negative and statistically significant just as in the EdgeRank model in Figure 11. We also exclude demographic fixed effects for space. Scanning through the results, we observe that the estimates are directionally similar, in most cases, with and without EdgeRank correction. However, the magnitudes often change. For example, consider the coefficients for message type Photo. In the model without EdgeRank correction, Photos are very likely to get comments (coefficient = 0.867) and Likes (coefficient = 1.011). After EdgeRank correction, the results are similar but the magnitude of the effect drops. This makes sense because we know that EdgeRank prefers Photos. Similarly, Status Updates continue to be more likely (than apps) to get comments and Likes but the effect size is smaller after EdgeRank correction. In some instances, there are directional changes for some coefficients. For example, the result that links are more likely to get Likes/comments relative to apps changes sign after EdgeRank correction. This highlights the importance of EdgeRank correction, an issue that most industry reports (e.g., Wildfire 2012) often overlook. For example, most industry reports’ ordering of engaging media type often list status update to be more engaging than videos. While we find this to be true before EdgeRank correction for Likes, we find that this is reversed after the EdgeRank correction.

We find that high reading complexity (SMOG) decreases both Likes and comments whereas shorter messages (MSGLEN) are Liked and commented on more, albeit with a small effect size. Having links (HTTP) is worse for engagement whereas asking questions (QUESTION) significantly increase comments but at the cost of Likes. Using blanks in the message to encourage comments has a similar effect of increasing comments but hurting Likes. Interestingly, while the odds ratio of comments increases by 69% if a message asks a question, it increases by 200% if blanks are included suggesting that blanks are more effective than questions if the goal is to increase comments. Asking for Likes increase both Likes and comments, whereas asking for comments increase comments but at the cost of Likes. It is
clear that even these simple content variables impact user engagement.

The next 16 variables in the table are the persuasive and informative content variables. Figure 12 charts the coefficients for these variables in a bar graph and demonstrates the sharp difference between persuasive and informative content types. Looking at comments, a striking pattern is that most informative contents have a negative impact whereas persuasive contents have a positive impact. The informative content variables with the most negative impact are PRICE, DEAL, and PRODMENTION. The persuasive content variables with the most positive impact are EMOTION and PHILANTHROPIC. Interestingly, HOLIDAYMENTION discourages comments. One possible explanation is that near holidays, all Facebook pages indiscriminately mention holidays, leading to dulled responses. For example, during Easter, the occurrence of holiday mention jumped to nearly 40% across all messages released that day compared to the average occurrence of about 1%. Looking at Likes, fewer persuasive content variables have positive impact but the results are qualitatively similar to that for comments. Among persuasive contents, EMOTION has the most positive impact on Likes whereas EMOTICON has the most negative impact. Most informative content variables continue to have a negative impact (i.e., reduce engagement), with PRICE and DEAL having the most negative impact. The results also highlight that there exist some differences between impact on Likes versus Comments.

Figure 13 shows the results on content effects by industry. Only the statistically significant results are graphed and all results are EdgeRank-corrected. The coefficients are very different across industries both in magnitude and, for some variables, in direction. For example, emotional and philanthropic content has the most positive impact on Facebook pages of type “Organizations” which include non-profits, educational organizations and religious groups. Further, while mentioning holidays has a negative impact on engagement for most industry types, it has a positive impact on engagement for Organizations. Similarly, looking at informative content, we observe that variables such as Price, Product Availability,

\[15\] We checked for correlation with other contents to investigate this matter but no correlation was over 0.02.
and Product Mentions generally have a negative impact on engagement for most industry
types, but have a positive impact for industry type “Celebrity.” Users seem more forgiving
of celebrity pages endorsing products and sharing price information.

Comparing Figures 3 and 13 also provides interesting comparisons of what each industry
is currently posting and what users engage with. For example, pages of types Places and
Businesses, Entertainment, and Consumer Products do not post emotional content much
though Figure 13 shows that emotional content induce higher Likes and Comments. Sim-
ilarly, while Places and Business pages tend to post more of deal content, only Consumer
Product pages seem to be benefiting from the deal content (in terms of obtaining more
comments). Places and Businesses pages also post larger percent of product availability
content while only the Consumer Product and Celebrity pages benefit from inclusion of
such content.

2.4.3. Robustness

Targeting by Firms

We concentrated on EdgeRank-induced selection as the main difficulty in inference since we
believe the specifics of the Facebook environment makes several other sources of confounds
second-order compared to the effect of EdgeRank. For instance, one concern may be that
firms may target content directly to specific users on Facebook. This is not possible because
in contrast to Facebook’s banner advertisements or sponsored posts, the Facebook organic
page environment does not allow companies to target specific audiences (the only factor
that can be controlled is the time-of-day of release of the message). Rather, all targeting
is implicitly implemented by Facebook via EdgeRank’s filtering. Another story may be
that firms observe that a particular type of content receives significant engagement, and
subsequently start posting similar content. Thus, new content reflects past engagement.
Our data shows significant within-variation in the attributes of messages launched over time
by a given firm, which is inconsistent with this story which would instead predict significant
<table>
<thead>
<tr>
<th></th>
<th>NO ER COMMENT</th>
<th>OR</th>
<th>ER COMMENT</th>
<th>OR</th>
<th>NO ER LIKE</th>
<th>OR</th>
<th>ER LIKE</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>12.309 (0.197)</td>
<td>14.083 (0.142)</td>
<td>-7.833 (0.089)</td>
<td>13.504 (0.065)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMOG</td>
<td>-0.045 (0.000)</td>
<td>0.956 (0.000)</td>
<td>-0.066 (0.000)</td>
<td>0.936 (0.000)</td>
<td>-0.029 (0.000)</td>
<td>0.971 (0.000)</td>
<td>-0.057 (0.000)</td>
<td>0.945</td>
</tr>
<tr>
<td>MSGLEN</td>
<td>0.000 (0.000)</td>
<td>1.000 (0.000)</td>
<td>-0.000 (0.000)</td>
<td>1.000 (0.000)</td>
<td>-0.000 (0.000)</td>
<td>1.000 (0.000)</td>
<td>-0.000 (0.000)</td>
<td>1.000</td>
</tr>
<tr>
<td>HTTP</td>
<td>-0.484 (0.002)</td>
<td>0.616 (0.002)</td>
<td>-0.324 (0.002)</td>
<td>0.723 (0.000)</td>
<td>-0.353 (0.000)</td>
<td>0.703 (0.000)</td>
<td>-0.180 (0.000)</td>
<td>0.835</td>
</tr>
<tr>
<td>QUESTION</td>
<td>0.449 (0.001)</td>
<td>1.567 (0.001)</td>
<td>0.527 (0.001)</td>
<td>1.694 (0.000)</td>
<td>-0.292 (0.000)</td>
<td>0.747 (0.000)</td>
<td>-0.185 (0.000)</td>
<td>0.831</td>
</tr>
<tr>
<td>BLANK</td>
<td>0.942 (0.003)</td>
<td>2.565 (0.009)</td>
<td>1.099 (0.003)</td>
<td>3.001 (0.002)</td>
<td>-0.716 (0.002)</td>
<td>0.489 (0.002)</td>
<td>-0.625 (0.002)</td>
<td>0.535</td>
</tr>
<tr>
<td>ASKLIKE</td>
<td>0.002 (0.010)</td>
<td>1.002 (0.010)</td>
<td>0.178 (0.010)</td>
<td>1.195 (0.003)</td>
<td>0.456 (0.003)</td>
<td>1.578 (0.003)</td>
<td>0.501 (0.003)</td>
<td>1.650</td>
</tr>
<tr>
<td>ASKCOMMENT</td>
<td>0.779 (0.021)</td>
<td>2.179 (0.021)</td>
<td>0.710 (0.021)</td>
<td>2.034 (0.011)</td>
<td>-0.000 (0.011)</td>
<td>0.914 (0.011)</td>
<td>-0.282 (0.011)</td>
<td>0.754</td>
</tr>
</tbody>
</table>

**Table 7: Aggregate Logistic Regression Results For Comments and Likes**

This table presents the aggregate logistic regression on comments and Likes for both EdgeRank-corrected (ER) and uncorrected (NO ER) for all data. OR means Odds Ratio and shows the odds ratio for the estimates left of the column.
Figure 12: Message Characteristic Coefficients for Comments and Likes: These bar graphs show the coefficients of logistic regression for both EdgeRank corrected and uncorrected models. Only the significant coefficients are plotted.

within-firm persistence in these attributes. Another concern is that engagement is driven by unmeasured message characteristics that co-occur with included message characteristics. To the extent that these unmeasured message characteristics drive engagement, they represent unobservables that are potentially correlated with included message characteristics and generate an omitted variables problem. This concern is plausible, but is second order in
Figure 13: Logistic Regression by Industry (Comments and Likes) This bar graphs show the coefficients of logistic regression for EdgeRank-corrected model. Only the significant coefficients ($p < 0.05$) are graphed. In the Like plot on the right, the coefficient for ASKCOMMENT for websites is $-4.8$ but zoomed in to optimize the clarity of the graph.

our view to the extent that we have included a very rich set of message characteristics. Our approach to this problem has been to convert unobservables into observables by collecting direct data on a relatively comprehensive set of message-characteristics.

To validate that our results do not primarily reflect selection of certain kinds of posts by high-performing firms, we evaluate the diversity of content posted by firms as well as serial correlation in posts by firm. To the extent that the diversity of posts by firms is limited, it raises the concern that certain kinds of content attributes were not used by some low or high performing firms. Similarly, if there is high serial correlation in posts by firms, it may reflect limited content diversity or that firms are strategically choosing posts based on the performance of a previous post. To evaluate this issue, each post by a company is represented by a binary vector of length sixteen (8 informative and 8 persuasive) in which 1 represents that the content attribute is present (and 0 otherwise). Next, all such vectors
by the same firm are simply added up to form a vector indicating overall content creation for each firm. The Herfindahl index\(^\text{16}\) is then calculated for each firm. The mean Herfindahl index is 0.089, and the median 0.088, which is just above the minimal possible value of \(\frac{1}{16} = 0.0625\). Other concentration measures, such as Gini coefficient, also report similar patterns, which suggest that firms do mix content. Similarly, to show the extent of serial correlation in post content released by firms, we again represent each post with a binary vector of length sixteen. For each firm, messages are ordered by release date, then the XOR function applied to all consecutive messages to measure content similarity. A value of 1 means that all content attributes that were used in a previous post are not present in the current post, and vice versa. The average value of this XOR function across firms was 0.32 and the median was 0.33. This suggests significant variation in content by firms in our dataset.

Lastly, firms may be strategic about the timing of the messages. For example, say a firm wants to give an emotional message better exposure than a humorous message. If the firm knows that more people are logged on Facebook at 3pm versus 8am, the firm can post the emotional message at 3pm and the humorous one at 8am. While our model does control for demographics and impression, if firms’ strategic targeting of specific content varies heavily based on time of the day, this can result in vast differences in engagement across the messages. We evaluate this possibility for all the messages in our dataset and investigate if firms are targeting certain time of the day to post specific content. Our data show no evidence to support this strategic timing behavior. Figure 14 presents the distribution of the time of post for each of the sixteen content attributes. Each line represents a different content. This graph shows that while firms are deciding when to post (such as at 5pm–6pm when people leave their workplace), they are not posting specific content at specific time. All distributions appear similar. In fact, none of the \(\binom{16}{2}\) pair-wise Kolmogorov-Smirnov tests were able to reject the null that the distributions came from the same distribution.

\(^{16}\)A measure of diversity that ranges from \(\frac{1}{16}\) to 1 where 1 means highly concentrated. For our case it can range from \(\frac{1}{16}\) to 1 where \(\frac{1}{16}\) means all sixteen contents are equally used.
The figure, coupled with high diversity of content, suggests that firms in our dataset were not strategically selecting certain content attributes. This may be due to the lack of social media analytics tools that provide content-level analytics. In fact, social media analytics tools available at the data collection only provided simple timing strategies such as what time of the day to post, as reflected in the data and controlled for in our model, but not when to post specific content.

While it is hard to formally rule out a selection bias, all of our analysis of content diversity and timing suggests that firms’ strategic targeting is a second order consideration for our study.

![Proportion of Content Posted Split into Hour-bin: Each line represents one of sixteen content.](image)

**Figure 14: Proportion of Content Posted Split into Hour-bin: Each line represents one of sixteen content.**

**Alternative Specifications**

We run a variety of alternative specifications to assess the robustness of our results. First, we replicate the results using only the set of 5,000 messages directly coded up by the Amazon Mechanical Turkers. Second, we assess the extent to which the parameters are stable when we drop subsets of attributes. Third, we include additional checks for robustness against selection. Our added checks use residuals from the first-stage as a “control function” in the second-stage. To see this, note the residuals in Equation 2.4, $\epsilon_{kjt}^{(d)}$, represent unobserved
reasons that users in demographic bucket $d$ would be more likely to be targeted a message $k$ by EdgeRank. As robustness, we ask whether our results on the effect of message attributes change when we control for these unobservable drivers of attractiveness of each bucket for that message. To do this, note that from our first-stage, we can obtain an estimate of the residual, denoted $\hat{\epsilon}_{kjt}^{(d)}$. We re-run our second stage estimation including the estimated $\hat{\epsilon}_{kjt}^{(d)}$s as covariates in $M_{kt}$ in Equation 2.6. We can interpret the revised results as the effect of message characteristics on engagement after “controlling for” the unobserved attractiveness of each bucket for that message. Results from these alternative models show that the main qualitative features of our results are robust across these specifications.

2.5. Discussion and Managerial Implications

2.5.1. Shares, Click-throughs, and Explanation of Results

As previously mentioned, most marketers' top goal on social media sites is to increase consumer engagement, with the percentage of marketers who say so varying from 60% to more than 90% across different surveys (Ascend2, 2013; Gerber, 2014; eMarketer, 2013a; SmartBrief, 2010; Ragan and Solutions, 2012). As a result, we focused our attention on post-level engagements — *Likes* and comments. However, an acknowledged limitation of our study is that we do not have the sales data that matches the scale of 782 firms. Many academic research (Kumar et al., 2013; Goh et al., 2013; Rishika et al., 2013; Li and Wu, 2013; Miller and Tucker, 2013; Sunghun et al., 2014; Luo and Zhang, 2013; Luo et al., 2013) and industry reports (comScore, 2013; Chadwick-Martin-Bailey, 2010; 90octane, 2012; HubSpot, 2013) present positive correlation and even claim causal relationships between increased social media engagement and firm performance (e.g., purchase intention, visit frequency, profitability, etc). This suggests that our results may extend to these other performance measures as well. While most firms in our dataset did not track purchase information at the post level, we additionally obtained cross-sectional data on shares and click-throughs for the posts in our dataset to further investigate the effect of content on
more direct outcome measures like click-throughs.\textsuperscript{17} Figure 15 show the results for shares and click-throughs.

The results for shares are very similar to the results for comments and Likes. The general trend of persuasive contents with positive coefficients and informative contents with negative coefficients are replicated. Again, the emotional and humorous content were associated with the highest level of shares, while the price and deal information were associated with the lowest level of shares.

Results for click-throughs follow a similar trend to other engagements except for one notable difference. The coefficients for deal information and holiday mention (which is positively correlated with the presence of deal information) are now highly positive. While deal information may not elicit likes, comments, and shares, we find evidence that they do increase click-through rates. Results for other content attributes are qualitatively similar to those for Likes and comments. The results highlight that different content has different engagement outcomes, and that managers should implement appropriate content strategy for their marketing goals.

2.5.2. Managerial Implications

Finally, we informally assess the extent to which the models we develop may be used as an aid to content engineering, and to predict the expected levels of engagement for various content profiles a firm may consider for a potential message it could serve to users. We present an illustration set of out-of-sample prediction of engagement with real messages. Our intent is to give the reader a rough sense for the use of our estimates as a tool to assess expected engagement for hypothetical content bundles.

Suppose a firm’s content marketing team has developed multiple alternative messages. The marketing team may be interested in choosing a message keeping in mind a specific engagement goal. To illustrate the above, we choose three messages released around the same time.

\textsuperscript{17}While click-throughs do not guarantee sales, many links point to the companies’ product pages, content, and registration pages, all of which are a type of ‘conversion’.
outside our sample. To emphasize that our second-stage model of engagement has predictive power, we choose these to be for the same firm, of the same message type and having roughly the same number of impressions (i.e., we are taking out the effect of EdgeRank).

Table 8 shows the messages, the actual lifetime engagement realized for those messages, and the content tags we generated for those messages. For each message, we use the coefficients from our second stage to predict the expected engagement. In the “Content Coef:” column, we present the latent linear index of the logistic function obtained by multiplying the coef-
ficients for the engagement model (Table 7, EdgeRank corrected) with indicator variables for whether each type of content attribute is present in these messages, and then adding these up. In the last two columns we present the predicted and actual ranks for the three messages in terms of their engagement. We see that the match is very good and that the model can be useful to select the message that maximizes engagement.

Now imagine that a firm starts with the second message as the base creative. Content engineering for this message using the model is straightforward. For instance, if the marketer is assessing the potential impact of adding a philanthropic aspect and asking user to like the post to message two, we can determine that it will increase the latent linear index for comments from 0.731 to $0.731 + 0.140 + 0.178 = 1.049$ (from the Table 7, Like, EdgeRank corrected), which increases the predicted comments rank of this message to 1, and increases the predicted odds ratio for comments by 37%. Similarly, if the marketer is considering asking for comments explicitly, this will increase the number of comments for the message obtained by increasing the latent linear index from 0.731 to $0.731 + 0.710 = 1.441$. In this sense, the model is able to aid the assessment of the anticipated engagement from various possible content bundles in a straightforward way.

<table>
<thead>
<tr>
<th>Sample Messages</th>
<th>Actual Comments, Actual Likes</th>
<th>Content Tags</th>
<th>Content Coef {Com, Likes}</th>
<th>Comments Rank</th>
<th>Likes Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Don’t forget in celebration of hitting over 70,000 we are giving all our awesome fans {exclusively} the ‘employee discount’ take 20% off your entire order on our website {<a href="http://anonymized.com%7D">http://anonymized.com}</a> with the code: SOMECODE and it is good until 3/16/12. Enjoy some shopping on us :)</td>
<td>HTTP, DEAL, PROLOCATION, PRODAVAIL, EMOTICON</td>
<td>{-0.596, -0.552}</td>
<td>Actual:3</td>
<td>Actual:2</td>
<td></td>
</tr>
<tr>
<td>Predicted:3</td>
<td>Predicted:2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Who is ready for a giveawayyyyy?! :) :) (35 mins from now!)</td>
<td>EMOTION, EMOTICON, QUESTION</td>
<td>{0.731, -0.142}</td>
<td>Actual:2</td>
<td>Actual:1</td>
<td></td>
</tr>
<tr>
<td>Predicted:2</td>
<td>Predicted:1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPLETE THIS SENTENCE: Crafting is best with</td>
<td>BLANK, SMALLTALK</td>
<td>{1.025, -0.746}</td>
<td>Actual:1</td>
<td>Actual:3</td>
<td></td>
</tr>
<tr>
<td>Predicted:1</td>
<td>Predicted:3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Predicted versus Actual Engagement Ranking for Three Illustrative messages: Note: we anonymized some parts of the messages for presentation.
2.6. Conclusions

We show through a large-scale study that content engineering in social media has a significant impact on user engagement as measured by Likes, comments, shares, and click-throughs for messages. Our analysis shows that persuasive content, such as emotional and philanthropic content, has a positive impact on engagement. This suggests that firms gain from sharing their brand personality and information about their social initiatives on social media. Further, we find that product informative content has a negative impact on user engagement.\(^{18}\) This presents a challenge to marketers who seek to build a large following on social media and who seek to leverage that following to disseminate information about new products and promotions. One takeaway from our study is that these strategies work when product informative content is combined with persuasive content. In addition, our results are moderated by the industry type suggesting that there is no one-size-fits-all content strategy and that firms need to test multiple content strategies.

Because of the scale of our study (nearly 800 firms and more than 100,000 messages analyzed), we believe our results generalize and have broad applicability. Nonetheless, it is important to recognize several limitations of our study. First, we note that the results from any study on consumer response to content depend on the mix of content used in the study. For example, we find that messages mentioning holidays, especially by consumer product companies, have a negative effect on engagement. This may be due to excessive use of holiday messages by firms. It is possible that the effect may be positive if firms use these kinds of messages in moderation. Similarly, we find that emotional messages have a positive impact on engagement. Here again, it is possible this effect may reduce in the future if firms start using emotional content excessively. Hence, it is important to interpret our results in the context of the content mix used by firms and redo the analysis in the event of large-scale changes in the content mix used by firms. Ultimately, we urge managers to strike the right balance between the informative content (meant to drive leads and sales)

\(^{18}\)With a caveat for click-throughs where a deal information was the best content.
and the persuasive content (meant to engage the consumers), especially since EdgeRank uses firms’ current engagement level to determine future reach.

We used several metrics for user engagement, namely Likes and comments on messages as well as whether users share messages with friends or visit the link in the message. Our use of Likes, comments, shares, and click-throughs is motivated both by the widespread use of these metrics as marketing goal in social media settings, and also the availability of data. Future studies that evaluate other measures of interest can add value, particularly in validating the generalizability of our findings and in exploring mechanisms underpinning the effects we describe. As noted in the introduction, we do not address the question of how engagement affects product demand and firm’s profits so as to complete the link between ad-attributes and those outcome measures. Such data are still not widely available at the scale needed for this study. Although it is not the focus of our study, it is worth highlighting that several extant studies have studied the link between Facebook engagements and sales (albeit at a smaller scale). For example, based on randomized studies, comScore (2012) reports a 38% lift in purchase for fans exposed to Starbucks advertising on Facebook through Facebook Pages or Facebook paid advertising. Similarly, studies such as Kumar et al. (2013); Goh et al. (2013); Rishika et al. (2013); Li and Wu (2013); Miller and Tucker (2013); Sunghun et al. (2014); Luo and Zhang (2013); Luo et al. (2013) show that social media can be used to generate growth in sales, and ROI, consumer participation, retention, and profitability, connecting social media metrics such as “comments” to financial metrics.

The competition for consumer attention across media outlets is intense, especially on social media platforms. Consumers, in turn, are overwhelmed by the proliferation of online content, and it seems clear that marketers will not succeed without engineering this content for their audience. We hope this study contributes to improve content engineering by firms on social media sites and, more generally, creates interest in evaluating the effects of content on consumer engagement.
3.1. Introduction

Online consumers are constantly guided by some form of recommendation systems (RS)\(^1\), be it for shopping or web browsing. For shopping, recommenders can be as simple as “the most popular” items sold on the site, or as sophisticated as a collaborative filter, CF, based on individual purchase history (e.g., Amazon.com’s “Customers who bought this item also bought”). These systems are so pervasive in e-commerce and web services that a majority of consumers now expect and prefer websites that provide personalized recommendations (Accenture, 2012). At the same time, 94% of companies agree that “personalization of the web experience is critical to current and future success” (Econsultancy and Monetate, 2013). The benefits of RS for consumers include lower search costs and higher product-fit. In addition, with the rise of the long tail phenomenon, where in e-tailers offer a broad range of niche items (Anderson, 2008), recommenders help consumers manage a potential choice overload by reducing the consideration set (via showing the most relevant items). For firms, RS has been shown to promote sales, web usage, and customer retention (Das et al., 2007; De et al., 2010; Thompson, 2008; Monetate, 2013).

In addition to their positive impact on volume, recommenders are widely believed to affect sales diversity. This refers to the market share distribution of products sold at the firm level and the variety of items purchased at the individual consumer level. One school of thought believes that recommenders lower search costs and contribute to a long tail phenomenon in which consumers are exposed to more niche items, increasing both individual product consumption diversity and firm-level sales diversity. An opposing theory suggests that common recommender designs such as collaborative filters can lead to reduction in aggregate sales diversity. It argues that because collaborative filters recommend products based on

\(^1\)We use the terms recommender systems, RS, and recommender algorithms interchangeably.
sales and ratings, they cannot recommend items with limited historical data. This leads to popularity bias in these designs. Understanding which viewpoint is correct is important for retailers, consumers, and producers. For retailers whose strategy is to offer variety — under the premise that consumers will find better-suited products and thus purchase more — to the extent recommenders increase concentration, they may be at odds with such a goal. Similarly for consumers and niche producers, if there exist better product matches outside of the most popular titles, these groups may be better served (or underserved) if these recommender systems increase sales diversity (concentration).

There are several reasons for the divergent viewpoints on the impact of RS on sales diversity. First, a potential driver may be the lack of data that provides a contrast between users exposed and unexposed to recommendations. Most retailers observe consumers after they arrive at the website (i.e., after exposure to recommendations) and cannot observe the contrast needed to answer this question. This is one reason why the question remains empirically unanswered despite years of debate. Second, the majority of research to date has studied specific recommender algorithms in isolation rather than comparing different kinds of recommenders. In e-commerce, there are many types of recommenders (Schafer et al., 1999). Most commonly used designs tend to be collaborative filtering algorithms. There can be multiple flavors of collaborative filters, including those based on purchases (“Customers who purchased this also purchased”) and views (“Customers who viewed this also viewed”). Other types of recommenders include Content-based (“Based on your consumption history”) and Social-network-based recommenders (“Your friends bought”). One possible explanation for the divergent viewpoints is that these different recommendations have considerably different impact on sales volume and diversity. To the extent that this is true, it alters how firms must choose recommender designs. For example, one firm may prefer a design that maximizes sales, whereas another firm may prefer a design that better exposes consumers to its breadth of product assortment.

Our study attempts to address these gaps by utilizing a field experiment to examine the
impact of the two most commonly used recommender algorithms on sales volume and diversity. By running different algorithms on the website of a top retailer in North America and randomly assigning recommender treatments to visiting consumers and tracking their views and purchases, we tease out the differential effects of recommender designs on both sales volume and diversity. Our findings help answer several critical questions related to recommender design. For example, is one recommender algorithm more effective for increasing the number of products consumers view or purchase? Do certain algorithms cause “popularity bias,” wherein a top few bestsellers are promoted? How exactly do sales diversity shift occur? What algorithms should a manager use on an e-commerce site, given product assortment decisions and sales goals?

We find that recommender design affects both sales volume and diversity. In particular, a collaborative filtering algorithm based on purchase data, “Customers who purchased this item also purchased,” was best at increasing the sales volume. Further, the algorithm increased individual consumption diversity but decreased aggregate consumption diversity. We show that this differential impact in individual versus aggregate diversity is mainly caused by users exploring into only a few ‘pathway’ popular genres like comedy. The result highlights a potential limitation in the ability of traditional collaborative filtering to aid discovery of truly niche items and genres. We also show that not all collaborative filtering algorithms are equal by contrasting the results with that for View-based collaborative filtering.

Our results help reconcile opposing theories on the impact of recommenders using a randomized field experiment. Further, they unlock important managerial insights on which designs better address their goals. We find that collaborative filters, especially those based on purchases, are very effective at driving an increase in sales. However, to the extent that firms are interested in promoting a broader product assortment, we advocate that firms should modify traditional collaborative filtering algorithms to ensure that relevant items with limited historical sales can be discovered by consumers.
3.2. Prior Work

Given the significant influence of RS on e-commerce and consumer purchase behavior, the RS literature has been growing steadily in the last couple of decades. In the 1990s, which marked the rise of RS in e-commerce, researchers focused on developing different recommender algorithms (for an extensive survey, see Adomavicius and Tuzhilin (2005)). In the 2000s, a stream of research investigated the impact of RS on sales volume, or firm performance, looking at factors such as profit, revenue, and consumer retention (Das et al., 2007; De et al., 2010; Thompson, 2008). In the late 2000s and 2010s, studies of RS expanded to examining consumer consumption patterns and firm sales diversity (Fleder and Hosanagar, 2009; Hinz and Eckert, 2010; Oestreicher-Singer and Sundararajan, 2012; Jannach et al., 2013; Matt et al., 2013; Hosanagar et al., 2014). In this section, we first provide a taxonomy of recommender systems. Second, we review what the recommendation system literature tells us about recommenders’ influence on sales volume and revenue, and then review the burgeoning literature on recommenders’ influence on sales diversity. Lastly, we discuss the gap in the literature and position our paper in it.

3.2.1. Overview and Taxonomy of Recommendation Systems

Before the advent and growth of personalized recommenders, the primary recommendation approach used by most firms involved the use of simple signals such as the “most popular” or “highest rated” items. A well-known current example is The New York Times’s “most emailed articles” feature. Such signals have been shown to influence consumer learning and choice in consumer behavior literature and in studies tied to observational learning theory (Bikhchandani et al., 1992). Similar types of learning from the masses, whether based on positive views or reviews, have been well documented in the literature (Salganik et al., 2006; Muchnik et al., 2013; Tucker and Zhang, 2011; Lee et al.). While these kinds of recommendations based on aggregate signals are privacy preserving and serve the mass market, they may limit exploration by consumers and are less useful for consumers with niche tastes.
Personalized recommenders seek to recommend items based on individual-level data. Within *Personalized Recommenders* systems, a broad taxonomy distinguishes three types of algorithms: Content-based, Collaborative Filtering, and Hybrid, which combines the first two (Adomavicius and Tuzhilin, 2005). Content-based systems analyze product attributes to suggest products that are similar to those that a consumer bought or liked in the past. Collaborative filtering recommenders, unaware of product attributes, recommend products either purchased or liked by similar consumers, where similarity is measured by historical purchase (or like) data.

In recent years, social-network-based recommenders have emerged that can either simply signal to consumers that their friends have bought certain items or can be used as an extreme form of collaborative filtering that assumes that friends are similar in their tastes (Victor et al., 2011). From a marketing perspective, some recent papers have looked at how to incorporate consumers’ expressed preference and learning (Ansari et al., 2000) and sensitivity of purchase probability (Bodapati, 2008) into recommender systems. For an extended survey from a computer science perspective, please refer to Adomavicius and Tuzhilin (2005).

3.2.2. Literature on the Impact of Recommenders on Sales Volume and Diversity

While there are a large number of studies on improving the actual algorithms mentioned above, we know little about how these algorithms affect consumers and markets. While the literature is unequivocal that recommender systems positively impact sales, it is unclear about which designs have a greater impact. For example, industry reports have claimed considerable increase in revenue and/or usage due to the use of recommendation systems in the cases of Amazon, Netflix, and Google News (Thompson, 2008; Das et al., 2007; Marshall, 2006). Recommendation engines have been reported to increase revenue up to 300%, conversion rates up to 150%, and consumers’ average order value up to 50%, according to a recent study (Monetate, 2013). De et al. (2010) show that the use of a recommendation system has a positive effect on the sales of both promoted and non-promoted products in
contrast to the use of regular search engines, which increase sales only for promoted products and decrease sales for non-promoted products. Similarly, Hinz and Eckert (2010) show, via agent-based modeling, that the use of recommenders can increase profits for retailers. De et al. (2010) use a supermarket case study to show that RS increases both direct revenue and indirect revenue, in which indirect revenue is obtained by consumers’ cross-buying behaviors. Jamnach and Hegelich (2009) carry out a case study to evaluate the use of recommenders in a mobile app market and find that the use of RS increases click-through rates and overall sales. In summary, there is ample evidence to support the positive effect of RS on sales. However, the literature lacks field evidence comparing different types of recommender algorithms and which types of designs are more effective.

Unlike the sales volume impact of RS, the sales diversity impact of RS is a subject of much disagreement in the literature. On one side, Brynjolfsson et al. (2011) and Anderson (2008) indicate that the use of RS will contribute to a long tail phenomenon in which niche items gain more market share, increasing both individual- and firm-level product sales diversity. On the other side, using theoretical models and simulations, Fleder and Hosanagar (2009) argue that the use of collaborative filtering will lead to a decrease in firm-level aggregate sales diversity but it can lead to an increase in individual product consumption diversity. They suggest that collaborative filters can lead consumers to new items but aggregate diversity may still decrease because they push similar users to the same set of products. Similarly, Jamnach et al. (2013) use a well-known MovieLens dataset and simulation study to show that many recommendation algorithms are biased toward broad-appeal items, causing a rich-get-richer situation that decreases aggregate product sales diversity. Celma and Cano (2008) examine the collaborative filtering recommendation of last.fm and find that the algorithm tends to reinforce the consumption of popular artists, supporting the hypothesis that collaborative filtering will decrease aggregate diversity while also showing that content-based algorithms are less biased toward popular artists. In contrast, Hinz and Eckert (2010) argue through agent-based modeling that recommenders will drive a long tail phenomenon by reducing search costs and shifting demand from broad-appeal items to niche
items. Echoing this idea, Oestreicher-Singer and Sundararajan (2012) argue that Amazon’s collaborative filtering recommender (e.g., “Customers who bought this also bought”) shifts demand from broad-appeal items to niche items, thereby increasing aggregate product sales diversity. Similarly, regarding YouTube, Zhou et al. (2010) find evidence that the recommender increases aggregate diversity. Via a lab experiment, Matt et al. (2013) argue that both collaborative filtering and content-based recommenders increase aggregate sales diversity and that the differences between different recommenders are small. Similarly, by studying a hybrid algorithm that is content-base heavy, Hosanagar et al. (2014) focus on individual consumption patterns and show that, while recommenders increase individual consumption diversity, they also increase commonality among consumers. They also find that aggregate diversity increases as a result. Lastly, Wu et al. (2011) use simulation and MovieLens data and find that content-based recommenders tend to increase the aggregate sales diversity, while collaborative filtering decreases it. Table 9 summarizes these academic papers and their main claims. In sum, there is no consensus among both popular and academic literature on how recommenders will affect sales diversity.

We believe this lack of consensus arises due to the following several reasons. Some studies evaluate one type of algorithm and are based on lab experiments or simulations calibrated to archival data, which makes generalization harder. Other studies measure non-purchase measure like purchase-intentions, use-intentions, and satisfaction rather than the actual views or purchases. The few based on field archival data are aggregate level data and are constrained by the limitations of observational data that make causal conclusions harder to derive. We carry out a randomized field experiment on a large e-commerce website using multiple recommender algorithms and investigate the differential effects on sales volume and diversity with direct individual-level view and purchase data. While our approach is not entirely free from mentioned problems, the strength of our approach is that 1) the field experiment conducted on a large e-commerce site allows us to observe recommenders’ effects more realistically, 2) we directly measure individual-level view and purchase data, and 3) we have clean identification as a result.
<table>
<thead>
<tr>
<th>Study</th>
<th>Method &amp; Data</th>
<th>Sales Volume</th>
<th>Sales Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>De et al. (2010)</td>
<td>Archival Data &amp; Econometrics</td>
<td>Increases sales</td>
<td></td>
</tr>
<tr>
<td>Hinz and Eckert (2010)</td>
<td>MovieLens Data &amp; Simulation</td>
<td>Increases sales and profit</td>
<td>Increase niche product consumption leading to increase in aggregate sales diversity</td>
</tr>
<tr>
<td>De et al. (2010)</td>
<td>Archival Data &amp; Case Study</td>
<td>Increases direct and indirect revenue</td>
<td></td>
</tr>
<tr>
<td>Jannach and Hegelich (2009)</td>
<td>Mobile App Market Data &amp; Case Study</td>
<td>Increases sales</td>
<td></td>
</tr>
<tr>
<td>Fleder and Hosanagar (2009)</td>
<td>Theoretical Models &amp; Simulation</td>
<td></td>
<td>Decrease in aggregate sales diversity but increase in individual sales diversity</td>
</tr>
<tr>
<td>Hosanagar et al. (2014)</td>
<td>Archival Data &amp; Econometrics</td>
<td>Increases individual consumption volume</td>
<td>Content-based RS increase aggregate sales diversity and increase overlap/commonality in consumption</td>
</tr>
<tr>
<td>Oestreicher-Singer and Sundararajan (2012)</td>
<td>Crawled Amazon Data &amp; Econometrics</td>
<td>Increases revenue</td>
<td>Recommender shifts demand to niche item increasing aggregate sales diversity</td>
</tr>
<tr>
<td>Jannach et al. (2013)</td>
<td>MovieLens Data &amp; Simulation</td>
<td></td>
<td>Different algorithms have different effects</td>
</tr>
<tr>
<td>Wu et al. (2011)</td>
<td>MovieLens Data &amp; Simulation</td>
<td></td>
<td>Mixed result based on different algorithms. Collaborative filtering decreases aggregate diversity while content-based increases it</td>
</tr>
<tr>
<td>Celma and Cano (2008)</td>
<td>last.fm and Allmusic.com API data &amp; Correlational Analysis</td>
<td></td>
<td>Collaborative filtering algorithm is linked to popularity bias suggesting decreased aggregate consumption diversity</td>
</tr>
<tr>
<td>Zhou et al. (2010)</td>
<td>Crawled YouTube Data &amp; Correlational Analysis</td>
<td>Recommender accounts for 30% of video views</td>
<td>Increases aggregate consumption diversity</td>
</tr>
<tr>
<td>Matt et al. (2013)</td>
<td>Lab Experiments</td>
<td></td>
<td>Increase in aggregate sales diversity for variety of different recommenders except for bestseller list</td>
</tr>
</tbody>
</table>

Table 9: Literature on Impact of Recommender Systems and Claims
3.3. Problem Formulation

This section formally sets up research problems and hypotheses based on the previous literature.

3.3.1. Research Questions

We are interested in studying the impact of recommenders on sales volume and diversity. Sales volume is measured in terms of number of purchases and value of purchases ("wallet size"). In addition, we measure the sales diversity of the products sold with a measure called the Gini coefficient. The Gini coefficient has been widely adopted in the long tail and the RS literature as a measure of sales diversity (Brynjolfsson et al., 2011; Fleder and Hosanagar, 2009; Hosanagar et al., 2014). It is computed based on the Lorenz curve. Let \( L(u) \) be the Lorenz curve denoting the percentage of the sales generated by the lowest \( 100u\% \) of items as shown in Figure 16. The Gini coefficient is defined as \( G \equiv \frac{A}{A+B} \). It ranges from 0, representing the least amount of concentration or highest diversity, to 1, representing the highest amount of concentration or lowest diversity. A Gini coefficient of 0 means that all products have equal sales, while values near 1 mean that a few broad-appeal blockbuster items account for most of the sales.

We approach this problem with a field experiment in which consumers visiting a website are randomly assigned to a control or treatment group. The treatment group is shown a panel of different recommendations, much like Amazon’s “Customers who bought this item also bought” recommenders. The control group is shown nothing. For each group, we analyze the following variables of interest.

1. Sales Volume Impact

   (a) Individual item view volume: This counts how many items individuals have clicked and viewed.

   (b) Individual item purchase volume: This counts how many items individuals
buy. We also look at individual wallet size, which measures the total individual purchase dollar amount.

2. Sales Diversity Impact

(a) Aggregate firm-level item (and genre) **view and sales** diversity: This measures how the recommenders affect product view/sales diversity at the aggregate level (for each treatment group) and is measured by the Gini coefficient. We repeat this at the genre level to investigate genre cross-pollination.²

(b) Individual item (and genre) **view and sales** diversity: This measures how the recommenders affect the diversity of products individuals view or purchase. Again, the Gini coefficient is used but it is computed based on an individual’s purchases. The analysis is repeated at the genre level.

²Measuring genres consumed is a conservative and more robust way of measuring diversity, and it is included in the analysis because 1) it is easier to interpret the purchase diversity of genres and 2) mathematically, the Gini coefficient changes are more conservative at the genre level, making the results more robust. The results are the same at the item level, with more group comparisons statistically significant. We present both in the Results section.
3.3.2. Treatment Groups

There are three groups in our study:

1. Control (no recommendations)

2. View-based collaborative filtering ("People who viewed this item also viewed")

3. Purchase-based collaborative filtering ("People who purchased this item also purchased")

We have two different treatment groups corresponding to two different recommender algorithms, plus a control group that was not shown any recommendations. The two treatments are two of the most commonly used types of collaborative filtering algorithms. One is based on views ("People who viewed this item also viewed"), while the other is based on purchases ("People who purchased this item also purchased"). Our data partner used the widely adopted open-source Apache Mahout framework (mahout.apache.org) for constructing the recommenders.

3.3.3. Study Design

Let $g_i$ represent group $i$ and let $f$ represent a function that calculates an aggregate measure of interest, $D_i$, for the given group (e.g., sales volume, Gini coefficient). We define the following quantity of interest:

| Aggregate Measure, $f$, of Group 1 | $D_1 \equiv f(g_1)$ |
| Aggregate Measure, $f$, of Group 2 | $D_2 \equiv f(g_2)$ |
| Difference in Aggregate Measures   | $D \equiv D_1 - D_2$ |

The difference in the aggregate measure, $D$, shows how different Group 1 is from Group 2. Let $\mu \equiv \mathbb{E}[D]$, with the distribution of $D$ unknown. All hypotheses testing in this paper takes the form:
Null Hypothesis $H_0$ \[ \mu = E[D] = 0 \]
Alternate Hypothesis $H_a$ \[ \mu \neq E[D] \]

This null hypothesis tests if variables of interest, such as group revenues for Groups 1 and 2, are distributed equally. Since we have one aggregate measure (or statistic) for each group, in order to produce a p-value, we utilize a permutation test technique (Good, 2005) that allows us to calculate a null distribution for a given aggregate measure. If the null hypothesis of equal distribution is true, all relabelings of individuals as Groups 1 and 2 are equally likely. A permutation test involves repeatedly and randomly relabeling individuals into Groups 1 and 2 (e.g., control and treated) to produce a null distributions for any test statistics. By comparing statistics from null distributions to the actual test statistics from the real distribution and tallying how often null distribution statistics exceed the actual distribution statistic, we can determine the p-value. For more details, see Good (2005). Note that we carry out the hypotheses tests as two-sided tests (equal or not equal rather than greater than or less than) to stay conservative. In our study, we use 1000 iterations to get an accurate p-value up to 0.001.

3.3.4. Hypotheses

We organize our hypotheses on the impact of recommenders on sales volume and diversity in this section. Our hypotheses are informed by the extant literature discussed in Section 3.2. While the hypotheses on sales volume are clearly driven by unequivocal results of previous studies, we take a particular stance on sales diversity since the existing literature disagrees on this matter.

Sales Volume Hypotheses There appear to be many reasons why recommenders affect sales volume. Many long tail and recommender studies (Brynjolfsson et al., 2006, 2011; Oestreicher-Singer and Sundararajan, 2012; Hinz and Eckert, 2010) postulate that personalized recommenders reduce consumers’ search costs by reducing the need to search for the right product. In doing so, recommenders make it easy for consumers to find the right prod-
ucts, which leads to more purchases. The impact on number of product views is less clear. On the one hand, a reduction in search cost can lead consumers to their desired product faster, resulting in fewer clicks and product views. On the other hand, recommenders may be effective in cross-selling, up-selling, or even in driving repeat visits, thereby resulting in increase in number of product views. We hypothesize that recommenders will contribute toward an increase in purchases as well as product views.

**Hypothesis 1**  
Collaborative filtering (CF) recommenders will increase the number of **product views** compared to no recommendation condition (Control).

**Hypothesis 2**  
Collaborative filtering (CF) recommenders will increase the number of **product purchases** compared to no recommendation condition (Control).

Although we have two different types of collaborative filters (Purchase-based or View-based), we do not make any hypotheses comparing the two. Intuition suggests that collaborative filtering based on purchase data will be more potent, since a purchase is a stronger signal than a view and is probably more accurate.

**Sales Diversity Hypotheses**  
Fleder and Hosanagar (2009) offer a clear conjecture on how collaborative filters affect product consumption diversity at both the aggregate and individual level. They argue that a collaborative filtering algorithm will show popularity bias by directing users to broad-appeal blockbuster items, leading to decreased aggregate firm-level product sales diversity. Likewise, Celma and Cano (2008) and Wu et al. (2011) also support the hypothesis that collaborative filtering will decrease aggregate product sales diversity. At the same time, Fleder and Hosanagar (2009) predict that individuals will be exposed to a greater variety of products, leading to increased individual product consumption diversity. Our hypotheses on sales diversity are based on the Fleder and Hosanagar (2009) paper, which presents a theoretical model specifically on collaborative filtering and provides conjecture on both the aggregate and individual diversity.

In summary, our hypotheses on the **sales diversity impact** of RS are as follows:

**Hypothesis 3**  
Collaborative filtering recommenders will decrease the aggregate product
sales diversity compared to no recommendation condition (control), i.e., $\text{Aggregate Gini(CF)} > \text{Aggregate Gini(Control)}$.

**Hypothesis 4** Collaborative filtering recommenders will increase the individual product purchase diversity compared to no recommendation condition (control), i.e., \( \text{Average Individual Gini(CF)} < \text{Average Individual Gini (Control)} \).

3.4. Data

Our dataset comes from a field experiment on a website of one of the top retailers in North America. The experiment was conducted for two weeks between August 8, 2013 and August 22, 2013. Focusing attention on a product category commonly used in the RS literature, this study examines the impact of recommenders on movie-related (Blu-ray disc and DVD) product views and purchases. The field experiment was run by the company using a state-of-the-art A/B/n testing platform. This platform implemented a session tracking technology whereby each visitor’s IP address is recorded and given a unique visitor ID. Then visitors’ behaviors are tracked over the period of the field experiment. This enables the website to track individuals’ viewing logs and purchases over many days. Whenever new visitors access the website for the first time, they are randomly chosen to be in the control or one of the treatment groups. Upon clicking and viewing a particular item, the visitors are shown the appropriate recommender panel, as seen in Figure 17. Figure 17 is a collaborative filtering recommender based on views (“People who viewed this item also viewed”). Similarly, there is also a collaborative filtering based on purchases, as mentioned in Section 3.3.2. Users in the control group do not see this panel. At the end of the experiment, we have each consumer’s view logs and purchase logs at the item level. The algorithms were retrained every three days to propagate the influence of users’ purchase history multiple times over the period of the experiment. About half of the users in the dataset were returning users. The website provides its own movie categorization, but in order to make the genre categorization more robust, we categorize each movie using IMDB.com’s\(^3\) categorization. For each movie,

\(^3\)World’s top movie information website, according to Alexa rank.
we obtain IMDB.com’s category information by asking at least three different users on Amazon Mechanical Turk ("Turkers") to provide primary genres from IMDB.com. Getting input from three Turkers ensures robustness, which is important in even this simplest look-up-and-copy/paste task. Table 10 shows each genre’s product page views and purchase numbers. All data were anonymized to ensure privacy.

Figure 17: Recommender Example: Example of a recommender shown to a consumer. This consumer was in the treatment group of collaborative filtering based on views.

When the company ran the field experiment, it wanted to test the recommenders with a very small fraction of its visitors since the company had never used recommenders on this particular website. Therefore, it randomly allocated 10% of its visitors to each collaborative filtering treatment group. Our analysis focuses on only those users who made at least one purchase during the study period. This is because it does not make as much sense to compute individual-level sales diversity measures (like the Gini coefficient) when no purchases have been made by the individual. There is no selection bias in this case since, ex ante, we randomized the treatment and control assignments for all visitors. A caveat to our study is that we investigate the effects on only those consumers who ended up making a purchase. Our resulting dataset has 572 unique users in the control group and about 70 to
Table 10: Movie Genres Viewed and Purchased: This table shows the number of views and purchases in each movie genres in our dataset for those who’ve viewed or purchased movies (DVDs, Blu-ray discs) on this e-commerce site.

90 who purchased movie-related products in each of the other groups. Multiple robustness checks that account for sample size differences and outliers are presented in Section 3.5.6, and all checks produced similar results.

Our field experiment dataset offers a clean way to tease out the causal impact of recommenders. However, it is lacking in that we do not consider all recommender designs used in practice, such as content-based recommenders. However, given that much of the debate in the recommendation systems literature relates to collaborative filters, we have the two most commonly used designs in our study.
3.5. Results

We have a total of three different groups in our field experiment, giving three unique pairs to compare for each variable mentioned in Section 3.3.1. First, we present descriptive stats and summary results with visualization. Then, for each variable of interest, we present tables that show the difference in aggregate measures of interest and statistical significance associated with the differences in the measure. All results presented here are robust from sampling, outlier, and other influences, as will be discussed in Section 3.5.6.

3.5.1. Descriptive Results and Visualization

Table 11 presents descriptive summary and results of our data.

Even at the summary data descriptive level, there are clear differences across the groups. Figure 18 shows summary statistics visually. Average individual movies and genres viewed are higher in the Purchase-based CF group (8.16 and 1.84, respectively) than in the other groups (around 6.5 and 1.65, respectively). This trend persists for purchases as well, with average individual movies at 2.5 and average individual genres being bought at 1.4, compared to other groups (ranging from 1.88–2.05 for individual movies bought and 1.29–1.35 for genres bought). The average individual wallet size is also different, with collaborative filtering groups spending more, at around 27–28 dollars per person compared to 25 dollars per person in the control. Lastly, the summary-level Gini coefficients are also different. At the individual level, the Gini coefficient is lowest for the Purchase-based CF, suggesting that consumers’ individual-purchase diversity is maximized with the Purchase-based CF. At the aggregate level, the Purchase-based CF Gini coefficient is the highest, suggesting a decrease in aggregate sales diversity. As we present later in this section, many of these differences are statistically significant. Figure 19 shows the Lorenz curve for aggregate firm-level sales diversity for movie genres, clearly illustrating the decrease in aggregate diversity for CF treatments. In the next section, we formalize the analysis with permutation test technique.

1) Control vs Purchase-based CF, 2) Control vs View-based CF, 3) Purchase-based CF vs View-based CF.
Table 11: Data Summary Statistics: Standard deviation is in parentheses
Figure 18: Average Individual Statistics: These graphs visualizes the average individual number of movies viewed/purchased, wallet size (total $ spent), and the Gini measure of genres viewed/purchased.

and provide group-level differences for each variable of interest with statistical significance.

3.5.2. Sales Volume Results

**Individual Item Views** Table 12 summarizes the results for the average number of items viewed per individual. The top rows present the average number of items viewed in each
group and the difference relative to the control group. The remaining cells evaluate whether the difference between two groups is statistically significant or not. The test statistic is $D \equiv f(g_{row}) - f(g_{column})$ with $f$ representing the average number of items viewed by individuals in a group; $g_{row}$ refers to the group identified in that row. The test statistic is followed by the p-value associated with the value. The p-value is computed using a permutation test as outlined above. For example, consumers in the control group on average viewed 6.54 items, while View-based-CF-exposed consumers viewed 6.75 items. The difference between the two is $-0.2070$, and the p-value associated with this difference statistic $D$ is 0.796. Consumers treated with the Purchase-based CF viewed 1.6 more items on average (a 25% lift) compared to the control, and this difference is statistically significant. The Purchase-based CF also performs better than the View-based CF, which only lifts views by 3% compared to the control. It is clear that 1) CFs can increase exploration on e-commerce sites, and 2) there is a difference between CF algorithms in terms of their impact on product views. The success of the Purchase-based CF may be because they help broaden the consideration set by recommending other relevant alternatives or because they are effective at cross-selling other product categories. The former may not necessarily create new purchases even if it

Figure 19: Lorenz Curves for Movie Genres Purchased: These Lorenz curves show that, on the genre level, firm-level aggregate sales diversity is decreased for both collaborative filtering algorithms.
helps improve product match for consumers, whereas the latter should contribute towards new purchases.

<table>
<thead>
<tr>
<th>Avg # of Items Viewed</th>
<th>6.5464</th>
<th>6.7534</th>
<th>8.1684</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Change from Control</td>
<td>3.1% ▲</td>
<td>24.7% ▲</td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Individual Item Views Comparison: Individual items (movies) viewed averaged within each group. The top row shows the actual average items viewed in each group. Each cell shows a \textit{ROW} – \textit{COLUMN} value with \textit{p}-value (e.g., Average items viewed in control — average items viewed in View-based CF = −0.2070 and \textit{p}-value obtained from permutation test is equal to 0.796).

**Individual Item Purchases** Table 13 presents the same information for item purchases. The impact of the treatments on purchases is directionally similar to their impact on views. The Purchase-based CF is the only group that is statistically significantly different from the control group. Consumers bought 0.66 more items on average (a 35% lift) under the influence of the Purchase-based CF than the control group. Consumers in the View-based CF group bought 0.17 more items on average (a 9% lift) than consumers in the control group (the difference is not statistically significant). The results again validate that CF can have a significant impact on purchases and that there are differences between CF algorithms. Further, we find that the Purchase-based CF is effective in creating new purchases.

<table>
<thead>
<tr>
<th>Avg # of Items Purchased</th>
<th>1.8809</th>
<th>2.0547</th>
<th>2.5473</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Change from Control</td>
<td>9.2% ▲</td>
<td>35.4% ▲</td>
<td></td>
</tr>
</tbody>
</table>

Table 13: Individual Item Purchases Comparison: Individual items (movies) purchased averaged within each group.

**Individual Wallet Size** Table 14 shows that, while nothing is statistically significant in the individual wallet-size comparison, CF average wallet sizes were higher than the control’s. Both CF groups yielded an average wallet size above $27 per person. The control group
yielded the smallest average wallet size, coming up just below $25.

<table>
<thead>
<tr>
<th>Avg Wallet Size</th>
<th>24.9761</th>
<th>27.8473</th>
<th>27.2794</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Change from Control</td>
<td>11.4% ▲</td>
<td>9.2% ▲</td>
<td></td>
</tr>
</tbody>
</table>

Control | View-based CF | Purchase-based CF |
--------|--------------|------------------|
-2.8712 | 0.296 | -2.3033 | 0.416 |
0.5679 | 0.946 |

Table 14: Individual Wallet-Size Comparison: Individual Wallet Size (total amount spent) averaged within each group.

**Discussion** Our results on the *sales volume impact* of recommenders are clear. Purchase-based CF exposes users to more items, and in fact, increases the sales volume. Specifically, Purchase-based CF drives a 25% lift in views and a 35% lift in the number of items purchased compared to the control group. In comparison, the View-based-CF group shows only a 3% lift in views and a 9% lift in the number of items purchased, which is not statistically significant. While the wallet-size analysis was not statistically significant, it suggests that CFs do increase the amount spent by consumers. It is highly likely that the lack of statistical significance is driven by the fact that purchases were infrequent and our data spans only a two-week period.

3.5.3. Sales Diversity Results

**Aggregate View Diversity** Table 15 shows the aggregate view diversity at the item and genre levels for the three groups. At the item level, both treated groups decrease in view diversity (i.e., increased Gini), and the results are statistically significant. Users as a whole view fewer items when shown collaborative filtering recommendations. At the genre level, only the Purchase-based collaborative filter decreased view diversity statistically significantly compared to the control group. The results stay the same when the analysis is repeated by first fixing the same number of randomly sampled users in each group instead of permuting with the entire sample.

**Aggregate Sales Diversity** Table 16 presents the first part of our main results on the *sales diversity impact*. Subtables 16a-16b show the impact of each recommender algorithm...
on aggregate sales diversity at the item and genre levels, respectively. At the item level, both treated groups show a statistically significant reduction in aggregate sales diversity. In tandem with the results from the aggregate view diversity, we see that for both treated groups, consumers as a group explored and purchased a less wide variety of items. Users in the treated groups ended up buying the same broad-appeal items, leading to some level of herding. This popularity bias persists at the genre level. These results support the theoretical results presented in Fleder and Hosanagar (2009) that conjecture that typical
collaborative filtering designs will show a popularity bias at the aggregate level because they make recommendations based on past purchases and/or views.

Tables 15 and 16 also show some differences in terms of how recommenders influence item diversity versus genre diversity. Collaborative filtering algorithms seem to cause heavier shifts in item diversity than in genre diversity both in percentage and absolute terms. This is in part due to the number of items, which are order of magnitude greater than the number of genres, thereby amplifying the effect of popularity bias of recommenders. Collectively, these tables suggest that CF algorithms change view and purchase diversity both at the item and genre level, but more so at the item level.

<table>
<thead>
<tr>
<th>Individual View Gini</th>
<th>0.9988</th>
<th>0.9988</th>
<th>0.9983</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Change from Control</td>
<td>0%</td>
<td>0.05% ▼</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>View-based CF</td>
<td>Purchase-based CF</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>0.00001 [0.934]</td>
<td>0.0004 &lt; 0.001</td>
</tr>
<tr>
<td>View-based CF</td>
<td></td>
<td>0.0004 [0.482]</td>
<td></td>
</tr>
</tbody>
</table>

(a) Individual Item View Diversity Comparison

<table>
<thead>
<tr>
<th>Individual View Gini</th>
<th>0.9338</th>
<th>0.9315</th>
<th>0.9299</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Change from Control</td>
<td>0.2% ▼</td>
<td>0.4% ▼</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>View-based CF</td>
<td>Purchase-based CF</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td>0.0022 [0.408]</td>
<td>0.0039 [0.164]</td>
</tr>
<tr>
<td>View-based CF</td>
<td></td>
<td>0.0016 [0.814]</td>
<td></td>
</tr>
</tbody>
</table>

(b) Individual Genre View Diversity Comparison

Table 17: Individual View Diversity

**Individual View Diversity**  Table 17 presents results for individual view diversity. The only statistically significant result is at the item level with the Purchase-based CF algorithm. The Purchase-based CF causes individual view diversity to *increase* (i.e., the Gini coefficient is lower than that of the control). Unlike aggregate view diversity, which *decreases*, individual view diversity *increases*. This means that individuals view a greater variety of items while consumers as a group view a more limited variety of items. While puzzling, this result is explained by Fleder and Hosanagar’s (2009) theory that recommenders “can push each person to new products, but they often push users toward the same products” because
one person’s product views or purchases feed into recommendations made to another user.

<table>
<thead>
<tr>
<th>Individual Purchase</th>
<th>0.9985</th>
<th>0.9984</th>
<th>0.9979</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini Item</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Change from Control</td>
<td>0.01% ▼</td>
<td>0.06% ▼</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>0.0001</td>
<td>0.426</td>
<td></td>
</tr>
<tr>
<td>View-based CF</td>
<td>0.0004</td>
<td>0.946</td>
<td></td>
</tr>
</tbody>
</table>

(a) Individual Item Purchase Diversity Comparison

<table>
<thead>
<tr>
<th>Individual Purchase</th>
<th>0.9418</th>
<th>0.9390</th>
<th>0.9389</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gini Genre</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Change from Control</td>
<td>0.2% ▼</td>
<td>0.3% ▼</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>0.0028</td>
<td>0.184</td>
<td></td>
</tr>
<tr>
<td>View-based CF</td>
<td>0.0008</td>
<td>0.996</td>
<td></td>
</tr>
</tbody>
</table>

(b) Individual Genre Purchase Diversity Comparison

Table 18: Individual Purchase Diversity

**Individual Purchase Diversity** Table 18 shows the results for individual purchase diversity. Similar to the individual view diversity, individual purchase diversity increases (i.e., the Gini coefficient decreases), suggesting that individuals buy a greater variety of items because of exposure to Purchase-based collaborative filtering recommenders. The directions for genre are the same but are not statistically significant.

Looking at Tables 15–18, we see a clear pattern emerging. Under the influence of the Purchase-based CF, individuals view and buy a greater variety of items than under no recommenders, but they collectively discover and buy a similar set of items, leading to a decrease in aggregate view and sales diversity. The View-based CF shows similar results but the effect sizes are smaller and sometimes not statistically significant. This suggests that, even within collaborative filtering algorithms, there are differences. Our results further show that recommenders influence both the consideration set (views) as well as the eventual conversion (purchase).
3.5.4. Results Summary

Table 19 summarizes our results and hypotheses supported. Our results confirm anecdotal and industry reports of the sales volume impact of the recommenders, specifically for the collaborative filtering algorithms. Collaborative filtering increases both the number of product (in this case, DVD and Blu-ray) views and the number of items purchased, suggesting that these recommenders can successfully engage users in additional search and exploration and eventually additional purchases. Further, we show that not all collaborative filtering algorithms perform the same. In our setting, Purchase-based collaborative filters had a more significant impact on sales than View-based filters.

Regarding the sales diversity impact of recommenders, we have shown that collaborative filtering causes individuals to discover a greater variety of products but pushes consumers to the same set of titles, leading to concentration bias at the aggregate level. Here again, we find that not all collaborative filtering algorithms are the same. Purchase-based collaborative filters have a greater impact. This may be because: 1) the algorithm based on purchases might simply be better at delivering the best fit products, 2) consumers might be more influenced by the “purchased also purchased” signal than the “view also viewed” signal, 3) or both. However, we lack the access to detailed recommender data and consumer information to deeper investigate and quantify which of these effects are in play. However, in the next section, 3.5.5, we investigate the source of diversity shift at the genre level, providing further insight into how the increase in individual diversity and the decrease in aggregate diversity can occur simultaneously.

3.5.5. The Source of Diversity Shift: Genre Cross-pollination Investigation

In this section, we attempt to visualize our results and also understand from where the shift in aggregate and individual diversity stems. To do so, we construct sample-size normalized genre-level co-purchase networks for both Purchase-based CF and the control. We create

\footnote{We believe that investigating and quantifying the two underlying recommender mechanism is a promising and interesting line of study.}
network graphs in which each node in the graph represents a movie genre, and the size of the node is proportional to the percent of overall sales that went to the genre. An edge between two nodes indicates that there were users who purchased from these genres, and the thickness of the edge is proportional to the number of such users who exist. Thus, the relative sizes of the nodes convey the extent to which sales were (un)evenly distributed at the aggregate, and the edges convey the extent to which individuals explored content in diverse genres. Figure 20 compares two network graphs side by side.

On visually comparing the two graphs, we note the following:

1. Relative size of the nodes show that the majority of purchases by the control is distributed across a few genres: action, drama and comedy. In the Purchase-based CF, however, comedy is much bigger than the rest, indicating that purchases were more concentrated in comedy. This might be because comedy titles were recommended more often by the recommendation algorithms or because consumers are more willing to explore and trust the recommender for comedy (perhaps due to less heterogeneity in taste across the users).

2. The Purchase-based CF graph is much better connected (i.e., denser) than the control. This indicates that there are more users who are buying different genres in the Purchase-based CF group, or, more specifically, there is greater individual cross-buying behavior. The connectedness of the Purchase-based CF graph reflects the increase in individual diversity that we noted previously. Individual users may be exploring more genres while sales may be simultaneously concentrated in a few genres.
at the aggregate level.

Figure 20: Co-Purchase Network Graphs of Genre Purchases under Control and Purchase-Based Collaborative Filtering.

Figure 21 presents the frequencies of genre purchases in the control set against its frequencies in the treatment set (Purchase-based CF) to show the shift-to-comedy effect caused by the recommender. If a recommender has no influence on genre cross-pollination, we expect to see the genre on or near the line (which has slope = 1). The comedy genre is distinctively away from and above the line, showing that this recommender pushed consumers to buy more comedy. This is interesting since according to Table 10, action has the highest number of views and purchases, suggesting that this is not merely a volume effect.

<table>
<thead>
<tr>
<th>Top Genres Market Size Difference</th>
<th>Purchase-based CF Stat - Control Stat — P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
</tr>
<tr>
<td>Top Genres</td>
<td></td>
</tr>
<tr>
<td>Market Size Difference</td>
<td>0.110</td>
</tr>
</tbody>
</table>

Table 20: Permutation Test Results for Co-purchase Network Comparisons: Purchase-based CF vs. Control

To formally test these differences, we use the same permutation technique to evaluate the market size of the top genres in each graph (market size of top N nodes). Table 20 shows the
difference between the market-size statistics for the Purchase-base CF and control group as well as the corresponding p-values obtained via permutation tests. We replicated the analysis for top \{1, 5, 10\} genres. We see a clear shift to top genres in Purchase-based CF. Under the Purchase-based CF, the top genre, comedy, took 11% more market share compared to the top genre in the control group, action. We also see a decrease in top genres market size differences (i.e., 0.11 \rightarrow 0.095 \rightarrow 0.081) as we include more number of genres, suggesting higher concentration of purchases for the top genres. In summary, a Purchase-based collaborative filtering algorithm shifts users to buy a few top genres at the aggregate level while increasing individual diversity through a cross-buying behavior that is aided by a few ‘pathway’ genres.

3.5.6. Robustness Checks & Other Measures

We also ran a series of robustness checks in regards to consumer samples. They include the following:

Figure 21: Genre Purchase Share Comparison on Purchase-based CF vs. Control
R1. Our analysis thus far is based on treatment groups of different sizes. We replicated the analysis by randomly sampling a fixed number of users in each group, ex ante before the permutation test, so that each group has an equal number of consumers.

R2. We removed the outlier consumers in terms of number of views and purchased items. We replicated the analysis excluding those users whose views or purchases exceeded 3 standard deviations from the mean.

R3. We replicated the analysis by using 1 or 0 (binary variable) for genre or items viewed (or purchased) instead of actual counts. This allows us to explore the notion of diversity from a perspective of number of unique items or genres viewed/purchased as opposed to proportion of sales.

R4. We replicated the analysis only on consumers who bought more than one item.

R5. Some groups (e.g., Purchase-based CF) see more views and purchases than others. In theory, this increase in volume should not affect sales diversity measures such as the Gini coefficient. To confirm this, we replicated the analysis after “Volume-Equalization” in which the number of items bought by individuals was normalized across groups to get rid of any effect from volume influencing the Gini coefficient.\(^6\)

In all of these robustness checks, the findings are qualitatively similar to our main result. Table 21 presents the details of hypotheses supported under the different robustness checks.

3.6. Discussion and Conclusion

With the advent of big data, recommenders and personalization technologies are fast taking over nearly every aspect of the web. Their use spans from the purchase of physical products (books, DVDs, clothing, electronics, etc) to digital media (movies, news), and even online services such as dating and peer-to-peer lending. Despite their ubiquity, we still have much to learn about how different recommender algorithms influence markets and society.

\(^6\)We followed the method used in Hosanagar et al. (2014).
Hypotheses Supported under Robustness Checks

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: CF Increases product views</td>
<td>BOTH</td>
<td>BOTH</td>
<td>DS−</td>
<td>BOTH</td>
<td>NA</td>
</tr>
<tr>
<td>2: CF increases the # of product purchases</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>NA</td>
</tr>
<tr>
<td>3: CF decreases aggregate product sales diversity</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
</tr>
<tr>
<td>4: CF increases individual product consumption diversity</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>BOTH</td>
<td>DS+</td>
</tr>
</tbody>
</table>

Table 21: Hypotheses under Robustness Checks: “Both” means that both directions and statistical significances were identical to the presented main results. “D” represents that direction was reproduced. “S−” signifies the loss of statistical significances, and “S+” signifies the gain of statistical significance. Lastly, “NA” indicates not applicable due to the nature of the robustness test (e.g., “Volume-Equalization” necessarily gets rid of volume differences).

Our study contributes to an emerging literature on the impact of personalization technologies by studying the impact of recommender algorithms on sales volume and diversity with movie products. We have three main findings. First, we show that recommenders have a positive impact on sales, which corroborates anecdotal evidence and prior findings in the literature. Second, we provide direct evidence from the field that collaborative filtering algorithms increase individual consumption diversity while decreasing aggregate consumption diversity and explain where this shift in diversity stems from at the genre-level. Third, for both sales volume and diversity, we show that different algorithms have different impact, extending baseline results for sales volumes while providing new insights for sales diversity. Furthermore, the result highlights a potential limitation in the ability of traditional collaborative filtering to aid discovery of truly niche items and genres. We reveal that users increase their purchase and view diversity by exploring into few top ‘pathway’ genres like comedy. Our study also helps resolve an ongoing debate among researchers on the impact of recommenders on sales diversity by providing an evidence from a randomized field experiment with individual view and purchase level data.

These results have significant managerial relevance. As the amount of consumer data available to firms grows exponentially, many retailers have aggressively adopted data mining and personalization technologies without deeply understanding how different designs may contribute toward (or deter) broader strategic goals. For example, a firm interested in exposing
consumers to a broader assortment of products may prefer a different design from another simply interested in maximizing sales. To the extent that a firm is interested in pushing the “back catalog,” it may seek to augment traditional collaborative filtering algorithms so that it is possible to identify relevant products with limited historical data (past views/purchases) and/or increase diversity, serendipity, or novelty of the recommended products using techniques from the extant literature (e.g., Adamopoulos and Tuzhilin (2013); Adomavicius and Kwon (2013); Oh et al. (2011)).

On a more general level, many firms have adopted theory-free predictive analytics approaches without trying to understand what drives changes caused by the online technologies that they have implemented. As data grows, this approach will increasingly be prone to issues tied to spurious correlations and a decrease in the signal-to-noise ratio. One promising alternative is to use theory-driven causal inference techniques to discern true effects. This is one of the first randomized field experiment explicitly looking at the differential effects of different personalization algorithms on sales volume and diversity. We look forward to additional studies documenting these differences in other field settings.

We conclude by discussing some limitations of our study and opportunities for future work. Our study focused on the two most commonly used collaborative filtering designs. It is worth investigating the impact of other recommender designs, such as content-based and social-network-based recommenders. Second, we evaluated the impact of collaborative filters in one product category, namely, movies. A promising extension of our work will be in creating an empirical generalization by including other product categories such as apparel and electronics in the study and investigating specific product characteristics that influence the sales volume and diversity. Third, a valuable addition to our work will be studies that develop consumer behavior theories on how and why people react differently to different recommender systems and signals. Lab studies can be highly valuable in this regard.
CHAPTER 4: When do Recommender Systems Work the Best? The Moderating Effects of Product Attributes and Consumer Reviews on Recommender Performance

4.1. Introduction

Recommender systems are now ubiquitous on the web. E-commerce sites regularly use such systems to guide consumers with prompts like “People who purchased this item also purchased...” to increase up-selling and cross-selling opportunities. Recommenders aid online shopping by reducing search cost (Anderson, 2008) and product uncertainty for consumers (Bergemann and Ozmen, 2006). As such, many existing studies have already shown that recommender systems increase revenue and profitability for firms Anderson (2008); Bodapati (2008); Das et al. (2007); Fleder and Hosanagar (2009); Hosanagar et al. (2014); Jannach and Hegelich (2009); Monetate (2013); Oestreicher-Singer and Sundararajan (2012); Thompson (2008). Consequently, according to a study by Econsultancy and Monetate (2013), 94% of e-commerce sites now consider recommendation systems to be critical competitive advantage to be implemented. At the same time however, the same study reveal that only about 15% of the company were getting good return on investment and 72% attributed failure to lack of knowledge on recommender systems. This is because recommenders almost always coexist with other factors and features on web that influence purchase decisions through product uncertainty levels. For example, different products have different search cost (Hann and Terwiesch, 2003) and product uncertainty (Dimoka et al., 2012), while user-generated reviews reduce product uncertainty. As such, effective implementation of recommenders must account for complicated interaction with these factors. However, there is a lack of literature on how the impact of recommenders are moderated by other factors such as types of items sold, item attributes, and consumer-generated reviews. In this study, through a randomized field experiment, we investigate how factors that influence product uncertainty

---

1Product uncertainty is defined as the consumer’s difficulty in evaluating product attributes and predicting how a product will perform in the future (Hong and Pavlou, 2014).
online, such as product attributes and consumer reviews, interact with a recommender system to affect conversion rate, defined as the percentage of product views that result in purchases.

Existing studies have shown that utilizing recommender systems in e-commerce settings lead to an increase in usage, revenue, and profitability - in short, an increase in sales volume. Anderson (2008); Bodapati (2008); Das et al. (2007); De et al. (2010); Dias et al. (2008); Fleder and Hosanagar (2009); Hosanagar et al. (2014); Jannach and Hegelich (2009); Monetate (2013); Oestreicher-Singer and Sundararajan (2012); Thompson (2008). Other studies have investigated the impact of recommenders on sales diversity (Hinz and Eckert, 2010; Matt et al., 2013; Hosanagar et al., 2014; Fleder and Hosanagar, 2009; Oestreicher-Singer and Sundararajan, 2012; Jannach et al., 2013), in which the focus was to study how the use of recommender systems influences the assortment of items viewed and purchased by consumers. While it is clear that the use of a recommender system generally leads to an increase in sales volume and influences sales diversity, there is a lack of investigation on how product-specific attributes or reviews influence the effectiveness of recommenders. Researchers and managers still don’t know under what conditions and for what products a recommender system works well. Specifically, there is a lack of actual field studies that investigate the interaction between other factors that influence product purchase decisions (e.g., product-level attributes and review data) and the efficacy of a recommender system to generate conversion. How do certain item attributes increase or decrease the effectiveness of recommender systems in causing purchases? For example, are recommenders substitutes or complements for high review ratings and review volumes? Will a recommender system cause more or fewer purchases for highly priced items? How about for hedonic vs. utilitarian product or search vs. experience products? Many of these highly insightful and managerially impactful questions are not answered or are partially answered due to limited data. The lack of access to a field experiment setting covering a wide range of products and the sheer amount of resources required to content-code attributes of a large number of products are just a few reasons for this gap. Answers to the questions above can guide
recommender implementation in e-commerce and provide insight into consumer purchase behavior in online settings.

Our study attempts to address these gaps by running a randomized field experiment on an e-commerce site of a top retailer in North America\(^2\). We run a randomized experiment with recommender treatment and control groups, then proceed to identify several key product attributes of more than 37,000 unique items viewed or purchased during the period of the field experiment. We utilize Amazon Mechanical Turk to efficiently content-code a large number of items and item attributes. After augmenting the dataset with the consumer review data pulled from APIs (Application Programming Interface), we run logistic regressions to tease out the moderating effects of product attributes in causing conversion under the use of recommenders.

Briefly, our main results show the follow. We first confirm that the use of a recommender increases the conversion rate in general (by 5.9%), but this increase is highly moderated by product attributes. For example, the higher the price, the lower the positive influence of recommenders. We also find that while the baseline conversion rate is higher for utilitarian products online, benefit from recommenders is higher for hedonic products compared to utilitarian products. We find that contrary to conjectures from existing literature, the search-experience attribute does not influence the power of recommenders. Furthermore, we find that the use of a recommender increases conversion rates as much as approximately 2 additional stars out of 5 in average review ratings. While the higher review volume increases conversion rates, once recommenders are accounted for, the volume no longer had any effect on conversion. Essentially, recommenders act as substitutes for high average review ratings. Besides these, we have many more insights with more details in the results section.

Our results provide both broad and specific insights for understanding the moderating effects of product attributes on the power of recommender systems. This study makes several

\(^{2}\)We are not allowed to disclose the identity of the company. But it is one of the biggest companies offline, also ranking top 5 in e-commerce revenue worldwide.
contributions. From an academic standpoint, ours is the first individualized field experiment study to look at the moderating effects of product attributes like price, hedonic-utilitarian quality, search-experience quality, and review data on a recommender with individual-level conversion field data. By working with a retailer that ranks top 5 in the world in e-commerce revenue and sells the most expansive list of product categories, we increase external validity. At the practice, our study has several managerial implications. First, managers can determine which specific products would be best served by recommenders and which would not. Second, managers will have insight into how other e-commerce features, such as product descriptions and user-generated reviews, interact with the power of recommenders. Managers can then optimize e-commerce sites appropriately and decide which features (e.g., reviews, more descriptions, recommenders) to implement in combination. Ultimately, we provide insight for better utilizing recommenders online for increased conversion rates.

4.2. Data

Our main dataset consists of complete individual-item level views and purchase transactional data from running a field experiment. The cooperating company that ran the experiment randomly assigned incoming new customers into either a treated group, in which the recommendation panel is shown, or a control group, in which the recommendation panel is not shown. We capture click-stream data as well as eventual conversion data. This dataset is augmented with 1) complete review data from the pages of all the products appearing in the dataset and 2) item attributes separately tagged via a survey instrument and workers on Amazon Mechanical Turk, an online marketplace for data tagging and cleaning.

4.2.1. Field Experiment & Data Description

With the cooperation of one of the top retailers in North America, we ran the field experiment on their e-commerce site for a two-week period in August 2013. The company has both an online and offline presence and is one of the top 3 in the North American region by size and revenue. It’s e-commerce presence is ranked top 5 in the world with more than $10
billion in e-commerce revenue alone in 2014. The company ran the field experiment using a state-of-the-art A/B/n testing platform. This platform implements a session tracking technology whereby each visitor’s IP address is recorded and given a unique visitor ID. Then, visitors’ behaviors are tracked over the period of the field experiment. This enables the website to track individuals’ viewing logs and purchases over the period of field experiment duration. Whenever new visitors access the website for the first time, they are randomly chosen to be in the control group or in the treatment group. Upon clicking and viewing a particular item, the visitors assigned to the treated group are shown a recommender panel, as seen in Figure 22. Visitors in the control group do not see this panel. There are many types of recommender systems and it is infeasible to run all types of recommender systems in the field experiment setting due to the amount of resources required to implement and opportunity cost for the retailer. In order to increase the external validity, we utilize the most common type of recommender system used in the industry, a purchase-based collaborative filtering algorithm - “People who purchased this item also purchased” (Adomavicius and Tuzhilin, 2005). The specific algorithm used in the study is obtained from the most widely used open-source machine learning framework called the Apache Mahout (mahout.apache.org).

The dataset, which spans 355,084 rows of individual-item transactional records, tracks 184,375 unique users split into 92,188 treated users and 92,187 control users. Users clicked and viewed details of 37,215 unique items and bought 3,642 unique items and a total of 9,761 items. In addition, we collected review data of all items appearing in the dataset.

3https://www.internetretailer.com/top500/?cid=2014-IRAGP

4Within Personalized Recommenders systems, a broad taxonomy distinguishes three types of algorithms: Content-based, Collaborative Filtering, and Hybrid, which combines the first two. Content-based systems analyze product attributes to suggest products that are similar to those that a consumer bought or liked in the past. Collaborative filtering recommenders, unaware of product attributes, recommend products either purchased or liked by similar consumers, where similarity is measured by historical purchase (or like) data. We discovered through talking to a large e-business analytics firm, which implements recommenders for many clients, that out of about 300 firms, only 3 utilized content-based recommenders. The rest utilized purchase-based collaborative filtering. A majority of companies utilize collaborative filtering algorithm simply because content-based recommender systems require expensive attribute tagging and content analysis. One prominent exception is Pandora.com (a music genome project) that managed to content-code a large library of songs.
Figure 22: Recommendation Panel: Example of a recommender shown to a consumer. Most commonly used recommender algorithm, “People who purchased this item also purchased”, is used.

The retailer’s description of the item, categorization including the subcategorization to the maximum depth, and more. Table 22 shows the top-level category appearance in the data and Table 23 gives the summary of the data. At the top level, the retailer has 18 categories including house appliances, automotive, electronics, movies, furniture, jewelry, and so on. We carefully chose the retailer with one of the most extensive coverage of SKUs and product categories to increases the external validity of the results.

<table>
<thead>
<tr>
<th>Products Appearance in Data by Categories as Classified by the Retailer</th>
<th>Top Level Categorization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appliances</td>
<td>Automotive</td>
</tr>
<tr>
<td>29545</td>
<td>5366</td>
</tr>
<tr>
<td>Grocery</td>
<td>Halloween</td>
</tr>
<tr>
<td>8422</td>
<td>6</td>
</tr>
<tr>
<td>Movies Music &amp; Books</td>
<td>Office &amp; Stationery</td>
</tr>
<tr>
<td>26000</td>
<td>12352</td>
</tr>
</tbody>
</table>

Table 22: Product Categories Occurring In the Dataset: The first level product categorization as classified by the retailer online. There are in total 4 levels of depths and subcategories. 1st depth has 18 categories, 2nd → 149, 3rd → 884, and 4th → 492.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>REC</td>
<td>Recommender system treatment condition. 1 means the user was randomly selected to be shown recommendations.</td>
<td>Treatment</td>
<td>0.503</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PRICE</td>
<td>Item price.</td>
<td>Site</td>
<td>85.94</td>
<td>120.69</td>
<td>0.01</td>
<td>998.00</td>
</tr>
<tr>
<td>DESLEN</td>
<td>Length of item description on the site.</td>
<td>Site</td>
<td>269.71</td>
<td>251.06</td>
<td>0</td>
<td>3882</td>
</tr>
<tr>
<td>AVGRATING</td>
<td>Average review star rating out of 5.</td>
<td>Site</td>
<td>2.44</td>
<td>2.22</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>RATINGNUMB</td>
<td>The number of reviews the item obtained.</td>
<td>Site</td>
<td>12.46</td>
<td>107.93</td>
<td>0</td>
<td>19407</td>
</tr>
<tr>
<td>BRAND</td>
<td>% of Amazon Mechanical Turkers who recognized the brand. Asked 5 Turkers per item.</td>
<td>AMT</td>
<td>0.53</td>
<td>0.35</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DURABILITY</td>
<td>Durability of the item. Likert scale from 1–7 with 7 being the most durable.</td>
<td>AMT</td>
<td>4.97</td>
<td>1.37</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>UTILHEDO</td>
<td>Classification into utilitarian or hedonic product. 1 if utilitarian, 0 if hedonic.</td>
<td>AMT</td>
<td>Util 18529 Hed 18596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEARCHEXP</td>
<td>Classification into search or experience product. 1 if search, 0 if experience.</td>
<td>AMT</td>
<td>Sea 15798 Exp 21327</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Views</td>
<td>For a given user-item session, the number of times the user viewed the item.</td>
<td>Site</td>
<td>1.3</td>
<td>0.79</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>Quantity</td>
<td>The number ordered.</td>
<td>Site</td>
<td>0.02</td>
<td>0.32</td>
<td>0</td>
<td>48</td>
</tr>
</tbody>
</table>

Table 23: Variable Descriptions and Summary for Content-coded Data.

### 4.2.2. Product Attribute Tagging on Amazon Mechanical Turk

Given the data from the field experiment, we still need to identify product attributes of interest. With more than 37,000 unique number of items, it is challenging to identify many product attributes at this scale. We have identified several product attributes motivated by extant literature to analyze for the products in our dataset. We discuss these attributes and relevant literature in Section 4.3. We now describe our methodology for identifying product attributes using Amazon Mechanical Turk (AMT). AMT is a crowd sourcing marketplace for simple tasks such as data collection, surveys, and photo and text analyses. To obtain product attributes for a given item, we create a survey instrument based on existing con-
structs, operating definitions, and measurement questions previously used in other studies. To ensure high-quality responses from the Turkers, we follow several best practices identified in literature (e.g., we obtain tags from at least 5 different Turkers choosing only those who are from the U.S., have more than 500 completed tasks, and an approval rate more than 98%. We also include an attention-verification question.) Please see the appendix for the measurement questions used and the complete list of strategies implemented to ensure output quality.

Ultimately, we achieve values greater than 0.8 for all the constructs in Krippendorff’s Alpha, a inter-rater reliability measure in which any value above 0.8 is accepted in the literature as a satisfactory outcome. We end up utilizing a several thousand unique AMT workers answering many questions about more than 37,000 unique items.

4.3. Product Attributes & Hypotheses

Extant literature in consumer economics, marketing, and information systems research have identified many product attributes that influence purchase decisions. Relating to products sold online on e-commerce sites, the literature has identified information uncertainty (Stigler, 1961; Arrow, 1963) related to product uncertainty and search cost (Degeratu et al., 2000; Bakos, 1997; Johnson et al., 2004; Wu et al., 2004; Hong and Pavlou, 2014; Kim and Krishnan, 2013; Lauraeus-Niinivaara et al., 2007) to be one of the main deterrents in product purchase decisions. Focusing on product-related uncertainty, the main aspects of product uncertainty online is description and performance uncertainty (Dimoka et al., 2012), defined “as the buyer’s difficulty in assessing the product’s characteristics and predicting how the product will perform in the future.” Similarly, Liang and Huang (1998) have shown that 1) different products do have different customer acceptance on the electronic market and 2) the customer acceptance is determined by the transaction cost, which is in turn

---

5 Another reliability measure, Cronbach’s Alpha, produced the same result.
6 We do not consider buyers experience and retailer uncertainty in this study. Buyer experience is not a concern since we randomize a large number of users into different groups. The retailer uncertainty is not a concern since our retailer is one of the most recognized retailers in the world. In fact, many company ranking lists rank our retailer as number one among US retailers.
determined by the uncertainty and asset specificity. Hann and Terwiesch (2003) have also shown that different products have different search costs associated with them. Lastly, connecting product type and complexity to recommenders on e-commerce sites, Xiao and Benbasat (2007), Aggarwal and Vaidyanathan (2005), and Senecal and Nantel (2004) suggested that product type and complexity may influence users’ acceptance of recommender systems. Thus in this paper, we analyze factors that influence product uncertainty in the online setting, which may influence recommender performance: product attributes and consumer-generated product reviews.

Product uncertainty can be ameliorated via product descriptions and reviews up to a certain point but this reduction also heavily depends on the type of product and the consumers’ willingness to search. For example, Nelson’s 1970s seminal work on economics of information and advertising (Nelson, 1970, 1974) classified products into search and experience goods. Search goods are dominated by characteristics and attributes that can be discerned prior to purchase and are often objective in nature. Experience goods are dominated by characteristics that can only be discerned by using the product or are subjective in nature. Nelson’s search and experience framework has been used to explain how people react to advertising, search for different products online, and ultimately make purchases (Klein, 1998; Klein and Ford, 2003). Another product attribute that may influence purchase decision is the hedonic-utilitarian framework. Hedonic (pleasure-oriented consumption) or utilitarian (goal-oriented consumption) purpose related to a product (Dhar and Wertenbroch, 2000; Khan et al., 2005) has been shown to change the way consumers shop online. For example, this attribute interacts with uncertainty reducing mechanisms such as reviews and descriptions online. Sen and Lerman (2007) show that online consumers trust negative reviews more for utilitarian products. There are many other attributes that influence purchase decision via difference in information cost and product uncertainty. As such, we posit that these product attributes will also influence the effectiveness of recommender systems, commonly acknowledged as an electronic word-of-mouth or another source of information for awareness and product fit. In this paper, we look at the impact of these product attributes
in an e-commerce setting in which recommenders are implemented.

Since it is infeasible to go through all of product attributes, we have focused our attention on identifying product attributes that 1) are shown in word-of-mouth and online review literature to influence consumers purchase behavior, 2) are clear and simple in concept for maximal managerial implication, 3) have strong theoretical background with existing and well-used operational definition and measurement survey questions. Following these criteria, we have identified several control variables as well as main variables of interest that may influence the effectiveness of a recommender. We next discuss each variable, related literature, how we tagged the attributes using extant operating definitions, and our hypotheses on how each will moderate the power of a recommender system. Details and sources of survey instruments for measuring product attributes are discussed in the Appendix.

4.3.1. Product Attributes

Hedonic VS. Utilitarian

A product characteristic often discussed and used to categorize products across industries is whether the product is dominantly a utilitarian product or a hedonic product (Dhar and Wertenbroch, 2000; Strahilevitz and Myers, 1998; Hirschman and Holbrook, 1982). The literature (Dhar and Wertenbroch, 2000; Strahilevitz and Myers, 1998; Hirschman and Holbrook, 1982) define utilitarian goods as those for which consumption is cognitively driven, instrumental, goal-oriented, and accomplishes a functional or practical task. Hedonic goods are defined as ones whose consumption is primarily characterized by an affective and sensory experience of aesthetic or sensual pleasure, fantasy, and fun. Broadly, the hedonic-utilitarian attribute has been shown to influence consumer product search behavior, purchase decisions, and even consumers’ value of products (Hirschman and Holbrook, 1982; Bart et al., 2014; Khan et al., 2005).

Connecting to online shopping, studies have shown that consumers are more goal-oriented
and utilitarian motivated online. Consumers with utilitarian motivation shop online for convenience, cost savings, and readily available information online (To et al., 2007). Since utilitarian goods dominantly consist of objective attributes that serve specific functions (e.g., hammer, memory card, and ink toners) and are apt for goal-oriented shopping, consumers may use online shopping for utilitarian products more than for hedonic products. As such, we posit that the baseline conversion rate is higher for utilitarian product.

Relating to recommender systems, extant literature have shown that the hedonic-utilitarian attribute moderates the trust and re-use intention of recommender systems. For example, Choi et al. (2011) suggests that consumers’ trust for recommender systems and re-use intention is increased when the recommender provides a “social presence”, defined as “the extent to which a website allows users to experience others as psychologically present”. This increase in trust and re-use intention is greater for hedonic products compared to utilitarian products. Extending along these lines, we draw from past advertising literature to theorize how hedonic-utilitarian attributes may moderate the power of recommender systems in directly increasing conversion rates. Studies have shown that the effectiveness of product endorsement depends on whether the product is utilitarian or hedonic (Feick and Higie, 1992; Stafford et al., 2002). When consumers are shopping for a utilitarian product, the purchase decisions are guided by information about objective functional attributes. As such, consumers prefer expert endorsers. However, for hedonic products with many subjective attributes and high heterogeneity in preferences, it’s been suggested that consumers prefer opinions of people who are more like them (Feick and Higie, 1992). The collaborative filtering algorithm implemented in our dataset provides recommendations to a consumer based on purchase histories of other consumers similar to the consumer and signal this clearly. Thus, we posit that conversion rates will be increased for hedonic products under the use of recommender systems since recommenders claim to reveal preferences of similar consumers. Thus, our hypotheses are as follows.

**Hypothesis 5** The base conversion rate for utilitarian goods will be higher in online settings.
Hypothesis 6 *The increase in conversion rate under the use of a recommender will be higher for hedonic goods, compared to utilitarian goods.*

To measure and classify an item into a hedonic or a utilitarian product, we surveyed the extant literature and found several operating definitions and measurement questions. One measurement survey defines hedonic and utilitarian values and for each value, asks to rate the product on a 1 to 7 likert scale. This results in two separate measurements for utilitarian and hedonic quality. Another scale condenses this into one scale starting from purely utilitarian to purely hedonic in intervals. We asked all three as seen in Table 24 to at least five different Turkers, then took mean values. Finally, based on these three dimensions, the k-means clustering algorithm (Hartigan, 1975) was used to classify products into two clusters: utilitarian or hedonic. The cluster means for each product are shown in Table 24. The Appendix has the full list of questions used, question sources, and the inter-rater reliability measure.

<table>
<thead>
<tr>
<th>Measurement Questions</th>
<th>Utilitarian Product Cluster Mean</th>
<th>Hedonic Product Cluster Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedonic Value</td>
<td>2.28</td>
<td>6.17</td>
</tr>
<tr>
<td>[1 NOT AT ALL HEDONIC to 7 PURELY HEDONIC]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Utilitarian Value</td>
<td>5.98</td>
<td>1.95</td>
</tr>
<tr>
<td>[1 NOT AT ALL UTILITARIAN to 7 PURELY UTILITARIAN]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Please give the scale on how much comparative utilitarian VS hedonic value the product offers. [1 PURELY UTILITARIAN to 7 PURELY HEDONIC]</td>
<td>2.19</td>
<td>5.97</td>
</tr>
</tbody>
</table>

Table 24: Utilitarian VS. Hedonic Product Cluster Means: Definition given is in the appendix.
Search VS. Experience

Philip Nelson’s seminal work on economics of information and advertising (Nelson, 1970, 1974) classified products into search and experience goods. Search goods consist of attributes that can easily be discerned before purchase and are dominated by attributes with lower informational search cost and objective attributes, such as the speed and memory of a computer. In contrast, experience goods consist of attributes that cannot easily be discerned before purchase and are dominated by attributes with higher information search cost and subjective attributes like taste of wine or the entertainment value of movies. Nelson originally theorized and calculated the total cost of the product as the sum of the product cost and the consumers’ search cost. Following this work, numerous studies in economics, marketing, and information systems have investigated how this search and experience classification of product influence consumers’ search, consideration set, and purchase behavior (Klein, 1998; Klein and Ford, 2003; Girard and Dion, 2010; Huang et al., 2009; Hsieh et al., 2005; Krishnan and Hartline, 2001; Hong and Pavlou, 2014; Dimoka et al., 2012). Specifically, in online settings, product information uncertainty and higher search cost for experience goods has been shown to be a major hurdle and challenge for e-commerce managers (Hong and Pavlou, 2014; Dimoka et al., 2012; Weathers et al., 2007; Girard et al., 2003). While experience goods like wine, cosmetics, apparel, etc are increasingly sold on e-commerce sites, these sites still find it challenging to satisfy consumers’ information needs to convert, or satisfy them enough to prevent high rates of return (Hong and Pavlou, 2014; Dimoka et al., 2012). A few studies have suggested several remedies like the use of search engines, multimedia product descriptions, and finally recommender systems to overcome high search costs (e.g., Hinz and Eckert (2010); De et al. (2013, 2010)). However, literature lacks studies on comparing search vs experience goods in the context of recommender systems. Traditionally, recommender systems were popularized on experience goods like movies, music, and books. However, now recommenders are being utilized for all types of products and we can compare the differential impact.
Nelson theorized that consumers’ search for experience goods will be characterized by heavier reliance on word-of-mouth and experience of other consumers since the cost of information via other routes are more costly (Nelson, 1970, 1974; Klein, 1998). Consequentially, Nelson hypothesized that experience goods sellers will focus on persuasive and brand-focused tactics such as word-of-mouth, testimonials, and celebrity endorsements while search goods sellers will prioritize their advertising with informative and easy to discern facts about the products. However, it is not clear how search-experience attribute will influence recommenders’ performance. Extant literature on the moderating influence of search-experience attribute on the power of recommenders is limited and conflicting. Senecal and Nantel (2004) found evidence that consumers are more influenced by recommendations for experience products than for search products. However, this study has a limited external validity due to the artificial nature of lab experiment in recommender settings and the fact that it is based on only two products, wine and calculators. Contrastingly, a study by Aggarwal and Vaidyanathan (2005), with again only two products, suggest a conflicting result. Aggarwal and Vaidyanathan (2005) claim that consumers perceived recommenders to be more effective for search goods than for experienced goods. Thus, the extant literature is lacking in both results based on realistic field data and based on an expansive list of products.

Ultimately, the power of a recommender to result in conversion for search or experience goods depends on consumers’ trust of the recommender system. If the consumers trust recommenders to serve as a replacement for costly search, the recommendations should be more effective when used for experience goods. Recent literature in recommender systems has dubbed the recommender agents as “digitized word-of-mouth” (Chen et al., 2009) where consumers adapt and trust recommender systems as “social actors” and perceive human characteristics (Benbasat and Wang, 2005; Xiao and Benbasat, 2007; Komiak and Benbasat, 2004). Essentially, consumers are increasingly trusting recommenders to replace searching when the search cost is high. Nelson’s theory suggest that consumers rely more on word-of-mouth for experience goods and recent literature have shown that recommender systems are accepted and trusted as a form of word-of-mouth. While the baseline conversion rate
for search goods online may be higher due to lowered search cost, product information uncertainty, and product fit uncertainty (Dimoka et al., 2012; Hong and Pavlou, 2014), recommenders may be better received by consumers for experience goods based on Nelson’s theory. In accordance with Nelson’s theory on experience goods and the role of recommender systems online, we develop the following hypothesis.

**Hypothesis 7** *The base conversion rate for search goods will be higher in online settings.*

**Hypothesis 8** *The increase in conversion rate under the use of a recommender will be higher for experience goods, compared to search goods.*

To measure and classify an item into a search or a experience product, we surveyed the extant literature and found several operating definitions and measurement questions. We found two sets of questions repeatedly used in the literature. One set of questions, used widely in marketing literature, asks the consumers to answer two questions: how well could you judge the attribute or quality of the product 1) before they have purchased it and 2) after they have purchased it. If the consumers can judge the attributes not so well before the purchase but well after the purchase, the literature has classified those products as experience goods while for search goods, consumers can judge the quality of the product well even before the purchase. Another set of questions asked similar questions related to the search cost. We combined these questions in the extant literature and asked in total 4 questions on the Likert scale. Once we obtained the answers for each product from at least five different Turkers, we took the mean value for each answer. Finally, we used the k-means clustering algorithm (Hartigan, 1975) to classify products into two clusters: search or experience. The cluster means for search and experience products are shown in Table 25. The Appendix has the full list of questions used, question sources, and the inter-rater reliability measure.
Measurement Questions

[1 NOT WELL/IMPORTANT AT ALL to 7 EXTREMELY WELL/IMPORTANT]

<table>
<thead>
<tr>
<th>Search Good Cluster Mean</th>
<th>Experience Good Cluster Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>How well could you judge the attributes or quality of this product even BEFORE you purchased or used it?</td>
<td>4.82</td>
</tr>
<tr>
<td>How well could you judge the attributes or quality of this product even AFTER you purchased or used it?</td>
<td>6.36</td>
</tr>
<tr>
<td>How important is it for you to see, touch, hear, taste, smell (whichever applies) this product IN PERSON to evaluate its attributes?</td>
<td>3.15</td>
</tr>
<tr>
<td>How well can you evaluate the product using only information provided by retailer and/or manufacturer about this product’s attributes and features?</td>
<td>5.04</td>
</tr>
</tbody>
</table>

Table 25: Search VS. Experience Product Cluster Means

Consumer Reviews

It is well documented in the literature that user-generated reviews influence online consumers’ purchase intentions (Chen et al., 2004; Chen and Xie, 2008; Duan et al., 2008; Sun, 2012; Chevalier and Mayzlin, 2006; Berger, 2014). However, results are mixed in that review ratings always do not influence consumers, while other studies show that the effect of reviews on sales are moderated depending on the nature of the product — which can increase search-cost — whether it’s a niche or experiential item (Li and Wu, 2013; Duan et al., 2008; Dai et al., 2014; Chen and Xie, 2008; Zhu and Zhang, 2010). Specifically, consumers tend to discount or even ignore review ratings when the volume of the review is low (Li and Wu, 2013; Duan et al., 2008; Chen et al., 2004). For niche or less-popular items, the impact of reviews can be greater (Zhu and Zhang, 2010). Specifically, high ratings have a more positive influence on consumer purchase intentions for niche items (Tucker and Zhang, 2011). Similarly, the herding effect for purchase has been found to be more salient for experience goods than for search goods in online settings (Li and Wu, 2013). Ultimately, all of these results are consistent with the search-cost argument in which consumers rely more on external informational sources like reviews when the search-cost is higher (e.g., niche item or experience items). Consumers rely on reviews as a source of information and do so selectively based on the search-cost related to products.
Recommender systems are electronic word-of-mouth (Chen et al., 2009) and reduce uncertainty and search-cost for consumers online (Clemons, 2008) just as consumer reviews do. However, it is not clear if recommender systems act as substitutes or complements to reviews since they serve similar yet slightly different purpose. Reviews mainly reduce uncertainty and provide product fit information while recommenders increase awareness and provide personalized product fit information. The cost of consumption is also different in that it takes a longer time to process review ratings (mean and variance) and to read the reviews compared to getting a straight forward recommendation - given that consumers trust recommenders. In fact, the acceptance and pervasiveness of recommenders in e-commerce have grown so much that a majority of consumers now expect and prefer websites that provide personalized recommendations (Accenture, 2012). Since recommender systems provide personalized fit information on top of consumer reviews, it’s likely that recommenders may serve as a substitute for consumer reviews and provide additional information and value for consumers. If the consumers do trust the recommenders as the Accenture survey suggests, it is possible that with the existence of personalized recommendations, consumers may discount other people’s reviews. Another reason consumers may discount higher review ratings when recommenders are present is because consumers may not agree with other consumers’ reviews. Indeed, Dai et al. (2014) claims that consumers rely less on other consumers’ reviews when shopping for experience products because consumers believe that other people’s reviews are not representative of their own evaluations. Since the cost of consumption is lower for recommender systems, in extreme cases and depending on products, consumers may not even bother to check the reviews. In summary, consumers may trust personalized recommendations from the website more than review data.

Based on discussed theory, we posit that while higher review ratings may increase conversion rates in the absence of a personalized recommender system, with the presence of a personalized recommender system, its positive influence may be lessened. It is likely that given the personalized recommendation by an algorithm, the high review ratings may have less impact on conversion. In other words, recommenders act as substitutes for reviews.
Our hypotheses are as follows:

**Hypothesis 9** The base conversion rate will be increased for products with higher review ratings.

**Hypothesis 10** The positive impact on conversion from high review ratings will be lessened under the presence of a recommender system.

All hypotheses are listed in Table 28.

### 4.3.2. Control Attributes

In addition to the attributes discussed above, we include the following control attributes in the model:

1. Durability: We asked 5 distinct Turkers to rate on a Likert scale from 1 to 7, with 7 being extremely durable, on how durable the product is.

2. Description Length: The retailer provides description of all products sold on the website. We get the length to proxy for the amount of information provided.

3. Brand Awareness Proxy: We asked 5 distinct Turkers if they recognized the brand of the item. We then take the percentage of the Turkers who answered “Yes” as a proxy measure for brand prominence.
4.4. Model & Results

4.4.1. Model

The conversion rate given recommendation treatment for user \(u\) and product \(i\)’s attributes are modeled as a logistic regression.

\[
\log\left(\frac{P(\text{conv}_{iu})}{1 - P(\text{conv}_{iu})}\right) = \beta_0 + \beta_1 \text{PRICE}_i + \beta_2 \text{REC}_u + \beta_3 \text{UTILHEDO}_i + \\
\beta_4 \text{SEARCHEXP}_i + \beta_5 \text{DURABILITY}_i + \beta_6 \text{BRAND}_i + \beta_7 \text{DESLEN}_i + \\
\beta_9 \text{RATINGNUMB}_i + \beta_{10} \text{PRICE}_i \times \text{REC}_u + \beta_{10} \text{UTILHEDO}_i \times \text{REC}_u + \\
\beta_8 \text{AVGRATING}_i + \beta_{11} \text{SEARCHEXP}_i \times \text{REC}_u + \beta_{10} \text{DESLEN}_i \times \text{REC}_u + \\
\beta_{10} \text{AVGRATING}_i \times \text{REC}_u + \beta_{11} \text{RATINGNUMB}_i \times \text{REC}_u + \epsilon_u
\]

We estimate the parameters \(\beta_{0-11}\) by maximum likelihood estimator.

4.4.2. Results

Table 26 provides results from running the logistic regression. We first discuss the baseline hypotheses and results before presenting the main results on interaction with the recommenders. The stand-alone main effect results confirm previous literature. The impact of price on conversion is negative and significant (-0.0016) while the effect of recommenders is positive and significant (0.17998). The result corroborates the extant literature in that using a recommender system indeed increases conversion rates, and thus the sales volume (Hosanagar et al., 2014; Lee and Hosanagar, 2014). In this experiment, the use of recommenders increased the baseline conversion rate by 5.9%. Description length of products provided by the retailer had no significant influence on base conversion rate. Keeping everything else the same, higher average product review ratings increase the conversion rate (0.10651) as shown previously by Chevalier and Mayzlin (2006) and Sun (2012). Higher durability was associated with lower conversion rate (-0.19567). This result is likely since high durability is correlated with higher price and lower purchase frequency, and thus higher perceived risk.
(Jacoby et al., 1971; Pavlou, 2003) and lower willingness to purchase, especially in online settings. Lastly, our proxy variable for brand prominence was directionally positive and weakly significant with p-value = 0.6. Next, we discuss our main hypotheses and results regarding interaction between product attributes (and reviews) and recommender systems.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Log odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.40089***</td>
<td>0.05973</td>
<td>0.033</td>
</tr>
<tr>
<td>PRICE</td>
<td>-0.00163***</td>
<td>0.00021</td>
<td>0.998</td>
</tr>
<tr>
<td>REC</td>
<td>0.17998**</td>
<td>0.05886</td>
<td>1.197</td>
</tr>
<tr>
<td>DESLEN</td>
<td>-0.00004</td>
<td>0.00008</td>
<td>0.999</td>
</tr>
<tr>
<td>AVGRATING</td>
<td>0.10651***</td>
<td>0.00818</td>
<td>1.112</td>
</tr>
<tr>
<td>RATINGNUMB</td>
<td>-0.00007</td>
<td>0.00015</td>
<td>0.999</td>
</tr>
<tr>
<td>UTILHEDO (UTIL=1)</td>
<td>0.26382***</td>
<td>0.03718</td>
<td>1.301</td>
</tr>
<tr>
<td>SEARCHEXP (SEA=1)</td>
<td>0.14281***</td>
<td>0.03561</td>
<td>1.153</td>
</tr>
<tr>
<td>BRAND</td>
<td>0.06783</td>
<td>0.03636</td>
<td>1.070</td>
</tr>
<tr>
<td>DURABILITY</td>
<td>-0.19567***</td>
<td>0.00841</td>
<td>0.822</td>
</tr>
<tr>
<td>REC X PRICE</td>
<td>-0.00115***</td>
<td>0.00031</td>
<td>0.998</td>
</tr>
<tr>
<td>REC X DESLEN</td>
<td>0.00027***</td>
<td>0.0001</td>
<td>1.000</td>
</tr>
<tr>
<td>REC X AVGRATING</td>
<td>-0.03562**</td>
<td>0.0114</td>
<td>0.965</td>
</tr>
<tr>
<td>REC X RATINGNUMB</td>
<td>0.00019</td>
<td>0.00016</td>
<td>1.000</td>
</tr>
<tr>
<td>REC X UTILHEDO</td>
<td>-0.16565**</td>
<td>0.05169</td>
<td>0.847</td>
</tr>
<tr>
<td>REC X SEARCHEXP</td>
<td>0.00654</td>
<td>0.04989</td>
<td>1.006</td>
</tr>
</tbody>
</table>

Table 26: Logistic Regression Results Table: ‘*’= p-value < 0.05, ‘**’= p-value < 0.01, ‘***’= p-value < 0.001

**Hedonic VS. Utilitarian**

The main effect of hedonic-utilitarian attribute (1 if utilitarian, 0 if hedonic) show higher conversion rate for utilitarian products online at 0.26382. The effect is statistically significant and greater than any other effects including the use of recommender systems. This supports our hypothesis that the base conversion rate for utilitarian goods will be higher in online settings keeping everything else constant. As To et al. (2007) suggests, consumers are utilizing e-commerce more for utilitarian purposes. Hedonic products often have attributes related to sense and beauty that consumers need to experience beforehand and is less bought online where price, convenience, and reduced search-cost may be the primary reasons for conversion. Interaction term with recommender treatment is negative and statistically significant at -0.16565. This suggests that while consumers purchase utilitarian products more in general in e-commerce settings, recommenders increase conversion more for hedonic prod-
ucts. The use of recommenders for utilitarian products still increases conversion since the main effect minus the interaction term is positive ($0.17998 - 0.16565 = 0.01433$).

The result is consistent with the story that consumers buying online are mainly motivated by utilitarian reasons of price, convenience, and reduced search-cost. Given that recommenders primarily serve as another source of information to increase the awareness set and to reduce search-cost, utilitarian products, which already have lower search-cost on the internet, benefit less from the use of recommenders. Recommenders are effective for hedonic goods.

**Search VS. Experience**

The main effect of the search-experience attribute (1 if search, 0 if experience) shows a higher conversion rate for search products online at 0.14281. The effect is statistically significant and positive, thus supporting our hypothesis that the base conversion rate for search goods will be higher in online settings. This corroborates existing theory that (Nelson, 1970; Pavlou, 2003; Dimoka et al., 2012) search goods with less informational cost attributes have less deterrent for purchase in online settings. However, the interacted term with recommender treatment was not statistically significant while directionally positive. The results do not support our hypothesis that the conversion rate will be higher for experience goods under the use of a recommender system. The results suggest that the original conjecture by Nelson (1970), that consumers will rely more on word-of-mouth and experience of others for experience goods, doesn’t seem to carry over to a recommender system. While recommenders are theorized as “digitized word-of-mouth” (Chen et al., 2009), it is possible that a simple signal such as “other consumers who’ve purchased this item also purchased” does not provide enough details or reduction in uncertainty to particularly work well on experience products. Another explanation may be that consumers do not believe other consumers’ preferences accurately reflect their tastes as suggested by Dai et al. (2014) in cases of reviews. Since our dataset spans expansive categories of products sold on websites, we
sought to replicate results of Senecal and Nantel (2004) (recommendations for experience products like wine were more influential than recommendations for search products like calculators) and Aggarwal and Vaidyanathan (2005) (that recommenders are received more favorably for search goods). Depending on the product category chosen, we were able to replicate the results that support both arguments. However, when everything in the dataset is considered, search-experience attributes do not seem to moderate the effectiveness of this particular recommender.

**Consumer Reviews**

The main effect of average review ratings had a positive impact on conversion at the baseline at 0.10651. This means that approximately 2 additional stars out of 5 in review ratings increases log-odds ratio as much as the use of recommender systems\(^7\). Contrary to a previous study (Duan et al., 2008) that showed that review volumes are associated with higher sales, our results show that once the recommenders are accounted for, the review volume does not have any impact on baseline conversion rates in e-commerce settings (RATINGNUMB coefficient is -0.00007 and statistically not significant). The interaction term with recommender treatment and average ratings suggest that the positive impact on conversion from high review ratings will be lessened under the presence of a recommender system with estimate at -0.03562. This supports our hypothesis that consumers rely less on high average ratings once the recommenders are introduced.

To further investigate the interaction between review ratings, review volume, and recommender systems, we run multiple specifications in Table 27. The model on the first column, with only review volume, corroborates the results by Duan et al. (2008), which claimed that high review volumes increase conversion. Column 2 confirms that higher average rating increases the baseline conversion rate. However, once recommenders are accounted for, the rating volume does not matter and the positive impact of high average rating is lessened.

\(^7\)That is, the increase in log-odds ratio from using a recommender, 0.179, is approximately twice that of 0.106. However, it is likely that increase in conversion is nonlinear for average star ratings from 0 to 5.
Ultimately, our results suggest that recommenders serve as substitutes for review volume and higher review ratings in causing conversion.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-3.93436***</td>
<td>-4.06221***</td>
<td>-4.11774***</td>
</tr>
<tr>
<td>RATINGNUMB</td>
<td>0.00012*</td>
<td>-0.00002</td>
<td></td>
</tr>
<tr>
<td>AVGRATING</td>
<td>0.05053***</td>
<td>0.07224***</td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>0.10843**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC X RATINGNUMB</td>
<td></td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>REC X AVGRATING</td>
<td>-0.04396***</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 27: Multiple Specifications for Review Related Variables: ‘*’ = p-value < 0.05, ‘**’ = p-value < 0.01, ‘***’ = p-value < 0.001

Lastly, we summarize our findings and hypotheses supported in Table 28. We also summarized other takeaways in Table 29.

<table>
<thead>
<tr>
<th>Attribute Construct</th>
<th>Hypotheses</th>
<th>Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedonic-Utilitarian</td>
<td>The base conversion rate for utilitarian goods will be higher in online settings</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Hedo-Util × Rec</strong></td>
<td>The increase in conversion rate under the use of a recommender will be higher for hedonic goods, compared to utilitarian goods</td>
<td>YES</td>
</tr>
<tr>
<td>Search-Experience</td>
<td>The base conversion rate for search goods will be higher in online settings</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Sea-Exp × Rec</strong></td>
<td>The increase in conversion rate under the use of a recommender will be higher for experience goods, compared to search goods</td>
<td>NO</td>
</tr>
<tr>
<td>Review Rating</td>
<td>The base conversion rate will be increased for products with higher review ratings</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Review Rating × Rec</strong></td>
<td>The positive impact on conversion from high review ratings will be lessened under the presence of a recommender system</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 28: Hypotheses and Results

4.4.3. Measurement Robustness

For both hedonic-utilitarian and search-experience attributes, we utilized the clustering algorithm to classify a product dichotomously into a hedonic or utilitarian product, as well as a search or experience product. The decision to use dichotomous classifications was for practical convenience and to use existing measurement strategies. While the literature has acknowledged the shortcomings of dichotomous classification schemes, it is still commonly used in the literature based on dominant attributes (e.g., Huang et al. (2009), Senecal
<table>
<thead>
<tr>
<th>Attribute Construct</th>
<th>Result Takeaways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Durability</td>
<td>The higher the durability, the lower the baseline conversion rate online.</td>
</tr>
<tr>
<td>Price</td>
<td>The higher the price, the lower the baseline conversion rate. Additionally, the higher the price, the lower the benefit of recommender.</td>
</tr>
<tr>
<td>Description Length</td>
<td>Description length did not influence the baseline conversion rate. However, longer description increased the benefit of a recommender.</td>
</tr>
<tr>
<td>Brand</td>
<td>Brand prominence showed weak positive effect ($p$-value = $0.06$) on baseline conversion rate.</td>
</tr>
<tr>
<td>Review Volume</td>
<td>The higher the review volume, the higher the conversion rate. However, once recommenders are accounted for, high review volume did not influence conversion.</td>
</tr>
</tbody>
</table>

Table 29: Other Takeaways

and Nantel (2004)). However, since these product attributes could be continuous qualities, we repeated analyses in which the search-experience and hedonic-utilitarian attributes are denoted by a scale from 1 to 7. We obtain qualitatively similar results.

4.5. Conclusion and Discussion

While recommenders are prevalent in e-commerce and have been shown to increase sales volume in multiple studies, effective use and implementation of recommenders still elude a majority of e-commerce managers and retailers as shown in studies such as Econsultancy and Monetate (2013). We believe that this is due to the lack of holistic investigation of conversion process that influence purchase decisions other than the recommenders. This study addresses this gap and adds empirical results.

This paper examined the interaction between a recommender system and product attributes along with reviews in e-commerce setting. Several product attributes were found to influence the power of recommenders in causing consumers to ultimately buy products. Our results reproduced several baseline hypotheses regarding the impact of product attributes on e-commerce shopping and extended existing baseline hypotheses to incorporate the impact on and interaction with recommender systems. The results show rich interaction between the effectiveness of recommenders and a variety of product attributes and review. We show
that recommenders act as substitutes for high average review ratings and review volumes. Additionally, we find that baseline positive impact on conversion from recommenders are reduced for utilitarian products compared to hedonic products while search-experience quality did not have any impact. We also find that the higher the price, the lower the positive impact of recommenders, while providing longer product descriptions increased the power of recommenders.

Given these findings, managers have several key takeaways for implementing effective recommender strategies. Our study suggests effective ways to utilize recommender systems. For example, since our results suggest that recommenders act as substitute for high review volume and higher average rating for conversion, e-commerce sites with low review volumes could prioritize recommender implementations. We also show that a longer product description increases recommender effectiveness. While sites selling utilitarian products may still benefit from the use of recommenders, the benefit was not as substantial as using it on hedonic products. Utilitarian product sellers may want to utilize the limited webspace for other content before a recommender system.

One shortcoming of our paper is that we used only one type of recommender system: purchase-based collaborative filtering. However, we carefully chose an algorithm (i.e., collaborative filtering over content-based) that is most widely used after researching industry reports and companies in this area, and utilized an open-source implementation (Apache Mahout) most widely used by e-commerce sites. We believe that our results have high external validity due to the retailer we worked with and the expansive list of products covered in the study.

---

8One of the largest e-business and A/B/n testing company that implements recommenders reported that out of about 300 company clients, only 3 were using content-based recommenders and most companies were using purchase-based collaborative filtering recommenders.
A.1. Survey Instrument

We are interested in identifying the content of social media messages. Please read the following short message posted on a social media site and answer the following questions. Thank you.

(Message Content)

Indicate the message characteristics by (Yes/No).

The Message,
1: contains an interesting or remarkable fact.
2: includes emotional content.
3: uses humor.
4: includes philanthropic, awareness or activist content (promotes social welfare, social change, or informs the reader about certain facts to improve people's lives).
5: mentions a company (organization) name or brand name (e.g. Nike, Red Cross, VitaminWater or musicians such as The Beatles).
If yes, please type the company or brand name as it appears in the message in the following box (only one per box)
Most relevant company/brand mentioned ____________
Second company/brand mentioned (if any)________
6: provides information about any type of discounts or freebies (e.g., deals, coupons, promotional offers, sweepstakes, rewards, free items).
7: makes a price comparison or a price-match guarantee against competitors' product(s) or service(s).
8: contains content other than about a product or company business (e.g. small talk, social pleasantness, casual banter, “Happy Halloween”, “what was your day like?”).
9: is targeted towards an audience segment. E.g., a particular demographic (race, gender, age, location) or people with certain qualifications or characteristics. (e.g. “all moms, do you like Gerber?” , “All chocolate lovers, get this deal!”).
10: provides information about the availability of a product or service (e.g. Only 5 deals remain, available only until tomorrow, new product coming tomorrow).
11: provides information about where or how to obtain a product or service (e.g. link, physical location to buy, general location)?
12: does the message have a dollar sign ($)?
13: Are your friends on social media likely to post messages such as the above?
14: The message is specific to a particular product or service (e.g. electronics, drinks, music, etc.).
If yes, please type the mentioned product or service:
Most relevant product mentioned ____________

Figure 23: Survey Form Used in Amazon Mechanical Turk

A.2. Amazon Mechanical Turk – Robust Content Extraction

Following best-practices in the literature, we employ the following strategies to improve the quality of classification by the Turkers in our study.

1. For each message, at least 9 different Turkers' inputs are recorded. We obtain the final classification by a majority-voting rule.

2. We restrict the quality of Turkers included in our study to comprise only those with at least 100 reported completed tasks and 97% or better reported task-approval rates.
3. We use only Turkers from the US so as to filter out those potentially not proficient in English, and to closely match the user-base from our data (recall, our data has been filtered to only include pages located in the US).

4. We refined our survey instrument through an iterative series of about 10 pilot studies, in which we asked Turkers to identify confusing or unclear questions. In each iteration, we asked 10–30 Turkers to identify confusing questions and the reasons they found those questions confusing. We refined the survey in this manner till almost all queried Turkers stated no questions were confusing.

5. To filter out participants who were not paying attention, we included an easily verifiable test question “does the message have a dollar sign ($)?”. Responses from Turkers that failed the verification test are dropped from the data.

6. In order to incentivize workers, we awarded additional bonuses of $2–$5 to the top 20 workers with exceptional accuracy and throughput.

7. On average, we found that message tagging took a little over 3 minutes and it typically took at least 20 seconds or more to completely read the tagging questions. We defined less than 30 seconds to be too short, and discarded any message tags with completion times shorter than that duration to filter out inattentive Turkers and automated programs (“bots”).

8. Once a Turker tags more than 20 messages, a couple of tagged samples are randomly picked and manually examined for quality and performance. This process identified about 20 high-volume Turkers who completed all surveys in less than 10 seconds and tagged several thousands of messages (there were also Turkers who took time to complete the surveys but chose seemingly random answers). We concluded these were automated programs. These results were dropped, and the Turkers “hard blocked” from the survey, via the blocking option provided in AMT.
We believe our methodology for content-classification has strong external validity. The binary classification task that we serve to the AMT Turkers in our study is relatively simpler than the more complex tasks for which AMT-based data have been employed successfully in the literature. The existing AMT literature has documented evidence that several of the strategies implemented above improves the quality of the data generated (Mason and Suri (2012); Ipeirotis et al. (2010); Paolacci et al. (2010)). Snow et al. (2008) show that combining results from a few Turkers can produce data equivalent in quality to that of expert labelers for a variety of text tagging tasks. Similarly, Sheng et al. (2007) document that repeated labeling of the type we implement wherein each message is tagged by multiple Turkers, is preferable to single labeling in which one person tags one sentence. Finally, evaluating AMT based studies, Buhrmester et al. (2011) concludes that (1) Turkers are demographically more diverse than regular psychometric studies samples, and (2) the data obtained are at least as reliable as those obtained via traditional methods as measured by psychometric standards such as Cronbach’s Alpha, a commonly used inter-rater reliability measure.

A.3. NLP Algorithm

This section provides detailed outline of the algorithm used in the paper. Figure 24 shows the process visually.

A.3.1. Training The Algorithm

1. The raw textual data of 5,000 messages in the training sample are broken down into basic building blocks of sentences using stop-words removal (removing punctuation and words with low information such as the definite article “the”), tokenization (the process of breaking a sentence into words, phrases, and symbols or “tokens”), stemming (the process of reducing inflected words to their root form, e.g., “playing” to “play”), and part-of-speech tagging (determining part-of-speech such as nouns). For reference see Jurafsky and Martin (2008). In this process, the input to the algorithm

115
is a regular sentence and the output is an ordered set of fundamental linguistic entities with semantic values. We use a highly regarded python NLP framework named NLTK (Bird et al., 2009) to implement this step.

2. Once the messages are broken down as above, an algorithm extracts sentence-level attributes and sentence-structure rules that help identify the included content. Some examples of sentence-level attributes and rules include: frequent noun words (bag-of-words approach), bigrams, the ratio of part-of-speech used, tf-idf (term-frequency and inverse document frequency) weighted informative word weights, and whether “a specific key-word is present” rule. For completeness, we describe each of these in Table 30. The key to designing a successful NLP algorithm is to figure out what we (humans) do when identifying certain information. For example, what do we notice about the sentences we have identified as having emotional content? We may notice the use of certain types of words, use of exclamation marks, the use of capital letters, etc. At the end of this step, the dataset consists of sentence-level attributes generated as above (the $x$-variables), corresponding to a series of binary (content present/not-present) content labels generated from AMT (the $y$-variables).

3. For each binary content label, we then train a classification model by combining statistical and rule-based classifiers. In this step, the NLP algorithm fits the binary content label (the $y$-variable) using the sentence-level attributes as the $x$-variables. For example, the algorithm would fit whether or not a message has emotional content as tagged by AMT using the sentence attributes extracted from the message via step 2. We use a variety of different classifiers in this step including logistic regression with L1 regularization (which penalizes the number of attributes and is commonly used for attribute selection for problems with many attributes; see (Hastie et al., 2009)), Naive Bayes (a probabilistic classifier that applies Bayes theorem based on presence or absence of features), and support vector machines (a gold-standard algorithm in machine learning that works well for high dimensional problems) with different fla-
vors of regularization and kernels.\textsuperscript{1} To account for imbalance in positive and negative class labels in some content, we utilized combination of class-weighted classifiers and resampling methods.

4. To train the ultimate predictive classifier, we use ensemble methods to combine results from the multiple statistical classifiers we fit in step 3. The motivation for ensemble learning is that different classifiers perform differently based on underlying characteristics of data or have varying precision or recall in different locations of the feature vector space. Thus, combining them will achieve better classification output either by reducing variance (e.g. Bagging (Brieman, 1996)) or reducing bias (e.g. Boosting (Freund and Schapire, 1995)). Please see Xu and Krzyzak (1992); Bennett (2006) for further reading on ensemble methods. This step involves combining the prediction from individual classifiers by weighted-majority voting, unweighted-majority voting, or a more elaborate method called isotonic regression (Zadrozny and Elkan, 2002) and choosing the best performing method in terms of accuracy, precision and recall for each content profiles. In our case, we found that support vector machine based classifiers delivered high precision and low recall, while Naive Bayes based classifiers delivered high recall but low precision. By combining these, we were able to develop an improved classifier that delivers higher precision and recall and in effect, higher accuracy. Table 31 shows the improvement of the final ensemble learning method relative to using only one support vector machine. As shown, the gains from combining classifiers are substantial. We obtain similar results for negative class labels.

5. Finally, we assess the performance of the overall NLP algorithm on three measures, viz., accuracy, precision, and recall (as defined in Footnote 4) using the “10-fold cross-validation” method. Under this strategy, we split the data randomly into 10 equal subsets before the step 2. One of the subsets is used as the validation sample, and the algorithm trained on the remaining 9 sets. This is repeated 10 times, each

\textsuperscript{1}We tried support vector machines with L1 and L2 regularization and various kernels including linear, radial basis function, and polynomial kernels. For more details, refer to Hastie et al. (2009).
time using a different subset as the validation sample, and the performance measures averaged across the 10 runs. The use of 10-fold cross-validation reduces the risk of overfitting and increases the external validity of the NLP algorithm we develop. Note, 10-fold cross-validation of this sort is computationally intensive and impacts performance measures negatively and is not implemented in some existing papers in business research. While the use of 10-fold cross-validation may negatively impact the performance measures, it is necessary to increase external validity. Table 31 shows these metrics for different content profiles. The performance is extremely good and comparable to performance achieved by the leading financial information text mining systems (Hassan et al., 2011).

6. We repeat steps 2–5 until desired performance measures are achieved.

Tagging New Messages

1. For each new messages repeat steps 1–2 described above.

2. Use the ultimate classifier developed above to predict whether a particular type of content is present or not.

One can think of this NLP algorithm as emulating the Turkers’ collective opinion in content-coding.
Figure 24: Diagram of NLP Training and Tagging Procedure: This diagram shows the steps of training the NLP algorithm and using the algorithm to tag the remaining messages. These steps are described in Appendix A.3.
<table>
<thead>
<tr>
<th>Rules and Attributes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bag of Words</td>
<td>Collects all the words and frequency for a message. Different variations include collecting top $N$ most occurring words.</td>
</tr>
<tr>
<td>Bigram</td>
<td>A bigram is formed by two adjacent words (e.g. “Bigram is”, “is formed” are bigrams).</td>
</tr>
<tr>
<td>Ratio of part-of-speech</td>
<td>Part-of-speech (noun, verb, etc) ratio in each message.</td>
</tr>
<tr>
<td>TF-IDF weighted informative word</td>
<td>Term-Frequency and Inverse Document Frequency weighs each word based on their occurrence in the entire data and in a single message.</td>
</tr>
<tr>
<td>Specific Keywords</td>
<td>Specific keywords for different content can be collected and searched. e.g., Philanthropic messages have high change of containing the words “donate” and “help”. For brand and product identification, large online lists were scraped and converted into dictionaries for checking.</td>
</tr>
<tr>
<td>Frequency of different punctuation marks</td>
<td>Counts the number of different punctuations such as exclamation mark and question mark. This helps to identify emotion, questions, appearance of deals etc.</td>
</tr>
<tr>
<td>Count of non-alphanumerics</td>
<td>Counts the number of characters that are not A–Z and 0–9.</td>
</tr>
</tbody>
</table>

Table 30: A Few Examples of Message Attributes Used in Natural Language Processing Algorithm
Table 31: Performance of Text Mining Algorithm on 5000 Messages Using 10-fold Cross-validation: This table presents metrics for performance of the classification algorithms used. The left 3 columns show the metrics for the final algorithm which combines classifiers via ensemble learning method while the right 3 columns show the metric for a support vector machine algorithm. Notice that the support vector machine classifier tends to have low recall and high precision. Naive Bayes tends to have high recall but low precision. Classifiers on their own cannot successfully overcome precision-recall tradeoff (if one is higher, one is lower). But combining many different classifiers with ensemble learning can increase both precision and recall. We obtain similar results for negative class labels.
B.1. Amazon Mechanical Turk Strategy & Survey Instrument for Item Attribute Tagging

Following best-practices in the literature, we employ the following strategies to improve the quality of attribute tagging by the Turkers in our study.

1. For each message, at least 5 different Turkers’ inputs are recorded.

2. We restrict the quality of Turkers included in our study to comprise only those with at least 500 reported completed tasks and 98% or better reported task-approval rates.

3. We screened out the workers by giving them a simple test to see if they understood the instructions. Those who failed were banned from participating.

4. We use only Turkers from the countries where English is the primary language to filter out those potentially not proficient in English.

5. We refined our survey instrument through an iterative series of about several pilot studies, in which we asked Turkers to identify confusing or unclear questions. In each iteration, we asked 10-30 Turkers to identify confusing questions and the reasons they found those questions confusing. We refined the survey in this manner till almost all queried Turkers stated no questions were confusing.

6. To filter out participants who were not paying attention, we included an easily verifiable attention question. Responses from Turkers that failed the attention test are dropped from the data.

7. On average, we found that survey took a little over 4 minutes and it typically took at least 1 minute or more to completely read the questions. We defined less than
30 seconds to be too short, and discarded any message tags with completion times shorter than that duration to filter out inattentive Turkers and automated programs (“bots”).

8. Once a Turker tags more than 100 messages, a couple of tagged samples are randomly picked and manually examined for quality and performance. This process identified several high-volume Turkers who completed all surveys in less than 15 seconds and tagged several thousands of messages (there were also Turkers who took time to complete the surveys but chose seemingly random answers). We concluded these were automated programs. These results were dropped, and the Turkers “hard blocked” from the survey, via the blocking option provided in AMT.

The existing AMT literature has documented evidence that several of the strategies implemented above improves the quality of the data generated (Mason and Suri (2012); Ipeirotis et al. (2010); Paolacci et al. (2010)). Snow et al. (2008) demonstrates that combining results from a few Turkers can produce data equivalent in quality to that of expert labelers for a variety of tagging and content-coding tasks. Similarly, Sheng et al. (2007) document that repeated labeling of the type we implement wherein each message is tagged by multiple Turkers, is preferable to single labeling in which one person tags one sentence. Finally, evaluating AMT based studies, Buhrmester et al. (2011) concludes that (1) Turkers are demographically more diverse than regular psychometric studies samples, and (2) the data obtained are at least as reliable as those obtained via traditional methods as measured by psychometric standards such as Cronbach’s Alpha or Krippendorff’s Alpha, commonly used inter-rater reliability measures.

The following table provides the construct we’ve used, literature sources we’ve adapted the measurement survey instrument and operating definitions, and inter-rater reliability measure achieved.
<table>
<thead>
<tr>
<th>Construct &amp; Measurement</th>
<th>Question Sources (Krippendorff’s Alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hedonic VS. Utilitarian (0.9455)</td>
<td>Adapted from Dhar and Wertenbroch (2000); Strahilevitz and Myers (1998); Bart et al. (2014); Khan et al. (2005); Babin et al. (1994)</td>
</tr>
<tr>
<td>Measurement Questions (Likert Scale from 1- Least 7-Most)</td>
<td>Product consumption is driven by different motives. A couple of example motivations are based on the idea of hedonic (want) consumption vs. utilitarian (need) consumption. Hedonic, Pleasure-oriented consumption is motivated mainly by the desire for sensual pleasure, fantasy, and fun (e.g., movies, perfume, art piece). Utilitarian, goal-oriented consumption is motivated mainly by the desire to fill a basic need or accomplish a functional task (e.g., paper clips, dishwashing agent, vacuum cleaner). Given the above definition of hedonic and utilitarian value of a product, rate the product above in the scale below on hedonic value and utilitarian value.</td>
</tr>
<tr>
<td>Hedonic Value [1 NOT AT ALL HEDONIC to 7 PURELY HEDONIC]</td>
<td>Utilitarian Value [1 NOT AT ALL UTILITARIAN to 7 PURELY UTILITARIAN]</td>
</tr>
<tr>
<td>Please give the scale on how much comparative utilitarian VS hedonic value the product offers. [1 PURELY UTILITARIAN to 7 PURELY HEDONIC]</td>
<td></td>
</tr>
<tr>
<td>Search VS. Experience (0.8433)</td>
<td>Adapted from Krishnan and Hartline (2001); Hsieh et al. (2005); Huang et al. (2009); Girard and Dion (2010); Klein (1998); Klein and Ford (2003)</td>
</tr>
<tr>
<td>Measurement Questions (Likert Scale from 1- Least 7-Most)</td>
<td>• How well could you judge the attributes or quality of this product even BEFORE you purchased or used it? [1 NOT WELL AT ALL to 7 EXTREMELY WELL]</td>
</tr>
<tr>
<td></td>
<td>[For example, some products are easy to judge the attributes/quality of BEFORE you’ve purchased or used them (e.g., Computers, Printer Ink) while others (e.g., Movies, Food, Wine) are not.]</td>
</tr>
<tr>
<td></td>
<td>• How well could you judge the attributes or quality of this product even AFTER you purchased or used it? [1 NOT WELL AT ALL to 7 EXTREMELY WELL]</td>
</tr>
<tr>
<td></td>
<td>[For example, some products are easy to judge the attributes/quality of AFTER you’ve purchased or used them (e.g., Movies, Food, Wine)]</td>
</tr>
<tr>
<td></td>
<td>• How important is it for you to see, touch, hear, taste, smell (whichever applies) this product IN PERSON to evaluate its attributes? [1 NOT IMPORTANT AT ALL to 7 EXTREMELY IMPORTANT]</td>
</tr>
<tr>
<td></td>
<td>[For example, you may want to touch a piece of clothing to determine the quality of fabric, but this may not be necessary for printer toner or vitamin.]</td>
</tr>
<tr>
<td></td>
<td>• How well can you evaluate the product using only information provided by retailer and/or manufacturer about this product’s attributes and features? [1 NOT WELL AT ALL to 7 EXTREMELY WELL]</td>
</tr>
<tr>
<td>Durability</td>
<td>Please rate how durable the product is Some products are extremely durable and do not quickly wear out (e.g., Cars and Mobile Phones) while others are less durable and wear out quickly (e.g., Food, Gasoline, Papers, Medications). Assume average usage and no accident.</td>
</tr>
<tr>
<td>Brand Prominence Proxy</td>
<td>Have you heard of the brand/company that made this product?</td>
</tr>
</tbody>
</table>

Table 32: Survey Instrument: We use existing and widely used operational definitions and measurement questions to tag the items in our dataset. Median Krippendorff’s alpha, a standard measure of inter-rater reliability measure is provided and are well above the acceptable measure of 0.8.


H. Dai, C. Chan, and C. Mogilner. ‘don’t tell me what to do!’people rely less on consumer reviews for experiential than material purchases. People Rely Less on Consumer Reviews for Experiential than Material Purchases (September 8, 2014), 2014.


eMarketer. Digital set to surpass tv in time spent with us media, August 2013b.

eMarketer. Global b2c ecommerce sales to hit 1.5 trillion dollar this year driven by growth in emerging markets, 2014.


L. R. Klein and G. T. Ford. Consumer search for information in the digital age: An


P. A. Pavlou. Consumer acceptance of electronic commerce: Integrating trust and risk


R. Snow, B. O’Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast - but is it good?


WSJ. Gm says facebook ads don’t pay off, 2012.


