MEASURING AND ANALYZING "VOICE":
IMPLICATIONS FOR CONSUMERS, MANAGERS AND REGULATORS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Dinesh Puranam

August 2016

Measuring and Analyzing "Voice":

Implications for Consumers, Managers and Regulators

Dinesh Puranam, Ph. D.

Cornell University 2016

My dissertation focuses on measuring "voice" from unstructured data and understanding its implications for consumers, firms and regulators. The main thrust of my essays is to answer interesting substantive questions of relevance to consumers (essay on Amazon's Vine program), marketers (essay on Uber Technologies) and regulators (essay on calorie posting regulation) using text data. I also seek to address empirical issues that arise in the extracting measures from text (essay on optimal number of topics). I am optimistic that my empirical research on "voice" can potentially answer substantive questions that are difficult to address with traditional data, in a rigorous modeling framework.

BIOGRAPHICAL SKETCH

Dinesh Puranam graduated with a master's in Economics from the Delhi School of Economics in 2002. Prior to joining the doctoral program at Cornell in 2011, he was an analytics and business process consultant for close to a decade. He has worked on assignments for P&G, GE, Discover Financial, Fingerhut and Hess Energy in China, Mexico, India, the UK and the USA.

To my parents

ACKNOWLEDGMENTS

I am extremely grateful to my Co-Chairs – Vrinda Kadiyali and Vishal Narayan and my committee members Claire Cardie and David Mimno – for steering me in the right direction and for making this dissertation a reality. I am indeed fortunate to have had such intellectually rigorous, accessible and patient faculty to support me in my research. It simply would not have been possible without them.

I thank Vrinda Kadiyali for being an extraordinary support – both intellectually and otherwise – over the last 5 years. I thank her for teaching me to stay anchored in important, research questions. She consistently went out of his way to create wonderful opportunities for me to grow intellectually and to find my feet in the profession. I thank her for instilling in me a love for great ideas, clean thinking and the entire research process.

Vishal Narayan was an outstanding mentor, champion and steady source of encouragement and constructive advice throughout my PhD. Over the course of several discussions with him I have acquired a deep appreciation of the rigors of the intellectual process and the advantages of considering multiple perspectives. I thank him for taking an active interest in me right from the beginning. He has continuously challenged me to be ambitious and creative about my work while reposing in me the faith to deliver.

Claire Cardie was exceptionally generous in supporting me in my forays in Natural Language Processing and Machine Learning. Her consistent guidance and support – be it choosing coursework or working on a research project- were crucial to my understanding of the discipline. Despite having a small army of students to guide, she always found time to discuss my research and offer insights.

David Mimno was a phenomenal teacher and guide - sharing with me his deep knowledge in methods, oftentimes at short notice, and to have many conversations to help me clear the cobwebs in my thinking. He contributed substantially helping me gain an appreciation for pitfalls and opportunities in empirical methods. He also taught me the value of not letting go on an interesting question.

I also thank Manoj Thomas for rounding out my training in Marketing by exposing me to the behavioral side of the discipline. He was always available to offer a fresh perspective on my work and to offer words of encouragement.

Many faculty members at the Johnson School helped me refine my thinking and ideas as I delved deeper into the research process. I would especially like to thank Sachin Gupta, Stijn Van Osselaer and Vithala Rao. Each of them generously spent time discussing my research and offered suggestions – many of which are reflected in this dissertation.

I thank my fellow graduate student Joowon Park for sharing those many coffee breaks that were unintentionally enlightening but always entertaining. I would also like to thank Myle Ott for the many fun and interesting conversations on Machine Learning.

A big thank you to Annie Johnston for being a phenomenal support in helping me through the graduate school's academic requirements and timelines. Her generosity and kindness at critical junctures indeed.

I thank my parents and my brother– for inevitably inquiring each week about my academic progress. Having researchers in the family adds a motivational twist to family conversations. I can never thank them enough.

In the end this entire endeavor would not have seen the light of day without my wife, Parul. Of course, I entered the tunnel (program) in the first place because of her. I simply followed her into graduate school (as I have followed her elsewhere) and have greatly enjoyed the experience. She has been and remains my most ardent champion. I look forward to following her on many more adventures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

My dissertation focuses on measuring "voice" from unstructured data and understanding its implications for consumers, firms and regulators. I define the term "voice" broadly to include all types of discussion captured in unstructured data- consumer opinions as captured in consumer reviews and blogs, intra firm discussions and, public communications by firms and regulatory institutions. Unstructured data includes text, audio and video data. I am optimistic that my empirical research on "voice" can potentially answer substantive questions that are difficult to address with traditional data, in a rigorous modeling framework.

My approach combines econometric methods and natural language processing technologies applied to text data to examine the influences on and of "voice". In my first essay, I consider the *effect of regulatory intervention on consumer opinion*. I extend the extant work on topic modeling to measure a specific topic (in this instance the topic is "health"). First, I introduce "seed" words as priors to measure the discussion of health. Second, I model the intuition that the overall review characteristics (rating and review length) and reviewer characteristics (reviewing experience) affect the mix of topics discussed in a review. Finally, I incorporate the measure within a differences-in-difference framework to identify the effect of the regulatory intervention.

In my second essay, I examine the effect of introduction of disruptive services (Uber Technologies) on consumer experience for traditional services (licensed taxi services). In contrast to the first paper, this paper examines the *effect of a disruptive entrant on "voice"*. I focus on specific aspects (that have been frequently discussed in the business press) that potentially affect consumer experience for traditional taxi services. I jointly model the ratings and review text by integrating two extant models – Supervised LDA (McAuliffe and Blei, 2008) and Factorial LDA (Paul and Dredze, 2012) to examine consumer experience. I extend the topic modeling framework to measure the topics of interest that focus on these multiple aspects. Further, the standard topic model does not require topics to be unique.

Consequently, a post processing step is often required to cluster topics together. To circumvent this problem, the model incorporates prior probability distributions for seed words. The model also incorporates the valence of a topic, so that positive and negative discussions of a topic can be identified and evaluated separately. I proceed to identify attributes that explain the changes in consumer experience for traditional taxi services.

In my third paper, I explore *how a firm's policy might influence the style of reviewer "voice"* (not necessarily the content of "voice"). Using "stylometric" cues, I examine the writing style of authors on Amazon.com before and after being enrolled to the "Amazon Vine" program. Enrollment to the program makes the reviewer eligible to receive free product samples and the reviewer's user name is tagged with a special badge on the site. The substantive question here is whether enrollment leads to reviewers writing differently. I test this hypothesis in a predictive framework using support vector machines on a data from over 2 million reviews. An initial analysis indicates that writing styles do change from the pre to the post period. There are several additional questions that remain to be explored. In particular, I plan to investigate stylistic differences (a) between reviews for purchased products versus for products received for free amongst Vine members and (b) between reviews by Vine reviewers and non-Vine reviewers. Another line of inquiry involves decomposing the "Enrollment" effect into a reputation/status effect (the influence of the status badge) and a product sampling effect (the influence of receiving goods for free). Finally, investigating the temporal dynamics of style for these reviewers can help understand whether getting enrolled in the Vine program affects review readers' purchase decisions.

In a fourth paper, I examine a key assumption in "voice" topic modeling – the choice of the number of topics. While there are instances where the choice of the number of topics is a matter of fit (as I found in my first essay) or is driven by the researcher (as specified in the second essay), it is often the case that the optimal number of topics is not readily identifiable using standard fit criteria. In this paper, I examine this phenomenon extensively and suggest alternative criteria that mitigate the issue to some extent.

In each of these projects, I have explored and extended current modeling approaches / frameworks to address a set of closely related substantive and methodological questions related to measuring "voice". My goal has been to not only address substantive and methodological questions of interest to academics, but also address managerially relevant questions. I am eager to pursue this programmatic approach to further extend our understanding of "voice".

A broad area of both substantive and methodological interest to me is the measurement of emergent or nascent voice. Methodologically, computer science continues to make significant progress in terms of algorithms that can deal with big data and summarize large volumes of data. However, these algorithms generally uncover themes that are dominant in the data. As researchers in Marketing, our focus lies on specific aspects that we would like to learn more about from this high volume of data. Alternatively, we are more interested in the subtle and emerging aspects that may be discussed in relatively low volume rather than the obvious aspects that dominate most of the data. A simple word count would be quite inadequate as that requires the researcher to undertake very strong (deterministic) assumptions on the usage of words. There are multiple applications for a reliable measure of "voice" in marketing (in addition to customer satisfaction) such as quality, new product attributes and brand perceptions. I am also actively tracking emerging natural experiments that may influence "voice" and can offer insight on substantive questions. Consequently, I hope to work specifically on models that offer a rigorous probabilistic framework that enables researchers in marketing to rigorously measure "voice" and draw robust inferences.

CHAPTER 1

THE EFFECT OF CALORIE POSTING REGULATION ON CONSUMER OPINION: A FLEXIBLE LATENT DIRICHLET ALLOCATION MODEL WITH INFORMATIVE PRIORS

Section 1. Introduction

In the face of rising obesity, Mayor Bloomberg of New York City pushed a regulation in 2008 that required chain restaurants (those with 15 or more units nationwide) to display calories for every item on all menu boards and menus in a font that was at least as prominent as price. Two years later, the Affordable Health Care Act of 2010 mandated that restaurants with multiple locations prominently display calories for every item on all menus. The desired impact of both of these laws was to make it easier for consumers to choose healthier foods, as posting calorific information should make health more salient in the minds of consumers when eating out. Implementing this national regulation "has gotten extremely thorny", in the words of the Commissioner of the Food and Drug Administration. "There are very, very strong opinions and powerful voices both on the consumer and public health side, and on the industry side, and we have worked very hard to figure out what really makes sense" (Jalonick 2013).

Past research has shown us that regulations pertaining to health claims on food labels affect consumer search and consumer behavior in various ways (Roe, Levy and Derby 1999, Bollinger, Leslie and Sorenson 2011, and Downs et al. 2013). Unlike these papers, we focus our research on consumers' post-consumption opinions of the product. Our data are 761,962 reviews of 9,805 restaurants in New York City, posted on a leading restaurant review site[1] in an 8-year period from the website's inception in October 2004 to December 2012. Online reviews remain in the public domain for long time periods and can be leading indicators of future trends in consumption behavior. We are unaware of studies that estimate the impact of regulation changes on consumer opinion or word of mouth.

We propose an automated and scalable probabilistic model that summarizes this large volume of free, unsolicited, rich user-generated reviews into a few interpretable topics. These

[1] As per Alexa.com, an independent research firm, this was among the top 70 most popular websites in the United States in July 2014. Owing to the legal terms of service of this website, we are unable to reveal its identity.

topics can offer managerial and policy insights into how consumer opinion or the "voice of the consumer" (Griffin and Hauser 1993, Lee and Bradlow 2011) was influenced by the implementation of a calorie posting regulation in New York City. Traditional approaches to measure the effects of regulations, such as surveys and focus groups, might be expensive, time consuming, and potentially subject to recall biases and demand effects (Netzer et al. 2012). Unlike such approaches which rely on primary data collected over a short period of time, our data are available over several years. Therefore, relative to primary data collection, our approach is especially useful for studying the impact of temporally distant events (such as past regulations) by comparing periods before and after such events. A long time series also allows us to study both short term and long term effects of the focal event[2].

Based on our data and methods, we pose and answer the following managerially- and policy-relevant questions about online reviews of chain restaurants:

a) What were the major topics or attributes about chain restaurants that consumers discussed in online reviews before and after the mandatory calorie posting regulation was enforced?

b) Was health a topic of discussion before and after the regulation? If so, what proportion of this discussion was on health? Which topics were discussed to a larger extent relative to health?

c) Was health a topic of discussion for a few reviewers, or was the discussion widespread across several reviewers? Did this distribution change after the regulation?

d) Does the regulation have a differential impact on areas with less healthy populations?

The following managerial and regulatory insights can be obtained from our analysis. First, should there be an increase post-regulation in health as a topic of discussion by a large set of consumers, this can be seen as a measure of the regulation's success in making health more salient in the minds and voices of consumers. Second, textual content posted in online consumer reviews affects subsequent demand (Archak, Ghose and Ipeirotis 2011, Ghose, Ipeirotis and Li 2012). That is, we might expect greater discussion of health across a very large number of online reviews to be accompanied with greater consumption of healthier foods. Third, changes in patterns of consumer opinion can provide continuous, timely and free inputs into more traditional forms of marketing research. Increased discussion of health in online reviews can serve as a basis

[2] It is plausible that despite large sample sizes, user-generated textual content might also suffer from biases; indeed, no market research technique is perfect. As such we do not propose to replace traditional approaches, but instead to augment them with data that are available for free, in larger quantities, and over longer period of times.

for commissioning more costly investigations into changes in consumer buying behavior e.g. the patterns of substitutions from less healthy options. Fourth, how widespread health discussion is (i.e. variance in how deeply reviews discuss health as a topic, conditional on mean level of health topics across reviews) can provide insights in to consumer segments (e.g. a small segment of reviewers dominating the health discussion versus a large segment of reviewers discussing health albeit to a small extent). Such information can also serve as the basis of studies aimed at identifying individuals who might influence restaurant choices of the population, and at ascertaining the demographic correlates of those individuals who are most vocal about health. Fifth, policy makers are interested in understanding if the regulation is more successful among less healthy populations. African-Americans have lower life expectancy[3] and greater rates of obesity[4], hypertension (Fuchs 2011) and diabetes[5]. Studies have found correlations between these health outcomes and food habits. Our approach enables us to estimate the impact of the regulation on health topics in neighborhoods with greater concentration of African-American populations. Note that there is also widespread media discussion of health and in particular this regulation; we confine ourselves to the analysis of consumers' post-consumption data for the reasons outlined above.

We now briefly discuss our model and research design. Our model belongs to a class of probabilistic topic models termed Latent Dirichlet Allocation (LDA) models (Blei, Ng and Jordan 2003), which have been developed by computer science (specifically in the machine learning discipline) researchers to analyze words in large sets of original text in order to discover the themes or topics within. We summarize a large collection of reviews into a few representative latent topics (e.g. price, service, menu item, cuisine) and characterize these topics by a probability distribution over all words in reviews. For each word in a review, a topic is chosen, and conditional on the choice of topic, a word is chosen. This process continues until the review adequately represents the topics of interest of the writer. Each review is composed of a random mixture of several topics (e.g. a restaurant review could be simplistically represented as 20% price, 20% service, and 60% Mexican Cuisine). This process represents a probabilistic interpretation of the data generation process for the observed reviews. We use state-of-the-art

[3] Source: http://www.cdc.gov/nchs/data/hus/hus14.pdf#064
[4] Source: http://frac.org/initiatives/hunger-and-obesity/obesity-in-the-us/
[5] Source: http://minorityhealth.hhs.gov/omh/browse.aspx?lvl=4&lvlID=18

tools to select topics that are coherent for ease of managerial and policy interpretability. Estimation challenges arise because a) we do not observe the topics, but rather infer them from the data, b) the same word could belong to different topics necessitating a flexible modeling approach, and c) the large scale of the data (761,962 reviews) necessitates scalable estimation techniques.

Since the scope of the regulation was limited to chain restaurants, we analyze data from chain and standalone restaurants separately, such that standalone restaurants serve as a useful contrast and as a natural control group.[6] To isolate the causal effects of the regulation on consumer opinions in chain restaurants, we control for differences in characteristics between chain and standalone restaurants, for differences in reviews of all restaurants between the two time periods (before and after the regulation), and for geographical differences in topic proportions (via zip code dummies). We conduct several additional tests for robustness of causal inference, including a variant of a regression discontinuity analysis (Thistlethwaite and Campbell 1960, Hartmann, Nair and Narayanan 2011), wherein we constrain the time period of analysis to a few months before and after the implementation of the regulation to minimize the effect of potential time-varying confounds. In sum, we combine current methods from computer science with causal inference techniques.

Situating our work within the existing research literature, our paper is somewhat related to the literature on the economic impact of numeric characteristics of online reviews (Godes and Silva 2013; Godes and Mayzlin 2004). However, there has been relatively less research on extracting useful information from large masses of text of reviews. Decker and Trusov (2010) use text mining to estimate the relative effect of product attributes and brand names on product evaluation. Ghose et al. (2012) combine text mining with crowdsourcing methods to estimate hotel demand. Methodologically, our work is closer to three papers. Lee and Bradlow (2011) automatically extract details from each review in terms of phrases. Each phrase is then rendered into a word vector which records the frequency with which a word appears in the corresponding phrase. Phrases are clustered together according to their similarity, measured as the distance between the word vectors. Netzer et al. (2012) use a similar approach, with the difference that they define similarity between products based on their co-mention in the data. Tirunillai and

[6] We refer to restaurants with less than 15 units nationwide as "standalone" (as opposed to chain) for ease of understanding. As mentioned, such restaurants were outside the scope of the regulation.

Tellis (2014) apply the LDA model on textual data from consumer reviews from five markets to infer the latent dimensions of product quality (e.g. portability of mobile phones), to extract the valence of these dimensions, to understand how brands within each market are positioned on these dimensions, and to estimate how these dimensions and brand positions vary over time.

Our model generalizes previous approaches in marketing for extracting topics in two important ways. First, managers or policy makers might have informed priors about how consumer opinions might change due to specific events. For example, the enforcement of a new sales tax law might alter the level of discussion of "price" as a topic when consumers review the focal product online. On the other hand, a new regulation pertaining to minimum wages in the retailing industry might alter consumer opinion about service (if wage increases lead to service improvements), or price (if wage increases translate to price increases). In the papers discussed above, words or phrases are allocated to topics, and then topics are interpreted by the researcher. In contrast our approach allows the analyst to pre-specify constructs or topics of interest, and to then track changes in consumer opinion as it pertains to those topics. This is achieved by specifying an informative prior distribution of topics over the words in the vocabulary. This enables us to parsimoniously integrate managerial intuition and interest with information contained in thousands of reviews. Combining managerial intuition with statistical modeling has a long tradition in marketing and psychology, and has even been shown to improve model fit compared with purely statistical modeling (Blattberg and Hoch 1990, Yaniv and Hogarth 1993, Wierenga 2006).

Second, current research in marketing and computer science assumes that the distribution of topics within a document is independent of the author's decision of how many words to write. We allow topic distributions to vary with the length of the document, its valence, and its author's experience. Other than improved model performance, allowing topic distributions to vary by the length of the review, and other characteristics has substantive implications. Authors of reviews which are focused solely on health are more likely to lead the consumer opinion on health, and be more important than the general reviewing population for targeting. To the extent that shorter reviews are more likely to discuss one or very few topics, the length of a review might be an important summary statistic of user-generated content to consider in identifying such reviews and reviewers.

From a substantive standpoint, our identification strategy (comparing reviews of chain restaurants before and after the regulation to reviews of the control group of standalone restaurants) helps us use textual data to make causal inferences of the impact of the regulation on consumer post-consumption reviews. We complement recent research on temporal dynamics in the ratings and textual content of online reviews (Tirunillai and Tellis 2014, Godes and Silva 2013), by inferring how levels of discussion of various topics vary over time (and across locations) due to an exogenous event. Our work also complements the academic research on the effect of the calorie posting regulation on consumer behavior (Bollinger et al. 2011, Downs et al. 2013). Such research provides insights from survey and transactional data from a single chain of restaurants (e.g. Starbucks for Bollinger et al. 2011). Our data are from 9,805 restaurants including 77 unique chains, and we focus on post-consumption opinions.

The backdrop of the regulation is the concerns from rising obesity, so we can safely conjecture that there was a need for regulation since consumers are likely not as health-aware. Therefore, our expectation is that prior to regulation, the discussion of health topics in reviews is likely to be small. And the purpose of our paper is to see if there are emergent voices discussing health topics more extensively after the regulation. In the context of the overall obesity levels in this country, our ex-ante expectation consequently is that the uptick in health discussion will not be large. Indeed, we find that health is discussed only in a small proportion of reviews (less than 7%). This proportion increased for chain restaurants after the regulation, but not for standalone restaurants, suggesting that the regulation served to grow the salience of health among a small segment of health-conscious consumers. Given the overall trends of increasing obesity in the United States, even small post-regulation increases in health topic discussion in restaurant reviews might be cause for celebration and offers the potential for significant long-term implications (see section 4 for details). We also find that much of the increase in health topic discussion post regulation can be attributed to a small segment of new authors. The increase in health discussion is somewhat greater in African-American neighborhoods, perhaps indicative of greater effectiveness of the regulation among more obese populations.

Next we discuss the model specification. Section 3 then presents the data, and discusses specific estimation challenges. Section 4 presents the results from the model, and their

implications, and compares model performance with models which do not incorporate the unique features discussed above. Section 5 concludes.

Section 2. The Model

2.1 Model Specification

We start with some standard definitions and notation. A corpus is a concatenation of all documents in the dataset; in our case, the corpus comprises all reviews. Each document d (d=1,…,D) is composed of n_d words. The number of total word instances in the corpus is N (e.g. for a corpus of 10,000 documents with 100 words each, N is 1 million). The corpus is therefore defined by an N-dimensional vector $\boldsymbol{w} = \{w_{11}, w_{12}, w_{di}, .., w_{Dn_D}\}$, where w_{di} is the i^{th} word of document d. A vocabulary is the set of all unique words across all documents (e.g. all unique words in all reviews posted on a website). It is defined by V unique words, each denoted by v. Each word in the vocabulary belongs to each topic, such that topic ϕ_k (k=1,..,K) is a probability distribution over all v in $\{1,2,..,V\}$. Element ϕ_{kv} denotes the probability of word v given topic k. Document d is a mixture of the K topics. θ_d is a K-dimensional vector that represents the proportions of each topic in document d. For a 3-topic model for example, document d might be summarized as $\theta_d = [0.25 \quad 0.50 \quad 0.25]$.

We briefly describe our textual data to enable better understanding of the model. The mean length of all 761,962 reviews is 126.7 words (SD=109.6). Each sentence is split into its component words using the Natural Language Toolkit's Tokenizer (Bird 2009). After eliminating stop words ("a", "the" etc.) and words that occurred less than 5 times in the entire corpus (Griffiths and Steyvers 2004, Lu et al. 2011) the number of unique words in the corpus is 44,276. Although the calorie posting regulation was implemented over a few months, we assume July 1, 2008 as the "implementation date" for comparing pre- and post-regulation consumer opinion. Robustness of our results to this assumption is discussed in the next section.

The LDA model assumes a generative process by which the textual data in each document is generated. The first step for generating word i in document d is to draw topic proportion θ_d from a Dirichlet distribution with K-dimensional parameter vector α_d. The second step is to choose z_{di}, the topic assignment for the word i in document d. This is drawn

from a categorical distribution with parameter θ_d. This is a particularly convenient choice of distributions as the Dirichlet distribution is conjugate to the categorical distribution, i.e. the posterior distribution of θ_d is also Dirichlet. Given the topic assignment z_{di}, the word *i* in document *d* is drawn from a categorical distribution associated with the assigned topic. To exploit conjugacy, each topic distribution is also specified Dirichlet (i.e. $\phi_k \sim Dirichlet(\beta)$). This process is repeated for each word *i* in document *d*. It ignores the order of words within a document, i.e. LDA is a "bag-of-words" model (Eliashberg, Hui and Zhang 2007, Netzer et al. 2012)[7]. The generative process for document *d* can be summarized as follows.

a) $\theta_d \mid \alpha_d \sim Dirichlet(\alpha_d)$

b) $z_{di} \mid \theta_d \sim Categorical(\theta_d)$

c) $w_{di} \mid (z_{di} = k), \phi \sim Categorical(\phi_k)$

d) $\phi = [\phi_1, \phi_2, \phi_3, ..., \phi_k]$

e) $\phi_k \mid \beta \sim Dirichlet(\beta)$

The researcher does not observe the topics, the (probabilistic) membership of words in each topic, the distribution of topics for each document, or the choice of topic that led to a specific choice of a word. The central computational problem is to use the observed documents to infer these distributions, i.e. to uncover the hidden topic structure that generated the observed set of documents. The process described above defines a joint probability distribution over both the observed and hidden random variables. We use this joint distribution to compute the conditional (or posterior) distribution of the hidden variables given the observed documents and words.

The model described so far ignores characteristics of the document which might affect the distribution of topics. Longer documents might discuss more topics and therefore might have more evenly spread topic distributions. It is also plausible that authors writing for the first time discuss fewer topics than more experienced authors, suggesting that documents with similar

[7] Modeling word order is computationally intensive and therefore rare, even in computer science. Such models have usually been limited to incorporating bi-grams (word pairs) or tri-grams (a triplet of words). Given the computational burden posed by estimating the hyperparameters, we chose to retain the standard "bag-of-words" assumption.

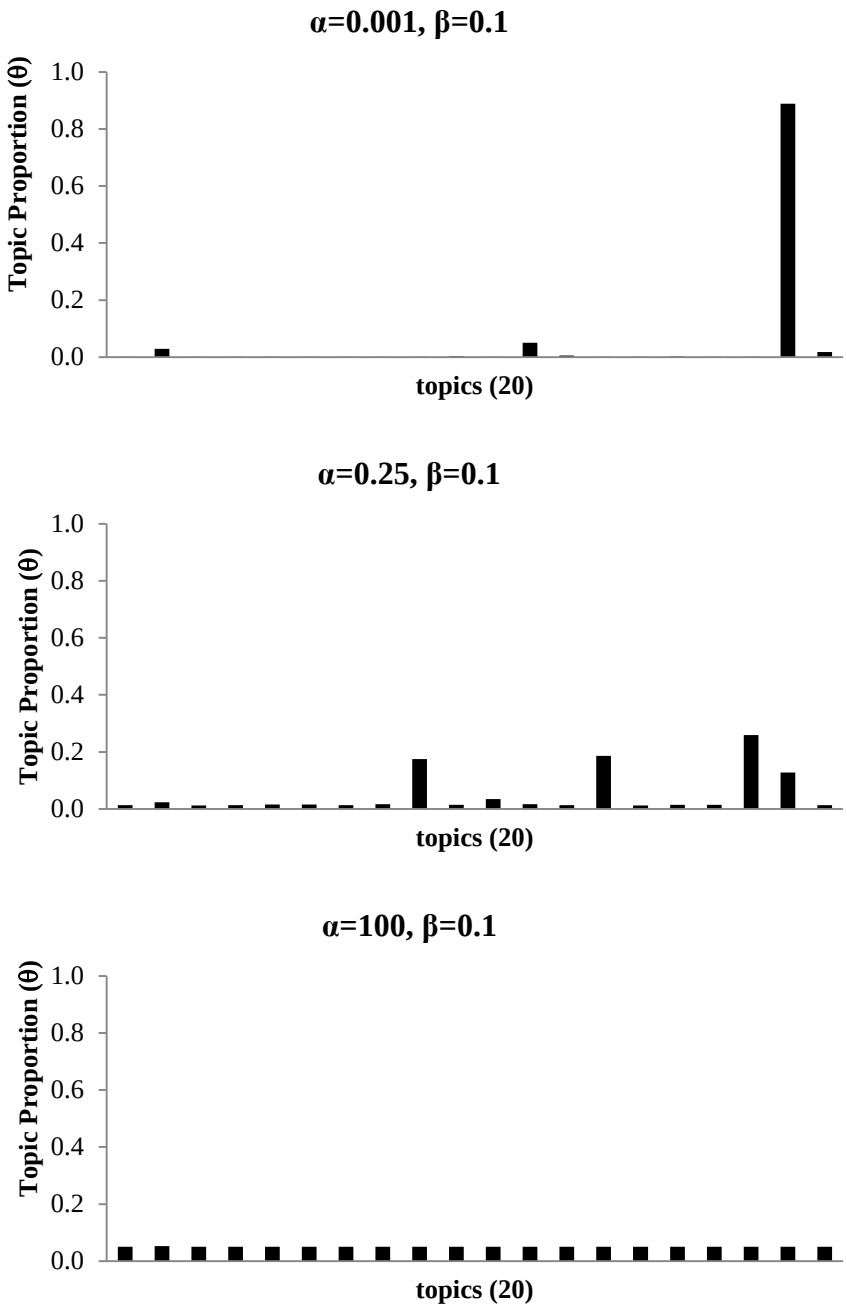
levels of author experience might have similar topic distributions. Online reviews tend to be disproportionately positive in valence, so it is possible that authors posting negative reviews elaborate on more topics to better justify the negative evaluation. We incorporate these intuitions into our model by allowing the hyperparameter α_d to vary with various observed characteristics of the document. More specifically, we allow α_{kd} for topic *k* in document *d* to depend on the document's length (measured by the number of words), its valence (as captured by a 5-point ordinal scaled numerical rating), and the past reviewing experience of the author (measured by the number of reviews in our dataset which were posted by the author of document *d*, before she posted document *d*). This leads to the following specification.

$$\alpha_{kd} = \exp(\lambda_{k0} + \lambda_{k1} n_d + \lambda_{k2} ex_d + \lambda_{k3} r_{1d} + \lambda_{k4} r_{2d} + \lambda_{k5} r_{3d} + \lambda_{k6} r_{4d}) \quad (1)$$

n_d is the number of words in document *d*; ex_d is the experience of its author; r_{1d}, r_{2d}, r_{3d} and r_{4d} are dummy variables which take the value 1 if the rating of document *d* is 1, 2, 3 and 4 respectively, and 0 otherwise. The exponential function preserves the positivity of the parameter. We allow the effects of metadata to vary across topics. To demonstrate how allowing for document specific hyperparameters affects the within-topic distribution, we estimated the model on online reviews of all chain restaurants in our dataset, assuming that α is known to the researcher, and studied topic proportions for different values of α. We estimated three models differing only in the value of α_{kd} (0.001, 0.25 and 100). For each model we assumed α_{kd} to be the same for all topics and documents. For all three models we assumed 20 topics, and set the hyperparameter β =0.1. Topic proportions for the three models (corresponding to three values of α_{kd}) are presented in Figure 1.1. Larger values of α are clearly associated with more evenly spread topic distributions. If longer documents involve discussion of more topics entailing a more evenly spread topic distribution, we expect λ_{k1} to be positive. Later in the paper, we demonstrate that this specification improves model fit, over the traditional specification which assumes the same α for all documents.[8]

[8] Topic proportions could vary due to unobserved restaurant characteristics. In another specification, we allowed λ_{k0} to vary across restaurants by estimating a restaurant specific intercept. Our results remain unchanged, suggesting that the observed metadata are sufficient to account for heterogeneity in topic proportions.

Figure 1.1



Distribution of topic proportions for a 20-topic model for various values of α

We now turn our attention to the central question of this research: how did the proportion of health topic discussion in online reviews change due to the regulation? Since the regulation was implemented for chain restaurants only, standalone restaurants fall outside of our domain of

substantive interest, except as a control group for the purpose of inferring the causal effect of the regulation on chain restaurants (which form the treatment group). For us to draw causal inference, it is imperative that we compare changes (before versus after the regulation) in the *same* construct across the groups. This construct is the health topic proportion estimated on the treatment group—i.e. chain restaurants. We first estimate changes in health topic proportions within the treatment group, and then compare this to changes in the *exact same topic* within the control group. To identify the causal effect of the regulation on the proportion of health discussion, we implement a "difference-in-differences" methodology. We calculate the causal effect of a treatment (i.e. the regulation) on the outcome variable (proportion of health discussion) by comparing the average change in the outcome variable for the treatment group (chain restaurants) to the average change for the control group (standalone restaurants). We regress the outcome variable on two main effects (the effect of belonging to the treatment group on the outcome, and the effect of the treatment on the outcome), the interaction of these two effects, and several control variables, as follows:

$$\theta_{kd} = \upsilon_{0k} + \upsilon_{1k} Chain_d + \upsilon_{2k} Post_d + \upsilon_{3k} Chain_d\_Post_d + \upsilon_{4k} ZipCode_d + \upsilon_{ad} + \upsilon_{fd} + \varepsilon_{kd} \qquad (2)$$

θ_{kd} is the proportion of topic *k* in document *d* (the outcome variable), $Chain_d$ is the dummy variable which accounts for the effect of belonging to the treatment group. It controls for unobserved factors which might affect topic proportions of chain and standalone restaurants differently. $Post_d$ is the dummy variable which accounts for the effect of the treatment. It is possible that at the time of the implementation of the regulation, there were unobserved events which affected topic proportions of all restaurants (including chain restaurants) in New York City. The main effect of $Post_d$ controls for how topic proportions for all restaurants changed after the implementation of the regulation. The coefficient of the interaction term ($Chain_d\_Post_d$) captures the crucial effect of the treatment on the treatment group. Further, we control for spatial variation (across locations in New York City), and temporal variation (over the duration of the data), in health topic discussion in reviews of all restaurants in our data. To control for spatial variation, we note that our reviews represent restaurants from 134 zip codes. $ZipCode_d$ contains 133 dummy variables, all of which are zero, except the variable corresponding to the zip code of the focal restaurant, which takes the value 1. The random effect

of author *a* of document *d* is captured by υ_{ad} [9] and the random effect of restaurant f is captured by υ_{fd}. Both effects are assumed normally distributed with zero mean.

We first estimate the model on reviews posted over a 4-year period (i.e. 16 quarters) starting July 1, 2006 (we show later in the paper that our results are consistent across different choices of time periods). This period was chosen so that the duration of time (in which reviews were posted) before and after July 1, 2008 (date of regulation implementation), is the same. 16 quarters correspond with 16 dummy variables, which form a vector of 16 elements, say Q_d. The variable in Q_d which corresponds to the quarter in which document *d* was posted takes the value 1. All other variables in this vector take the value 0. We note that the 8 elements of Q_d corresponding to the post-regulation period sum to the variable $Post_d$. Also, the 8 elements of this vector corresponding to the pre-regulation period sum to (1-$Post_d$). To avoid collinearity issues, we define $quarterID_d$ as a subset of Q_d formed by removing two dummy variables, one from the pre-regulation period, and one from the post-regulation period. Error terms are assumed IID and normally distributed. All parameters are topic-specific.

2.2 Model Estimation

We first describe how we estimate each parameter, and then our method of seeding. We estimate the hyperparameters α_d and β_d, the document level topic proportions θ_d, the vector of word level assignments of topics z_d, the topic level parameter ϕ_k and the parameters associated with the regression model in Equation 2. Assuming documents are conditionally independent and identically distributed, the likelihood of the data conditional on the hyperparameters is calculated as follows:

$$(L \mid \alpha, \beta) = \prod_{d=1}^{D} \int_{\phi} \int_{\theta} p(\boldsymbol{\theta}_d \mid \alpha_d) \times p(\phi \mid \beta) \prod_{i=1}^{n_d} \sum_{k=1}^{K} \prod_{v=1}^{V} [\phi_{k,v} \times \theta_d^k]^{I(w_{id}=v)} d\phi d\theta \quad (3)$$

where *I* is the indicator function, and α_d is a *K*-dimensional vector with element $\exp(x_d^t \lambda_k)$. We face two estimation challenges: this function does not have a closed-formed

[9] There are 1.5 reviews per author on average, which restricts our ability to identify author specific effects. To deal with this issue, we assume the same random effect for authors who posted the same number of reviews in our data.

analytical solution due to the product term involving $\phi_{k,v}$ and θ_d^k (Dickey 1983), and the dimensionality of our parameter space is very high (a common feature of problems associated with "big data"). The dimensionality problem is owing to the large number of unique words in the corpus (*V*), the potentially large number of topics and the large number of documents (note that θ_d is document specific). Following the computer science literature (Griffiths and Steyvers 2004), instead of estimating ϕ_k or θ_d as parameters we first estimate the posterior distribution of the assignment of words to topics, $P(z \mid w)$ based on the equation $P(z \mid w) = P(z, w) / \sum_z P(z, w)$.

The numerator of the right hand side of this equation can be factorized and simplified as $P(z, w) = P(w \mid z)P(z)$. We now turn our attention to $P(w \mid z)$ and $P(z)$. Given the conjugacy between the distribution of observing word *w* given topic *k* (assumed to be a Categorical Distribution as described in section 2.1) and the Dirichlet prior ($\phi_k \mid \beta \sim Dirichlet(\beta), \forall k$), the posterior distribution of $P(w \mid z)$ is as follows (Griffiths and Steyvers 2004):

$$P(w \mid z) = \left[\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right]^K \prod_{k=1}^{K} \left[\frac{\prod_{v=1}^{V} \Gamma(\beta + n_{kv})}{\Gamma(V\beta + n_k)} \right] \tag{4}$$

$\Gamma(.)$ is the standard gamma function. Here, n_{kv} is the number of times the word *v* in the vocabulary is assigned to topic *k* in the corpus; n_k is the number of words in the corpus which are assigned to topic *k*. Similarly, the conjugacy between the topic assigned to each word in a document (assumed to be a Categorical Distribution) and the Dirichlet prior ($\theta_d \mid \alpha_d \sim Dirichlet(\exp(x_d^t \lambda_k))$) yields document specific topic assignments. These topic assignments are conditionally independent across documents and can be multiplied to yield:

$$P(z) = \prod_{d=1}^{D} \left[\frac{\Gamma(\Sigma_k \exp(x_d^t \lambda_k))}{\Pi_k (\Gamma(\exp(x_d^t \lambda_k)))} \right] \left[\frac{\prod_{k=1}^{K} \Gamma(\exp(x_d^t \lambda_k) + n_{kd})}{\Gamma(\Sigma_k \exp(x_d^t \lambda_k) + n_d)} \right] \tag{5}$$

n_{kd} is the number of words in document *d* assigned to topic *k* and n_d is the length in words of document *d*.

Note that while $P(z,w)$ may be factored and computed as described above, $\sum_z P(z,w)$ cannot be computed directly because it does not factorize and involves K^N terms, which is again computationally challenging. We thus adopt an MCMC approach which relies on Gibbs sampling of the latent topic assignment variable z (Griffiths and Steyvers 2004).[10] The full conditional distribution of z is free of ϕ_k and θ_d, enabling us to estimate ϕ_k and θ_d by averaging the means of the posterior Dirichlet distributions across iterations from a single MCMC chain[11]. The Gibb's sampling scheme for (2) can be derived from (4) and (5) as follows:

$$P(z_i = j \mid z_{-i}, w)_d \propto \frac{n^w{}_{j,-i} + \beta}{V\beta + n_j - 1} \frac{n^d_{j,-i} + \exp(x_d^t \lambda_j)}{\Sigma_k \exp(x_d^t \lambda_k) + n^d - 1} \tag{6}$$

In each Gibbs iteration, the probability of assigning topic *j* to the word *w* in position *i* in document *d* is proportional to the product of two ratios. The first ratio is the number of times word *w* was assigned topic *j* as a proportion of the total words assigned to topic *j* (adjusted for smoothing) in the entire corpus. The second ratio is the number of words in document *d* assigned to topic *k* as a proportion of the total number of words in the document *d* (adjusted for smoothing). Any topic assignment is thus a function of both the corpus and the document.

We now describe how the vector λ_k, which captures the effects of metadata, is estimated. λ_k is an M+1 dimensioned vector of parameters specific to topic *k*, with prior distribution $N(0, \sigma^2 I)$[12]. Consequently, the joint probability of *z* and λ can be written as follows:

$$P(\mathbf{z}, \lambda) = \prod_{d=1}^{D} \left[\frac{\Gamma(\Sigma_k \exp(x_d^t \lambda_k))}{\Pi_k (\Gamma(\exp(x_d^t \lambda_k)))} \right] \left[\frac{\prod_{k=1}^{K} \Gamma(\exp(x_d^t \lambda_k) + n_{kd})}{\Gamma(\Sigma_k \exp(x_d^t \lambda_k) + n_d)} \right] \prod_{k=1}^{K} \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\lambda_{k,m}}{2\sigma^2} \right) \tag{7}$$

---

[10] Variational inference (VI) methods are also commonly employed in computer science and statistics for large-scale problems with intractable integrals. Whereas Monte Carlo methods provide numerical approximations of the exact posterior by sampling, VI methods provide a locally optimal but precise analytical solution to an approximation of the posterior. We estimated the model using a VI method and obtained almost identical results with comparable computational speed. We chose Monte Carlo methods since they are more common in the marketing literature.

[11] For example, $P(\phi_k \mid z, w) = Dirichlet(\beta + n_k^1, ..., \beta + n_k^v)$ and the estimated mean vector of this distribution from a single MCMC iteration is $(\frac{\beta + n_k^1}{V\beta + n_k}, ..., \frac{\beta + n_k^v}{V\beta + n_k})$.

[12] Directionally results are not sensitive to the choice of variance for this prior. We used different variance values (0.5, 5, 25), but our findings remain the same.

We maximize the likelihood of the topic assignments for each word in the corpus with respect to the parameters $\lambda_{k,m}$ using the Stochastic Expectation Maximization (Mimno and McCallum 2008). Consequently, the estimation algorithm consists of alternating between the MCMC iterations using (6) and the stochastic EM step where (7) is maximized with respect to the parameters $\lambda_{k,m}$. However, Mimno and McCallum (2008) assume that the hyperparameter β is known, presumably for computational ease. It seems implausible in our context of regulation change that we would know the distribution of words over topics *a priori*. Therefore we explicitly estimate the hyperparameter β, in addition to λ, to address this issue. Details for the estimation of this parameter appear in Online Appendix 1.

Finally, we estimate Equation 2 within each iteration of the MCMC sampler. In each iteration, we first obtain an estimate of the posterior mean of θ_{kd}, and estimate the coefficients by regressing this iteration-specific estimate on the covariates. Posterior estimates of θ_{kd} vary across iterations depending on their inherent variability. Low (high) posterior variance of θ_{kd} would entail low (high) variability across iterations. In this way we naturally accommodate the variability of θ_{kd} in our analysis and inference.

The estimation algorithm is implemented in Mallet[13] and was modified to accommodate seeding, to estimate β and to generate output relevant for our analysis. The MCMC chain ran for 15,000 iterations, with the first 1500 iterations for "burn-in". We then estimate all hyperparameters (λ and β) every 100 iterations. The last 5,000 iterations (using a sampling lag of 10) yielded 500 samples that were used to compute the moments of the posterior parameter distributions.

2.2.1 *Seeding*. A unique feature of our model (in contrast to extant models for analyzing textual data in marketing) is that it permits the researcher to specify words to belong to a topic, such that the "seeded" topic becomes a topic of central interest. The posterior parameter distributions can then be used to infer changes to the distribution of this topic across documents and over time. In our analysis, we seed a topic—which we simply label "health"—by allowing the prior distribution ϕ_k of this topic over the vocabulary to contain the following words or "seeds" with

[13] http://people.cs.umass.edu/~mccallum/mallet/

high probability: calorie, calories, fat, diet, health, healthy, light, fit, cardio, lean and protein. This list is based on a review of Section 81.50 of the New York City Health Code that articulates the regulation. The words "calorie", "calories" and "health" are the most frequently occurring health related words in the policy document. Words such as "light", "fit", "lean" and "protein" appear related to health, and occurred with high frequency in our corpus.

We now discuss how we initialize our dataset and our method of incorporating seeds. We randomly assign each word in each document to one of *K* topics. As an example, for a document with say 20 words, the first word is assigned to topic 1, the second to topic 5, the third to topic 189 and so on. Note that in this stage of the estimation process, topics don't actually exist – we are simply labeling words to topic indices. Using this initial allocation of words to topics, we make an initial estimate of ϕ_{kv}, the probability of word *v* given topic *k*. We count how many times each word in the vocabulary was assigned to topic *k*. For example the word "salad" maybe assigned to topic 0 10 times, to topic 5 87 times and so on. We then compute n_k, the total number of times any word was assigned to topic *k*. Suppose for illustrative purposes that n_k for topic 0 is 1000. Dividing the number of times each word was assigned to topic 0 by n_k yields a probability distribution over the vocabulary. For example, if "salad" is assigned to topic 0 10 times (i.e. n_{kv} =10), the probability of "salad" in topic 0 is 0.01 (n_{kv}/n_k =10/1000).

To incorporate seeds, we randomly choose a topic (as topics are exchangeable this does not result in any loss of generality). For this "seeded" topic, we then increment the counts of n_{kv} for each seed word *v*, thus increasing the prior probability of each seed word *v* in that topic. These incremental counts are referred to as pseudocounts. Intuitively, we are simply saying that these words were a priori assigned to this topic more often than a random initialization would indicate. As the Gibbs iterations proceed, the counts n_{kv} and n_k are updated, such that they could overwhelm the pseudocounts (i.e., their posterior estimates might differ considerably from the prior values)[14]. After some experimentation, we chose 5 as the value of this pseudocount in our

[14] For example, we included the word "health" as a seed word for the health topic because it had appeared in the calorie posting regulation document and was also consistent with our intuition. However, "health" is not in the top 10 most probable words in the posterior distribution of the health topic. "health" features in another topic with the following top 10 words: health, department, properly, unclean, dirty, roach, grade, equipment, nyc, contact. This

application. This choice is sufficiently flexible to allow for the possibility that the posterior estimate of ϕ_k will contain a) seeds with low probability, and b) other words with high probability. Given the large volume of data, this choice of pseudocounts does not affect the results. To verify this, we included a low frequency health-related word "cardio" as a seed. "Cardio" receives a low posterior probability assignment in the health topic and is not in the top 20 words used to describe the topic. In Section 4.2 of the paper, we discuss the robustness of our results to the choice of seeds.

Section 3: Analysis and Evaluation

We start by reporting the top 10 words across all reviews for chain restaurants in our data. In Figure 1.2 A, we present their ranking, based on the frequency with which they appear in each of the four groups of interest: chain restaurants pre-regulation, chain restaurants post-regulation, standalone restaurants pre-regulation and standalone restaurants post-regulation. For example, the word "salad" is the 8th most frequently appearing word in the reviews of chain restaurants posted before the regulation. We find the top 10 words are dominated by popular menu items (burger, fries), product attributes ($ possibly denotes price) and words connoting valence (good, love). None of these words explicitly reference health. Given the general view that health is not a very important consideration when eating out, this is not surprising. Although the words fries and steak (possibly connoting high calorie foods) rank higher (based on frequency of occurrence) in reviews of chain restaurants than in reviews of standalone restaurants, so does the word "salad," which is perhaps a healthier option. We do not discern any trends in frequency changes after the regulation from this figure, possibly because it pertains to just 10 out of 44,276 unique words in our corpus. Extending this analysis to 100 words did not reveal any other substantively useful insights.[15]

Next we present more granular time series plot by month of the top 10 most frequently occurring words in our data, as well as key words denoting health such as calorie, calories, fat

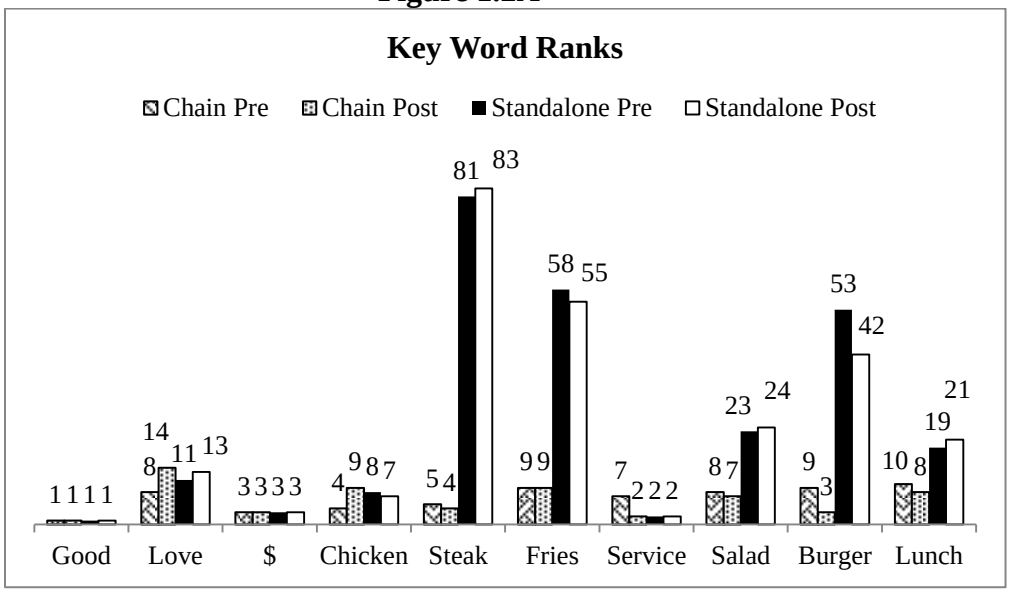
---

topic relates to health safety and hygiene and is unrelated to the calorie content of food. The data was able to overwhelm the prior and assign "health" to another topic.

[15] Since the regulation pertains to posting of calorific information, we investigated the occurrence of the words calorie and calories in our data. These words represent 0.07% of all words in reviews of chain restaurants before the regulation, and 0.10% of all chain restaurant reviews after the regulation. However, they represent just 0.01% of all reviews of standalone restaurants, both before and after the regulation. In other words, calories were discussed more in chain restaurants than in other restaurants before the regulation, and this trend intensified after the regulation.

and nutrition. We first report raw frequency of occurrence of these words in reviews of chain restaurants (Figure 1.2B). To enable better comparison across months, we also present the frequencies of occurrence for each month divided by the number of reviews of chain restaurants posted in that month (Figure 1.2C).

Figure 1.2A



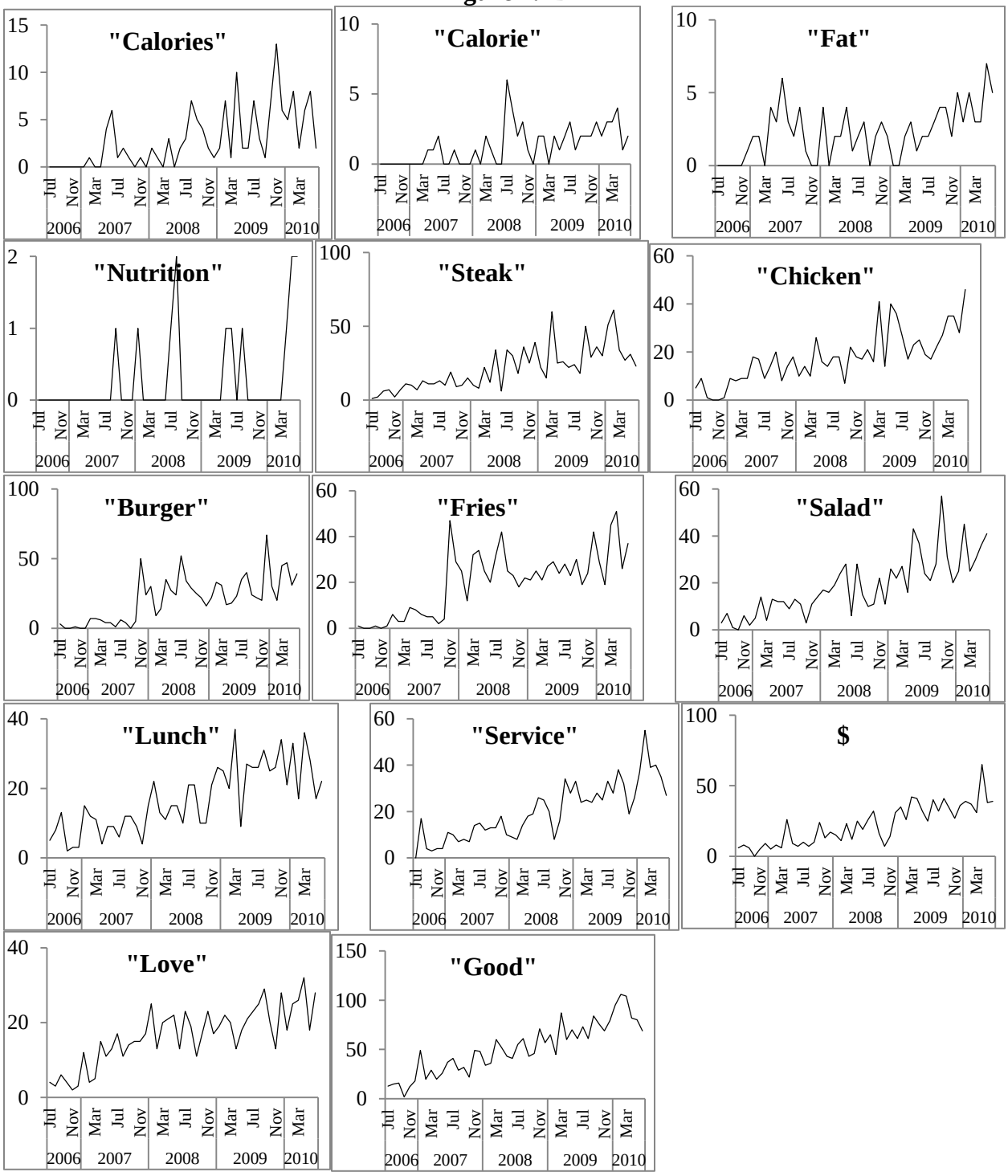
Most Frequently Occurring Words by Restaurant Type and Time Period

Plots of raw frequency counts reveal that most words occurred with very low frequency at the beginning of the data period, and showed an increasing trend over the entire data period. This correlates highly with the number of reviews posted. Plots of frequencies per review are more insightful. The words "calories" and "calorie" occurred with greater frequency (per review) post regulation than earlier, providing model free evidence of the success of the calorie posting regulation. On the other hand, words potentially connoting unhealthiness such as "fat", "burger" and "fries" seem to occur with lower frequency (per review) post regulation.

Although this analysis is useful to obtain a preliminary sense of the data, it cannot be used to draw any meaningful or robust substantive inferences in changes of consumer opinion due to the regulation. Since our objective is to infer topics of discussion from the data, analyses pertaining to counting specific words in the corpus, and how these frequencies vary over time are not helpful either. Except for the "health" topic, it is *a priori* unclear which words or topics to look for in the corpus. Second, even if a reliable list of topics were available, any choice of

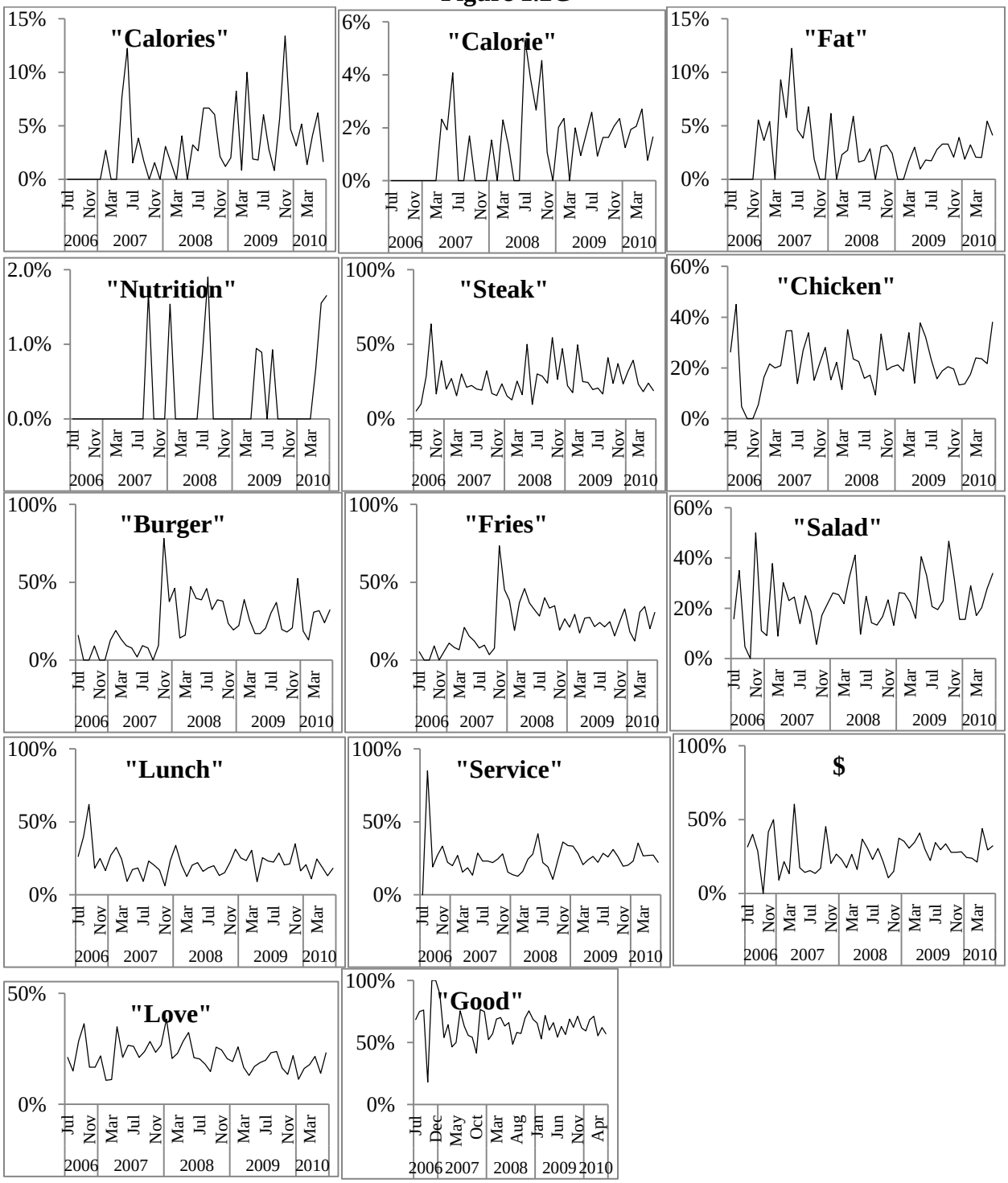
words for measuring the level of discussion of specific topics would be subjective; results pertaining to levels of topic discussions and their changes are sensitive to such choices. LDA offers a data-based, replicable, objective and principled methodology of inferring topics from text corpuses.

Figure 1.2B



Monthly Frequency of Occurrence of Words in Reviews of Chain Restaurants

Figure 1.2C



Monthly Frequency of Occurrence of Words in Reviews of Chain Restaurants Per Review,

A major challenge in all topic models is the interpretability of estimated topics. Models with large numbers of topics typically fit the data better and are able to support finer-grained distinctions in the text. However, some topics are more interpretable than others in the judgment of domain experts; furthermore the number of less interpretable topics often increases with the number of topics (Mimno et al. 2011). Measures of model performance such as out-of-sample fit, although commonly employed in marketing, correlate poorly with human judgments of topic interpretability (Chang et al. 2009). This has led to increased interest among computer scientists in developing automated metrics which are better able to predict topic interpretability. A useful insight from this research is that if a topic is highly interpretable (to humans), pairs of words which are associated with this topic with a high probability should frequently co-occur in several documents of the corpus. For example, a topic in which the words "healthy" and "vegetables" are highly probable is likely to be more interpretable or "coherent" if both of these words occur in several restaurant reviews. Mimno et al. (2011) provide evidence for this result, and use it to develop a "topic coherence" metric C_k for each topic. Topics scoring higher on this metric are more interpretable by human judges. It is defined as follows.

$$C_k = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^k, v_l^k) + 1}{D(v_l^k)} \tag{8}$$

where $V^k = (v_1^k, ..., v_M^k)$ is the list of M most probable words in topic k, $D(v)$ is the number of documents in which the word v appears and $D(v, v')$ is the number of documents which contain at least one occurrence each of both v and v'.

We now discuss how we choose the number of topics (K) and label each of them. Several statistical approaches exist for this purpose. Similar to cluster analysis, we maximize the dissimilarity between topics (Deveaud et al. 2012, Cao et al. 2009) by computing a distance between every pair of topics where each topic is a probability distribution over the vocabulary. We employ the Jensen-Shannon statistic (Lin 1991, Steyvers and Griffiths 2007) which is similar to the Kullback-Leibler divergence statistic (Kullback and Leibler 1951), except that it is symmetric (i.e. the order of distributions does not matter) and always takes finite values; these are both desirable properties. On estimating our model for various values of K, we found that this statistic is maximum at K=200. All results therefore pertain to 200-topic models. Not all topics are of substantive interest, and thus we follow the computer science literature and restrict

substantive inferences to a few coherent topics only (Mimno et al. 2011, AlSumait et al. 2009). Specifically, we present 20 topics in Table 1.1 A: the seeded "health" topic discussed earlier, and 19 topics with greatest values of the topic coherence metric (henceforth referred to as the top 19 topics).[16] Coherence scores of all other topics are available from the authors. Following convention in this literature, each topic is represented by listing the most probable words in the topics (Chang et al. 2009, Blei et al. 2003). We extend this principle to label topic *k* in terms of the two distinct words which have greatest posterior probability of belonging to that topic (as per ϕ_k). Although other words associated with the topic are likely meaningful, we choose this method for its objectivity, conciseness and because it does not require human intervention.

Although we use the Jensen-Shannon statistic to decide the number of topics, this in itself does not guarantee that all inferred topics are managerially relevant. Our topic of interest is quite focused—i.e. contains very few substantively relevant words. Seeding enables us to study the issue of interest in a focused and practical, yet statistically robust manner. Therefore, seeding is a very useful tool to investigate potentially small but emerging trends in the data, and allows managers and regulators to measure topics of special interest to them.

Section 4. Results and Implications

4.1 Results

We first discuss the major topics of discussion in our data. We then discuss the relative importance of the health topic. Next, we discuss how the discussion of the health topic is distributed across reviews, with an eye to inferring segments of the reviewing population which are more vocal about health. We draw causal inferences about how the relative importance of health and other topics changed over time due to the calorie posting regulation. Finally we discuss how review and author characteristics affect topic proportions.

16 In further analysis to test robustness of topics, we measured how far away the most probable words of the topics are from uniform distributions. The closer a topic's top words follow a uniform distribution, the less likely that the topic is informative. Empirically we expect the Zipf's law to apply; most of the probability mass in each topic is allocated to a few words. Employing this measure did not change the results in the paper.

Table 1.1A: Major Topics, Associated Words and Topic Proportions

ID	Name	Coherence Score	Top 10 words in decreasing order of posterior probability of belonging to the topic	Mean Topic Proportion in Chains
1	"Health"	-564.96	calories,calorie,fat,menu,healthy,count,low, counts,muscle,diet	0.41%
2	juice:orange	-296.324	juice,orange,cheap,plastic,buy,business, buying,cups,freshly,buckets	0.21%
3	steak:potatoes	-307.567	steak,potatoes,filet,good,spinach,cooked, mashed,dessert,sides,salad	2.20%
4	artist:content	-331.456	artist,content,kevin,trumps,canada,puerto, sexual,art,silent,rico	0.04%
5	account:free	-341.24	account,free,cosi,page,wireless,wifi,location, access,sns,service	0.09%
6	steak:wolfgangs	-342.367	steak,wolfgangs,bacon,lugers,peter, steakhouse,porterhouse,spinach,sides, creamed	1.00%
7	didnt:back	-350.135	didnt,back,wanted,ordered,wasnt,place,asked ,walked,time,friend	3.62%
8	mini:home	-353.3	mini,home,nanny,delta,kelly,doughnuts,qs, american,daddys,wins	0.04%
9	nom:bubba	-363.044	nom,bubba,cola,sq,coca,gump,sarah,elite, $2500,yada	0.06%
10	capital:grille	-365.492	capital,grille,week,cake,chocolate,dessert,cre me,sirloin,chowder,kona	0.29%
11	place:good	-371.465	place,good,youre,people,eat,theyre,make, time,find,eating	4.88%
12	bbq:chicken	-373.028	bbq,chicken,hawaiian,salad,ll;,rice,hawaii, spam,katsu,macaroni	0.35%
13	sandwich:potbelly	-380.581	sandwich,potbelly,sandwiches,bread,peppers ,hot,wreck,subway,italian,turkey	1.02%
14	table:waiter	-381.833	table,waiter,service,server,waitress,seated, drinks,ordered,meal,check	1.78%
15	pizza:dominos	-385.368	pizza,dominos,papa,johns,order,delivery, sauce,crust,slice,pizzas	0.64%
16	time:back	-385.705	time,back,make,give,ill,experience,eat,meal, made,long	4.20%
17	chipotle:burrito	-387.749	chipotle,burrito,rice,beans,salsa,chicken, bowl,cream,sour,guacamole	1.02%
18	stix:chinese	-389.72	stix,chinese,owned,nebraska,lincoln,mi, suppose,blimpie,fck,hours	0.06%
19	games:game	-390.189	games,game,fun,play,tickets,dave,busters, arcade,kids,playing	0.59%
20	burger:fries	-390.897	burger,fries,guys,burgers,toppings,free, cheeseburger,cajun,regular,joint	2.10%

In Table 1.1A, we present the top 10 words in decreasing order of posterior probability of being in each topic (for the health topic and the top 19 topics), as inferred from the analysis of all reviews of chain restaurants. These topics perform very well on well-established coherence metrics.[17]

Each topic is labeled by concatenating the two most probable words in the topic. First, we find that a substantial number of topics are focused on specific menu items (e.g. steak:potatoes, bbq:chicken, burger:fries). Second, several topics are focused on specific restaurant brands: Wolfgang (topic 6), Potbelly (topic 13) and Dominos (topic 15). Third, different aspects of service are captured across topics. Topic 5 captures service aspects unrelated to food (e.g. the words wireless, wifi, location, access). Topic 14 alludes to non-food related restaurant services (e.g. waitress, seated, check, server). Lastly, other than the first topic that we seeded with health related words, there are topics that might be important in understanding changes in consumer opinion due to the calorie posting regulation. Greater discussion of the topics "steak:potatoes" and "burger:fries" after the regulation might be an early signal that consumers think and write more about high calorie food than they did prior to the regulation.

Next we present the top 19 topics based on the means (across reviews of chain restaurants) of the posterior mean of the topic proportions θ_d. Since there are 200 topics, in the absence of any information we might expect the proportion of each topic to be about 0.5%. Service related topics (e.g. place:good, time:back) and topics associated with American staple fast-foods such as burgers, fries and steak get discussed to a greater extent than the average topic (see Table 1.1B). The seeded health topic is discussed at about an average level. We note that several topics are common across the two tables, suggesting that more interpretable topics are discussed to a larger extent.

To understand how widespread the discussion of health is, we compute the proportion of chain restaurant reviews for which the health topic proportion is greater than the baseline topic proportion of 0.5%[18]. We find that just 6.8% of such reviews contain the "health" topic to an

[17] To improve interpretability, it might be tempting to combine topics which appear similar. This can better be achieved by estimating models with fewer topics than by manually combining topics post estimation, since manual combinations might be subjective. However, such models would offer poorer fit. We follow the standard approach of drawing substantive inferences from the best fitting model.

[18] To aid understanding, we present an example of a review with a high estimate of the proportion of the health topic: "For anyone who is living a healthy lifestyle you need to come and sample Muscle Maker you will be back.

extent greater than 0.5%. In as many as 63% of all reviews, the proportion of this topic is less than 0.05%. This is a general pattern in the data; the discussion of a specific topic is skewed such that a small proportion of all reviews account for most of the discussion.

Table 1.1B: Major Topics, Associated Words and Topic Proportions (based on topic proportions)

ID	Name	Mean Topic Proportion	Top 10 words in decreasing order of posterior probability of belonging to the topic
1	"Health"	0.41%	calories,calorie,fat,menu,healthy,count,low,counts,muscle,diet
2	place:good	4.88%	place,good,youre,people,eat,theyre,make,time,find,eating
3	time:back	4.20%	time,back,make,give,ill,experience,eat,meal,made,long
4	didnt:back	3.62%	didnt,back,wanted,ordered,wasnt,place,asked,walked,time,friend
5	taste:hot	2.30%	taste,hot,flavor,fresh,meat,cheese,made,side,sauce,delicious
6	steak:potatoes	2.20%	steak,potatoes,filet,good,spinach,cooked,mashed,dessert,sides,salad
7	burger:fries	2.10%	burger,fries,guys,burgers,toppings,free,cheeseburger,cajun,regular,joint
8	menu:options	1.85%	menu,options,tasty,choose,option,great,pretty,good,choice,side
9	place:service	1.85%	place,service,time,worst,bad,eat,horrible,terrible,awful,money
10	good:service	1.80%	good,service,place,pretty,bad,decent,bit,wasnt,average,slow
11	table:waiter	1.78%	table,waiter,service,server,waitress,seated,drinks,ordered,meal,check
12	chain:place	1.68%	chain,place,prices,nyc,places,good,quality,youre,decent,fast
13	great:good	1.60%	great,good,place,service,recommend,amazing,worth,highly,price,expensive
14	taste:good	1.60%	taste,good,meat,bland,tasted,flavor,didnt,wasnt,dry,ordered
15	price:worth	1.59%	price,worth,pay,meal,expensive,drink,cost,money,dollars,$10
16	great:place	1.54%	great,place,delicious,time,love,awesome,amazing,back,eat,perfect
17	great:service	1.53%	great,service,good,nice,friendly,staff,experience,excellent,attentive,nyc
18	lunch:line	1.52%	lunch,line,long,time,lines,rush,location,busy,order,wait
19	chipotle:burrito	1.47%	chipotle,burrito,burritos,qdoba,mexican,chicken,chips,tacos,guacamole,bowl
20	good:pretty	1.47%	good,pretty,nice,place,bad,service,theyre,bit,tasty,quick

As examples, burger:fries and steak:potatoes are discussed in just 17.4% and 19.4% of all reviews of chain restaurants (to an extent of at least 0.5%).

The staff is helpful and friendly. The Arizona Rocky Balboa and Cajun Chicken with whole wheat penne are my favorites. If your (sic) on a low carb or low sodium or any kind of diet (except a high fat diet) they have something for you."

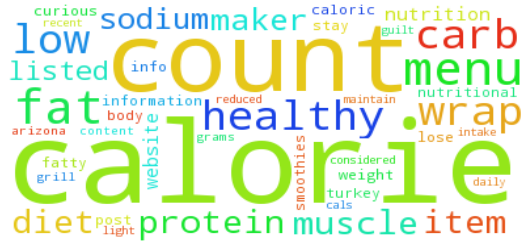
Figure 1.3 visually represents the relative importance of major words in each of the top 20 topics. Online Appendix 2 lists the rest of the topics and 3 provides further details of how widespread the discussion of each topic is.

Next, we present in Figure 1.4 the temporal trends over a 48-month window of the posterior 95% credible intervals of the mean of topic proportions (within a month) for the health topic separately for a) chain restaurants, b) standalone restaurants and c) the difference between the two. July 2008 is the approximate date of regulation implementation (marked by a vertical line).

Figure 1.3: A Visual Representation of Major Topics
The size of a word indicates the relative probability of the word in relation to other words.

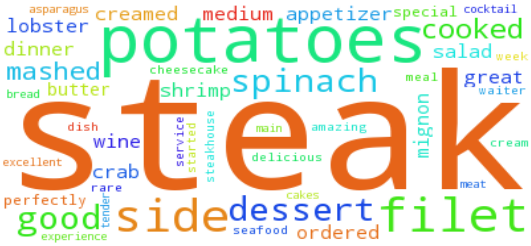
Figure 1.3

HEALTH

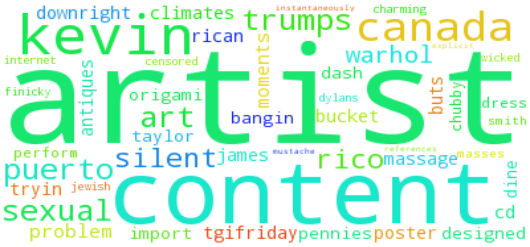


JUICE:ORANGE

STEAK:POTATOES



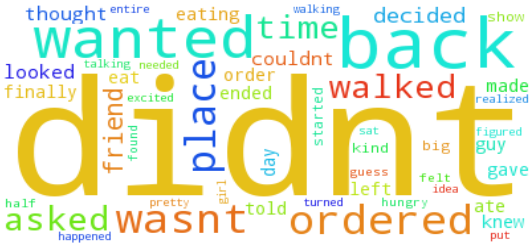
ARTIST:CONTENT



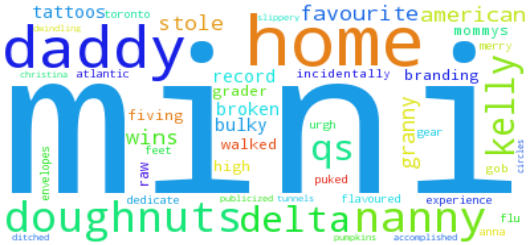
ACCOUNT:FREE

STEAK:WOLFGANGS

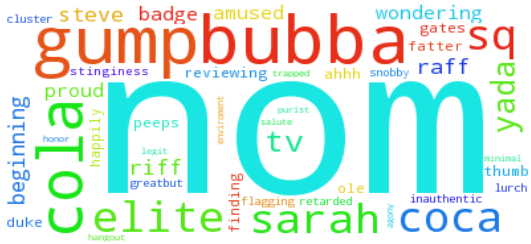
DIDN'T:BACK



MINI:HOME



NOM:BUBBA



CAPITAL:GRILLE

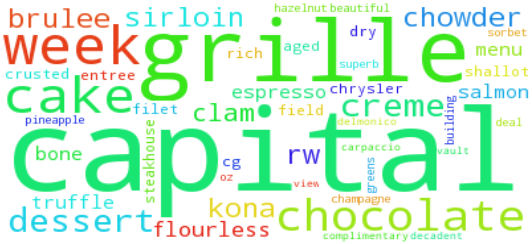
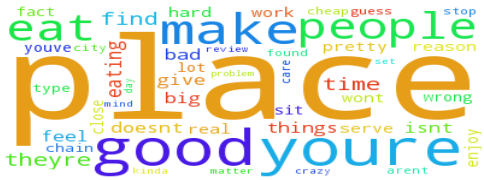
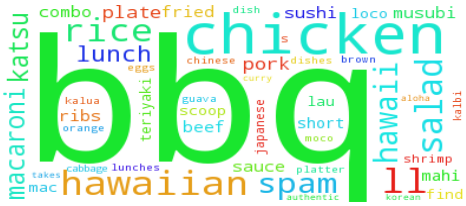


Figure 1.3 (Continued)

PLACE:GOOD



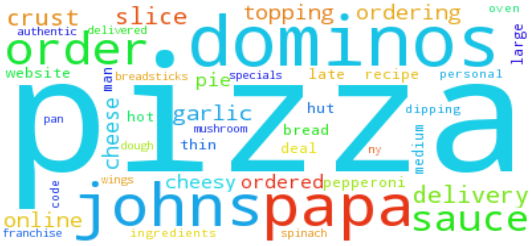
BBQ:CHICKEN



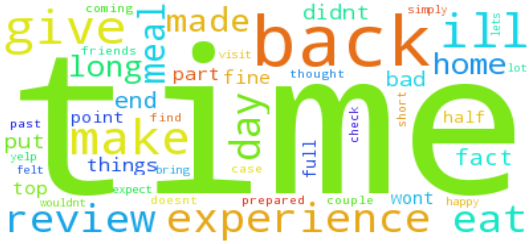
SANDWICH:POTBELLY

TABLE:WAITER

PIZZA:DOMINOS



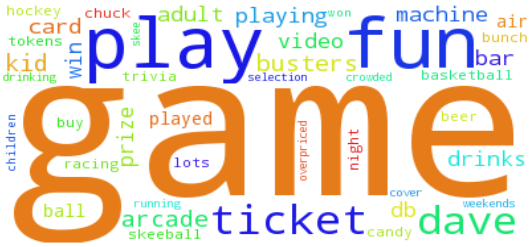
TIME:BACK



CHIPOTLE:BURRITO

STIX:CHINESE

GAMES:GAME



BURGER:FRIES

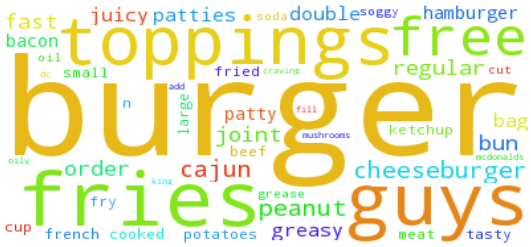
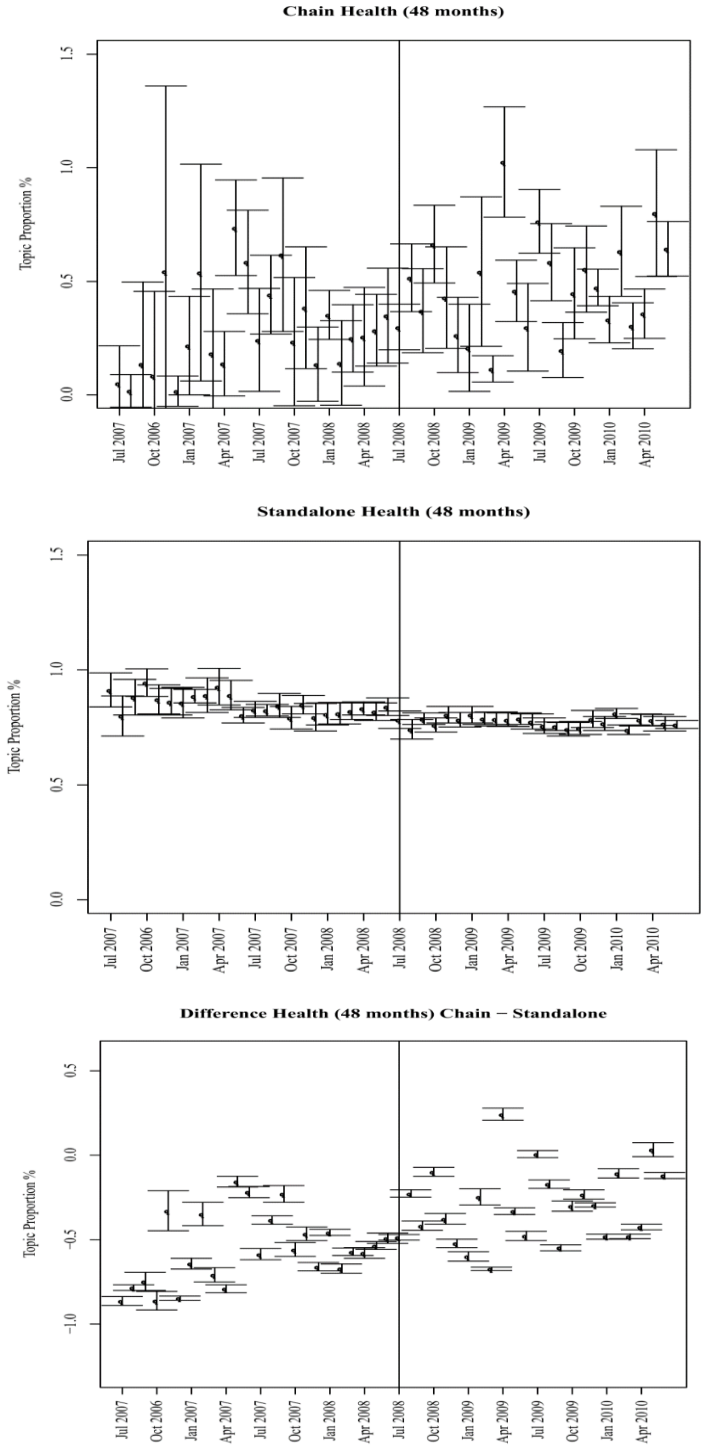


Figure 1.4



Temporal trend of the posterior 95% credible intervals of the mean of topic proportions (within a month) for the health topic.

This visual representation suggests a minor increasing trend for health topic proportions for chain restaurants, a diminishing trend of health topic proportions for standalone restaurants, and much smaller credible intervals for topic proportions for standalone restaurants (this is enabled by the large number of reviews for standalone restaurants), and d) an increasing trend of the difference in topic proportions of the chain and standalone restaurants. Whether this increase in the difference of health discussion between the types of restaurants is statistically significant, and whether this trend changes post-regulation, is difficult to infer from these charts. Temporal trends in of topic proportions for the top 19 topics are presented in Online Appendix 4.

Next we present the results of the difference-in-difference analyses for the health topic as specified in Equation 2, with key parameters presented in Table 1.2. Several insights emerge. First, after controlling for chain characteristics (in comparison to standalone restaurants), temporal trends, restaurant locations, and time-specific shocks, we find that the proportion of the health topic in chain restaurants increases to a greater extent than in standalone restaurants, after the regulation (the coefficient of $Chain_d \_ Post_d$ is positive). This can be construed as evidence in support of the success of the regulation.

The proportions of topics connoting high calorie foods such as "steak:potatoes" and "burger:fries" also increase after the implementation of the regulation. Therefore unfortunately, the relative proportion of discussion of high calorie foods was high before the regulation, and became even higher after it was implemented. From the coefficients of *Post*, it is evident that topics related to brands such as Potbelly (sandwich:potbelly) and Chipotle (chipotle:burrito) garner lower proportion of online reviews post July 1, 2008. Such trends can serve as informative signals for brand managers of the focal and competing brands. Lower online discussion of a brand might be a precursor to decreasing demand for it. However, we prefer not to draw inferences about health discussion from topics other than the seeded health topic. The health topic, irrespective of whether it is seeded or not, contains the words calorie, calories and health with high probabilities, and therefore seems pertinent to study the effect of the calories posting regulation. Now consider the econometrics of inferring topics. Our approach of inferring topics is based on maximizing the dissimilarity between topics by computing a distance between every pair of topics where each topic is a probability distribution over the vocabulary. It follows

that the 199 topics other than the health topic *are least similar* to the health topic, and perhaps not suitable for drawing inferences about health.

Table 1.2: The Effect of Calorie Posting Regulation on Topic Proportions (Posterior Means and Posterior SDs of Coefficients of Equation 2)

		Mean Topic Proportions		
Topic		Coefficient of Chain	Coefficient of Posting	Coefficient of Chain x Posting
"Health"	M	**7.70**	**(1.18)**	**4.66**
	SD	**1.78**	**0.24**	**1.90**
pizza:bbq	M	-1.69	0.45	3.96
	SD	8.56	1.16	9.14
burger:fries	M	**165.21**	**3.01**	**38.39**
	SD	**11.21**	**1.53**	**11.96**
good:place	M	**-508.14**	**-49.44**	-27.62
	SD	**26.35**	**3.59**	28.10
sandwich:bread	M	**52.92**	**8.88**	**42.86**
	SD	**6.63**	**0.90**	**7.07**
salad:soup	M	**115.75**	**(1.97)**	**(49.58)**
	SD	**7.17**	**0.98**	**7.65**
steak:lobster	M	40.92	**30.74**	**59.00**
	SD	20.50	**2.80**	**21.87**
table:waitress	M	**121.64**	**6.19**	**-69.28**
	SD	**9.69**	**1.32**	**10.33**

Note: M and SD stand for the posterior mean and posterior standard deviation of the coefficient estimate. Coefficients whose 95% credible intervals do not contain zero appear in bold.

Next we investigate the source of increase in the proportion of health topic discussion in reviews of chain restaurants. Specifically, we ask if this increase is driven by a small number of authors very vocal about health, or by a relatively large number of reviewers who do not write as much about health. For this purpose, we focused on authors who posted at least one review of a chain restaurant in the data period. We classified each such author as a "high" or "low" discussant of "health", based on whether the mean health topic proportion of his or her reviews exceeds the overall mean health topic proportion (0.41%) in all reviews of chain restaurants. In Table 1.3, we present the number of "high" and "low" authors across the two periods. We find that in both time periods, less than 15% of all authors of reviews of chain restaurants are "high" discussants, suggesting that the increase in health topic proportion post regulation emanated by a small group of authors[19]. We then looked at the health topic proportions of reviews posted in

[19] It is plausible that "low" health discussants contribute to the health discussion by posting more reviews (per author) than "high" health discussants. However, we find the opposite: "low" health discussants post an average of 1.43 reviews in the data

each of the four cells. We find that the health topic proportion substantially increased post-regulation in reviews posted by "high" health discussants (which are much fewer in number), and decreased for the much larger segment of "low" health discussants. This provides further evidence that the increase in health topic proportion can be attributed to a small segment of authors who generally write more (than the average) about health.

We next examine if this increase in health topic discussion post-regulation is driven by authors who also posted reviews before the regulation or from new authors. We start by noting that in the post period, 825 authors posted reviews with above-average health topic discussion. Of these, as many as 802 authors did not post any review in the pre-implementation period. Similarly, of the 5,683 authors who posted reviews with below-average health topic discussion in the post period, 5,546 did not post any reviews in the pre-implementation period. This suggests that both the overall health discussion in the post implementation period, and any increase from the pre period, is driven by reviewers who did not review chain restaurants on the focal website before the regulation. In summary, we find that increase in health topic discussion post regulation is due to reviews posted by a small proportion of authors, who did not post reviews before the regulation.

Table 1.3: Author Level Analysis of Increase in Health Topic Discussion

Number of Authors Discussing Health Before and After the Regulation

	Pre-Regulation	Post-Regulation
"High" health discussants	112	825
"Low" health discussants	711	5,683

Mean (across reviews) of Health Topic Proportions Before and After the Regulation

	Pre-Regulation	Post-Regulation	% Change
"High" health discussants	1.61%	2.36%	46.79%
"Low" health discussants	0.13%	0.05%	-63.08%

Finally, we discuss how the hyperparameter α_{kd} which affects topic distributions for the health topic is affected by observed characteristics of reviews and of the author. Parameter estimates pertaining to Equation 1 appear in Table 1.4. As expected we find that longer reviews

period, compared to 1.83 reviews posted by the average "high" health discussant. So a typical "high" health discussant not only discusses more about health in a typical review, but also posts more reviews.

have lower values of this parameter, suggesting more evenly spread distribution of the health topic. The more negative the valence of the review, the lower is the value of α_{kd} suggesting that authors posting more negative reviews are more balanced in their discussion of the health topic versus other topics. Reviews posted by authors with more prior experience in writing reviews tend to be more focused on fewer topics of discussion[20].

Table 1.4: Effect of Review and Author Characteristics on Hyperparameter α_{kd} for the Health Topic.

	Intercept	Review Length	Author Experience	Rating=1	Rating=2	Rating=3	Rating=4
Posterior Mean	-0.790	-0.562	0.005	-0.401	-0.322	-0.106	-0.016
Posterior SD	0.002	0.006	0.002	0.001	0.001	0.001	0.001

4.2 Dynamic Topic Model

The proposed model assumes that the topics in themselves do not change over the duration of our data (i.e., the probability of word *v* given topic *k* is invariant over time). It may well be possible that the words representing a topic may evolve with time due to changes in word usage[21]. In order to verify this, we estimated a Dynamic Topic Model (Blei and Lafferty 2006) by allowing topic *k* in *t* time period $\phi_{t,k}$ to depend on the natural parameter of the same topic in the previous time period as follows: $\phi_{t,k} \mid \phi_{t,k-1} \sim N(\phi_{t,k-1}, \sigma^2 I)$. We incorporated the review specific characteristics within the model as outlined in section 2.1. This model can potentially reveal interesting patterns of topic evolution. It can also alleviate any potential biases that might arise from making the assumption that topics do not evolve over time on a continuous basis.

In deciding the time period of analysis, we note that our data are 10,823 reviews of chain restaurants over the 8-year period from January 2005 to December 2012. Of these reviews of chain restaurants, 10,779 reviews were posted by 7,156 authors in the period 2006-2012. Reviews were posted sparsely across months in 2005 and consequently we drop 44 reviews from

[20] The author topic model (Rosen-Zvi et al. 2004) extends the LDA model to include authorship information. It assumes that each author has a different mix of topic proportions. A document with multiple authors is then modeled as a distribution over topics that is a mixture of the author-specific distributions. Since all documents in our application are authored by a single individual only, we did not adopt this approach.

[21] We thank our anonymous reviewer for highlighting this possibility.

2005 for the estimation of the dynamic topic model. We use reviews for every month post 2005 up until December, 2012 for estimation. Sparsity of reviews at the weekly level also led us to use month as our unit of analysis. We modeled topic dynamics from January 1st 2006 till December 31st 2012 at a monthly level. To estimate this model, we closely followed the variational inference technique in Blei et al. (2003).[22] We found that an 8-topic model maximizes model fit as per the Jensen-Shannon statistic. So this model in effect requires us to estimate 8 topics in each of 84 months (which is effectively 8×84 = 684 topics), which is far greater than the 200 topics in the static model. Table 1.5A shows the top 20 words by posterior means of the probability (ϕ_{kv}) in the health topic and the corresponding probabilities for every 20th month (in addition to the 84th month). We note that there is little temporal change in the set of top 20 words in the health topic (based on the posterior probability of belonging to the topic). Differences in adjacent time periods are even less noticeable. To ensure that this insight is not specific to the seeded health topic, we present the evolution of top 20 words for a topic that is ranked high by both coherence and importance metrics in our dataset (burger:fries) in Table 1.5B. Again, we do not find much evidence of topic evolution over time (18 words are common to the top 20 words across the first and last month). We find that the same pattern holds if we extend this analysis to the top 50 words. We infer that topics do not evolve to a large extent during the data window.

Based on the estimates of the dynamic model, the resulting measure of health discussion in chain restaurants increases from 0.38% in the pre period and to 0.45% in the post period, a finding consistent with our static model. Given the limited evidence of dynamics in topics across the time periods of interest, we conclude that our assumption of static topics is reasonable. Finally, the dynamic model (Perplexity = 477.93) and the static model (Perplexity = 487.76) offer similar fit to the data. Perplexity is a measure of fit for LDA models and is described in more detail in section 4.5.

[22] This model takes advantage of the multinomial distribution's representation via its mean parameterization. A standard result from the analysis of exponential distributions is that the i[th] component of the natural parameter β_k of a topic (a multinomial with V dimensions) with mean parameter $(\pi_{1,k}, \ldots., \pi_{i,k}, \ldots \pi_{V,k})$ is $\beta_{i,k} = \log(\frac{\pi_{i,k}}{\pi_{V,k}})$. The mean parameters ($\pi_{k,t}$) for a topic in any period can be recovered using this result.

Table 1.5A: Temporal Evolution of "Top 20" Words in the "Health" Topic

	Time Period (in months)					
	$t = 0$	$t = 20$	$t = 40$	$t = 60$	$t=80$	$t =84$
Greatest	calories	calories	calories	calories	calories	calories
Words in Decreasing order of Posterior Probability of a Word Being in Health Topic	fat	fat	fat	fat	Fat	fat
	light	light	light	light	light	light
	calorie	calorie	calorie	calorie	calorie	calorie
	count	count	count	count	count	count
	menu	menu	menu	menu	menu	menu
	low	low	low	low	low	low
	salad	salad	salad	salad	salad	salad
	good	good	good	good	good	good
	yogurt	yogurt	yogurt	healthy	healthy	healthy
	healthy	healthy	healthy	yogurt	yogurt	yogurt
	salads	salads	salads	salads	salads	salads
	place	lean	lean	lean	lean	lean
	lean	place	place	place	place	place
	burrito	burrito	burrito	dressing	dressing	dressing
	dressing	dressing	dressing	burrito	burrito	burrito
	counts	counts	counts	counts	counts	counts
	cheese	cheese	eat	eat	eat	eat
	eat	eat	cheese	cheese	cheese	cheese
Least	fruit	fruit	fruit	love	love	love

Table 1.5B: Temporal Evolution of "Top 20" Words in the "burger:fries" Topic

	Time Period (in months)					
	t = 0	*t* = 20	*t* = 40	*t* = 60	*t=80*	*t* =84
Greatest	Burger	burger	burger	burger	burger	burger
Words in Decreasing order of Posterior Probability of a Word Being in Health Topic	Fries	fries	fries	Fries	fries	fries
	Burgers	burgers	burgers	burgers	burgers	burgers
	Guys	guys	guys	Guys	guys	guys
	Shake	shake	shake	shake	shake	shake
	White	white	white	white	toppings	toppings
	Castle	castle	toppings	toppings	white	white
	Toppings	toppings	castle	castle	cheeseburger	cheeseburger
	Cheeseburger	cheeseburger	cheeseburger	cheeseburger	castle	castle
	Patties	patties	patties	shack	shack	shack
	French	french	french	french	french	french
	Shack	shack	shack	patties	patties	patties
	Patty	patty	patty	Patty	patty	patty
	Onions	onions	onions	onions	bacon	bacon
	Bacon	regular	regular	bacon	onions	onions
	Regular	bacon	bacon	regular	regular	regular
	Joint	joint	cajun	cajun	cajun	cajun
	Cajun	cajun	joint	Bun	bun	bun
	Bun	bun	bun	Joint	joint	joint
Least	Peanuts	peanuts	greasy	greasy	greasy	greasy

4.3 Robustness Checks

We now investigate the robustness of our estimate of a causal effect of the regulation on health topic proportion to several features of the model and the data.

Treatment Timing and Duration of analysis.

Section 81.5 of the New York City Health Code[23] makes the calorie posting regulation effective from May 5, 2008. This code states that the Health Department of NYC would begin citing violations of this requirement from May 5, 2008 and may impose monetary penalties from July 18, 2008. It is plausible that restaurants made changes prior to the regulation date of July 18, 2008 (e.g. healthier menus or lower calorie ingredients) in anticipation of calorie posting. Such changes could have affected health topic proportions even before the regulation was

[23] Source: www.nyc.gov/html/doh/downloads/pdf/cdp/calorie_compliance_guide.pdf (accessed in March 2016)

implemented. We estimated the model including Equation 2 for different temporal breaks (both before and after July 1, 2008), and find that our results hold (Table 1.6A).

Table 1.6A: Differences-in-Differences Analyses with Varying Policy Implementation Dates

		Dependent Variable = Health Topic Proportion		
Model		Coefficient of Chain	Coefficient of Posting	Coefficient of Chain x Posting
Policy Date 1st July, 2008	M	**7.70**	**-1.18**	**4.66**
	SD	**1.78**	**0.24**	**1.90**
Policy Date 1st October, 2008	M	**7.80**	**-0.99**	**4.70**
	SD	**1.60**	**.22**	**1.74**
Policy Date 1st April, 2008	M	**7.80**	**-1.68**	**4.44**
	SD	**1.94**	**0.26**	**2.05**

Note: M and SD stand for the posterior mean and posterior standard deviation of the coefficient estimate. Coefficients whose 95% credible intervals do not contain zero appear in bold.

To further account for the possibility that factors other than the regulation might affect topic proportions of chain restaurants, but not those of standalone restaurants, we conduct a regression discontinuity analysis. Such analysis elicits causal effects of interventions more cleanly by assigning a threshold above or below which an intervention is assigned. Such a threshold in our context is simply the time of implementation of the calorie posting regulation (July 1, 2008). The treatment (mandatory calorie posting) is assigned to chain restaurants only after this cutoff. By comparing observations lying closely on either side of the threshold, it is possible to estimate the local treatment effect in contexts in which randomization was unfeasible. As a result we estimate the regressions discussed above not for all reviews in our data period but for reviews posted in a period of say X months before and after the date of implementation. The smaller the time period of analysis around the date of implementation, the less likely is the occurrence of any events which potentially affect topic proportions of chain restaurants only. We estimate the regressions for X=24 months (i.e. on all reviews posted 24 months prior to and following the regulation), X=18, X=12, and X=6 months. Although the regression coefficients vary in magnitude (Table 1.6B), the coefficients' signs remain the same, showing a positive effect of the regulation from the 12th month onwards. The null effect for shorter timer periods could be driven by implementation delays that prevent a clean or widespread implementation of the policy on the date of enforcement, or by lower statistical power, or consumers taking time to absorb the information from calorie posting.

Table 1.6B: Differences-in-Differences Analyses with Varying Time Window Around Policy Implementation

		Dependent Variable = Health Topic Proportion		
Model		Coefficient of Chain	Coefficient of Posting	Coefficient of Chain x Posting
+/-24 months	M	**8.05**	-0.35	**4.04**
	SD	**1.83**	0.27	**1.90**
+/-18 months	M	**9.19**	0.19	2.34
	SD	**1.88**	0.30	2.35
+/-12 months	M	**8.40**	**0.59**	2.85
	SD	**2.20**	**0.35**	2.83
+/- 6 months	M	**12.45**	**1.94**	-5.43
	SD	**2.99**	**0.49**	3.96

Note: M and SD stand for the posterior mean and posterior standard deviation of the coefficient estimate. Coefficients whose 95% credible intervals do not contain zero appear in bold.

Differences between treatment and control groups. We analyze if our results are driven by two differences between the treatment group (chain) and control group (standalone) in the pre-regulation period: the large difference in mean health topic proportion between the two groups (0.85% versus 0.35%), and the large difference in sample size (there are 61.7 times more reviews of standalone restaurants than those of chain restaurants). For this purpose, we sampled pre-regulation reviews of standalone restaurants to form an alternate control group such that a) the mean of health topic proportions of this control group is the same as that of that of all pre-regulation reviews of chain restaurants (treatment group), and b) the number of reviews in the new control group are the same as those in the control group. We estimated the regression model on this data and again found a positive and significant coefficient of the interaction effect of *Chain* and *Posting* (Posterior mean = 2.13; Posterior SD = 0.77). Although this analysis is subject to limitations (ideally selection to the control and treatment groups should be on a random basis), it shows that our results are not affected by differences between chain and standalone restaurants[24].

Increase in number of authors and reviews. There are many more unique authors and reviews in the post-period as compared to the pre-period for both chain and standalone restaurants (Table 1.3). One concern is that this growth in volume of new authors or reviews might be leading to the

[24] Our identification strategy might not be as robust if consumers who visit chain restaurants did not visit standalone restaurants. Further analyses of authors of reviews of chain restaurants in our data revealed substantial that most authors of reviews of chain restaurants also post reviews of standalone restaurants, suggesting overlaps in consumer segments across the two types of restaurants.

increased discussion of "health" in chain restaurants. It is unlikely this sharp increase in number of new authors is due to the regulation itself. Indeed, we find in our data that the number of authors posting reviews for the first time does increase substantially over time for both types of restaurants.

Our main concern is whether the regulation led to change in this trend for chain restaurants which was different from the change in this trend for standalone restaurants. To analyze this, we regressed the number of new authors in month *t* for chain and standalone restaurants ($t = 1,\ldots,48$) against month *t*, *t_Chain* (interaction of t and the dummy variable for chain restaurants) and *t_Chain _Post* (interaction of *t_Chain* with a dummy variable which takes the value 1 from July 2008).

As expected, we find an increasing temporal trend in the number of new authors for all restaurants (coefficient of $t = 39.46$, $p < 0.01$), and a greater temporal increase in the number of new authors for standalone restaurants (compared with chains) over the entire data period (coefficient of *t_Chain* = -36.37 $p < 0.01$). Most importantly we find that the trend in the number of new authors after the regulation does not differ across types of restaurants (coefficient of *t_Chain _Post* = -1.33, $p < 0.384$). Given this similarity in trends after the regulation, we conclude that the regulation did not have a causal effect on the number of new authors of chain restaurants[25].

We repeated the same regression analysis with the total number of authors as dependent variable, and obtained the same result. Regressing number of reviews also leads to the same inference. Combined with the results obtained from the LDA model, we infer that the regulation affected the content of discussion in reviews of chain restaurants, but not the number of authors or reviews.

Choice of seeds. To understand if our results are driven by our choice of seed words from the regulation document, we estimated the model using another set of seed words. Specifically, Wikipedia (one of the most widely used sources of information globally) contains a description of the word "nutrition" and lists a table of nutrients which need to be consumed for healthy living (http://en.wikipedia.org/wiki/Nutrition). We used all the words describing the nutrients in this table as seeds. These seed words are calories, fat, protein, vitamin, cholesterol, calcium,

[25] These results are robust to the inclusion the following additional covariates in the regression: t^2, t^3, *Chain* and *Post*. Inclusion of more covariates leads to larger standard error of the coefficient of *t_Chain _Post*.

magnesium, sodium, potassium, iron and iodine. We re-estimated the model using these words as seeds, and found that our results are very similar to those obtained by using seeds from this source instead of those from the regulation document. Topic proportions did not vary based on the choice of seeds[26].

4.4 Implications for Managers and Policy Makers

We first discuss implications for policy makers interested in promoting healthy habits among consumers eating out. We find that health is not a prominent topic of discussion among hundreds of thousands of reviewers of restaurants in New York City. With over 57% of all adults in the city being overweight or obese[27], this is worrisome. Most reviewers of restaurants discuss health to a very low extent or not at all. Interestingly, much of the discussion of health is skewed towards a small segment of reviewers, who can be readily identified online. They could serve as useful starting points for initiatives to identify influencers or evangelists who might be successful in changing online public opinion about health.

We find that the calorie posting regulation was successful in increasing in the proportion of discussion of health related words among online reviews of chain restaurants. Regulators are interested in understanding the effect of the regulation on the overall level of discussion of health. This is a function of a) changes in the volume of reviews posted after the regulation, and b) changes in the mean proportion of health topic discussion across reviews. In this research we focus on the latter, since this is methodologically much more difficult to address, and more novel to the literature. Yet we note that the volume of reviews of chain restaurants increased over 7 times after the regulation (from 1,287 reviews in the 24 months before the regulation period to 9,536 reviews in the 24 months after). A small increase in health topic proportion per review, combined with a very large increase in the volume of reviews, suggests a substantially positive overall effect of the regulation on the level of online health discussion. Managers of restaurants with healthier offerings might be encouraged by this trend and managers of restaurants with less healthy offerings might consider conducting more market research to determine whether and how to alter their strategy. This is an encouraging sign of success of the regulation, and it

[26] Since our interest is specific to discussion of health, seeding only one topic is sufficient in our context. Recent advances in the LDA literature (Jagarlamudi, Daumé and Udupa 2012) allow for seeding multiple topics, such that each topic is a mixture of two distributions: a seeded topic and a regular topic. Such methods are more appropriate for contexts with multiple topics of interest.

[27] Source: http://www.health.ny.gov/statistics/prevention/obesity/county/newyorkcity.htm

provides a basis for conducting further (and costlier) studies into consumption of healthier products as a logical next step. We find that the increase in health discussion after the regulation was largely driven by reviewers who were not active in posting reviews before the regulation, but posted many more reviews after the regulation. Several consumers discussed health in online restaurant reviews for the first time (on our website) after the regulation. Policy makers might be interested in understanding what kinds of reviewers are more vocal about health after the regulation. We find that it is not the most prolific reviewers, but a small number of new reviewers, who were responsible for greater health discussion after the regulation.

While these results are econometrically significant, are they economically significant? Our estimate of a 17.1% increase in health topic proportion (from 0.35% to 0.41%) is consistent with research based on transaction data. Bollinger et al. (2011) estimate a 6% decrease in calories per transaction at Starbucks after the regulation, but no change in overall revenues. Irrespective of the data source and research methodology, such small effect sizes might suggest that the regulation was not a success. However, anecdotal evidence suggests that this effect might be material in economic and social terms. Small changes in consumer behavior have been known to bring about major changes in obesity levels. Kuo et al. (2009) estimate that even if 10% of restaurant patrons in Los Angeles county were to reduce calorie consumption by 100 calories per meal, as much as 40.6% of average annual weight gain in the entire county population would be averted. Reduction in obesity levels has monumental social and economic significance in the US. Over 250,000 deaths in the US every year are attributable to obesity (Allison et al. 1999). Obesity related costs in the US in 2008 were estimated to be a staggering $147 billion (Finkelstein et al. 2009), and are still rising.

Another key finding is that topics pertaining to health, price and service garner a smaller proportion of online reviews than those pertaining to brands and menu items. To the extent that these topics are correlated with product attributes which consumers use for choice decisions, this serves as a free and externally valid input into product management decisions. For making trade-offs between investing in service or menu redesign, it is useful for managers to know that menu items get discussed far more than service. Among menu items, the fact that steaks, burgers and sandwiches are discussed more than salads and appetizers is an indication of the relative popularity of various food items for eating out in New York City.

It is important for health regulators to understand whether health regulation has a greater effect on relatively less healthy consumers. In the absence of disaggregated data on health measures at zipcode level, researchers have often used race as a proxy (Boardman et al. 2005), given greater prevalence of food-correlated diseases (specifically obesity, hypertension, diabetes) among African American people. We separated reviews from African-American majority zipcodes[28] (henceforth referred to as African American neighborhoods), from other reviews. We find that the mean level of health discussion of chain restaurants in African American neighborhoods increased from 0.33% before the regulation to 0.63% afterwards, providing evidence for greater success of the regulation among neighborhoods inhabited by less healthy consumers on average[29]. This is especially encouraging since health discussion of standalone restaurants in African American neighborhoods *fell* from 0.87% to 0.76%, which is consistent with the general view of declining salience of health. African American consumers are known to get a greater share of their calories from fast food restaurants – typically organized as chains[30], suggesting that the regulation made calorific information more salient for African American consumers.

In the other neighborhoods, health discussion of chain restaurants increased to a much lesser extent (from 0.35% to 0.42%) and fell from 0.85% to 0.77% in standalone restaurants. A regression analysis of health topic discussion on dummies for chain (1 if chain), post (1 if post regulation, 0 otherwise), race (1 if African American neighborhood, 0 otherwise), interactions of chain_post, chain_race, post_race and chain_post_race, along with restaurant and author specific random effects, revealed that the crucial coefficient of chain_post_race is positive and significant at the 10% level (M = 0.88, SD = 0.51). Although the evidence that increase in health discussion of chain restaurants is greater for African American neighborhoods is not very compelling from a statistical standpoint, it does strongly indicate that increase in health discussion is not localized in non-African American neighborhoods which typically tend to be more healthy, and have better access to healthier restaurants.

[28] Source: 2010 census conducted by the US census bureau.
[29] This analysis assumes that restaurants located in African American neighborhoods are disproportionately frequented by African American consumers. We were unable to find any evidence to the contrary.
[30] Source: http://www.cdc.gov/nchs/data/factsheets/factsheet_nutrition_data.pdf

Finally, our analysis reveals useful insights for brand managers of restaurants. Topics in Table 1.1A reveal words which are commonly used along with certain brand names in consumer reviews. We note that Subway is the only brand among the top 10 words for the topic "sandwich:potbelly" suggesting that Potbelly and Subway are perceived to be similar by consumers. This could potentially serve as a useful input for future store choice decisions where one brand might not want to locate itself very close to the other. Food items frequently mentioned with a brand indicate which items a brand is associated with. Based on this, Chipotle is more strongly associated with burritos and chicken, and not as much with tacos or beef. This could serve as in input into a formal menu planning exercise – more items related to burritos and chicken might make these brand-product associations stronger.

4.5 Model Comparison

We assess improvement in model performance due to the incorporation of two features which are unique to the marketing literature: a) allowing the researcher to seed certain topics with specific words which are considered substantively important, and b) allowing the distribution of topics within a document to be affected by the characteristics of the document (length, rating and author experience). Model A is an unseeded model, i.e. we do not impose any prior distribution ϕ_k of any topic to contain any word with high probability. Model B is identical to the proposed model with the exception that we assume that θ_d is drawn from a Dirichlet distribution with parameter α which is invariant across reviews, and does not depend on review characteristics. For model comparison, we compute the perplexity score - the likelihood of observing a collection of words given a model of how the words were generated. It is monotonically decreasing in the likelihood of the data, such that models with lower perplexity fit the data better. Such an approach is commonly used for model comparison in the Natural Language Processing literature (Blei et al. 2003, Arora, Ge and Moitra 2012), and is defined as follows:

$$perplexity(D_{test}) = exp\left(-\frac{\sum_{d=1}^{D}\log \quad p(w_{test}{}^{d} \mid D_{train})}{\sum_{d=1}^{D} N_d}\right) \tag{9}$$

Where, $p(w_{test}^{\ d} \mid D_{train}) = \prod_{i=1}^{n_d} \sum_{k=1}^{K} \prod_{v=1}^{V} [\phi_{k,v}^{\ train} \times \theta_d^{k\ \ test}]^{I(w_{id}=v)}$ (10)

w_{di} is the i^{th} word of document d. Perplexity scores for Model A, Model B and the proposed model are 481.73, 531.40 and 487.76 respectively. Our seeded model is comparable in fit to the more flexible unseeded model (Model A)[31]. Therefore, seeding enables incorporating of managerial intuition and offers much richer managerial insights, at very little cost in terms of model fit. This is to be expected since our seeding approach is sufficiently flexible to allow to seed words to get "flooded" with other words. In terms of substantive insights, we find that a separate health topic emerges even in the unseeded model. The top 10 words by posterior probability for this topic are calories, calorie, healthy, menu, fat, count, chicken, low, protein and meal. Health topic proportions of chain and standalone restaurants estimated from Model A follow the same temporal trends as those in the proposed model.

We find that Model B, which ignores the review characteristics, performs worse than the proposed model. This provides empirical validity to the notion that longer reviews and reviews with more positive valence have a more even distribution of topics. To the extent that the topics discussed in reviews are indicative of reviewer's attribute preferences, review length and valence can serve as an easily measurable and observable segmentation variable. To the best of our knowledge, current targeting technologies for social media do not consider the length of content posted online. Our findings suggest this might be a fruitful area for further research.

Section 5. Conclusion

The growth of the internet has led to the availability of very large quantities of data that are often less structured than data collected offline. Such data are often in the form of opinions of consumers (e.g. blogs, product reviews), are from an increasingly representative subset of the population, are in the public domain, and are available for long periods of time (e.g. 8 years in this research). This provides an unprecedented opportunity for marketers to not only understand what consumers are saying about their products at a given point in time, but also to continuously track changes in consumer opinion over time. However, a major challenge for researchers is that

[31] We prefer a seeded model to an unseeded model since it allows us to focus the analysis of the corpus on a few specific words of interest. We do not depend solely on the data to generate the topic of interest. Managers and regulators can use seeded models to measure policy-relevant variables of interest, or to monitor the effects of their actions. However, drawing inferences from seeded models might not be advisable if they offer much lower fit.

much of these data are textual. It is perhaps for this reason that much of the research based on user-generated online content has focused on numerical descriptors of these data or simpler measures like word count. Techniques to analyze large volumes of text are at a nascent stage even in computer science. Yet, there is considerable interest from practitioners in using these data to gain usable knowledgeable. A recent report by the McKinsey Global Institute (Manyika et al. 2011) suggests that analyzing such data will become a "key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus."

Early research using online textual data in marketing has been focused on inferring market structure and product attributes in specific product categories; to ascertain the extent to which these correlate with consumer level data collected from more traditional experimental and survey based techniques; and to incorporate measures of such data in demand models. We extend this work by using textual data to address an issue that has perhaps been infeasible otherwise: how can researchers track changes in consumer opinion over time, and assess the impact of exogenous events on such changes? Specifically, we assess the impact of a regulation to post calories in chain restaurants on consumer opinion pertaining to chain restaurants. Across marketing and computer science, we were unable to find other research that uses textual data to infer the effect of any factor on consumer opinion. We find significant changes in proportions of various topics of discussion due to the implementation of the regulation. Methodologically, we extend the Latent Dirichlet Allocation set of models in computer science. We look forward to several strategy- and policy-relevant applications as well as more sophisticated models in this area of topic detection and measurement.

References

Allison, D. B., K. R. Fontaine, J. E. Manson, J. Stevens, T.D. VanItallie. 1999. Annual Deaths Attributable to Obesity in the United States. *The Journal of the American Medical Association*. 282(16), 1530--1538.

AlSumait, L., D. Barbará, J. Gentle, C. Domeniconi. 2009. Topic Significance Ranking of LDA Generative Models. *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*. 67--82.

Archak, N., A. Ghose, P. G. Ipeirotis. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*. 57(8), 1485--1509.

Arora, S., R. Ge, A. Moitra. 2012. Learning Topic Models - Going beyond SVD. *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*. 1--10.

Bird, S., E. Loper, E. Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Blattberg, R. C., S. J. Hoch. 1990. Database Models and Managerial Intuition: 50% Model+ 50% Manager. *Management Science*. 36(8), 887--899.

Blei, D. M., A. Ng, M. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. Volume 3, 993--1022.

Blei, D. M., J. D. Lafferty. 2006. Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning*, 113-120.

Boardman, J. D., Saint Onge, J. M., Rogers, R. G., Denney, J. T. 2005. Race differentials in obesity: the impact of place. *Journal of health and social behavior*, *46*(3), 229-243.

Bollinger, B., P. Leslie, A. Sorensen. 2011. Calorie Posting in Chain Restaurants. *American Economic Journal: Economic Policy*. 91-128.

Cao, J., T. Xia, J. Li, Y. Zhang, S. Tang. 2009. A Density-Based Method for Adaptive LDA Model Selection. *Neurocomputing*. 72(7-9), 1775--1781.

Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, D. M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information Processing Systems*, 1-9.

Decker, R., M. Trusov. 2010. Estimating Aggregate Consumer Preferences from Online Product Reviews. *International Journal of Research in Marketing*, 27(4), 293--307.

Deveaud, R., E. SanJuan, P. Bellot. 2012. LIA at TREC 2012 Web Track: Unsupervised Search Concepts Identification from General Sources of Information, *Proceedings of the 21th Text REtrieval Conference* (TREC 2012), Gaithersburg, USA, November 7-9.

Dickey, J. M. 1983. Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. *Journal of the American Statistical Association*. 78(383), 628--637.

Downs, J. S., J. Wisdom, B. Wansink, G. Loewenstein. 2013. Supplementing Menu Labeling With Calorie Recommendations to Test for Facilitation Effects. *American Journal of Public Health*. 103(9), 1604--1609.

Eliashberg, J., S. K. Hui, J. Zhang. 2007. From Story Line to Box Office: A New Approach for Green-lighting Movie Scripts. *Management Science*. 53(6), 881--893.

Finkelstein, E. A., J. G. Trogdon, J. W. Cohen, W. Dietz. 2009. Annual Medical Spending Attributable to Obesity: Payer and Service Specific Estimates. *Health Affairs*. 28(5), 822--831.

Fuchs, F. D., 2011. Why do Black Americans have Higher Prevalence of Hypertension: An Enigma Still Unsolved. *Hypertension*. 57, 379-380.

Ghose, A., P. G. Ipeirotis, B. Li. 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-generated and Crowdsourced Content. *Marketing Science*. 31(3), 493--520.

Godes, D., D. Mayzlin. 2004. Using Online Conversations to Study Word-of-Mouth Communication. *Marketing Science*. 23(4), 545--560.

Godes, D., J. C. Silva. 2013. Sequential and Temporal Dynamics of Online Opinion. *Marketing Science*. 31(3). 448--473.

Griffin, A., J. R. Hauser. 1993. The Voice of the Customer. *Marketing Science*. 12(1), 1--27.

Griffiths, T., M. Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*. 101(Suppl 1), 5228--5235.

Hartmann, W., H. S. Nair, S. Narayanan. 2011. Identifying Causal Marketing Mix Effects Using a Regression Discontinuity Design. *Marketing Science*. 30(6), 1079--1097.

Jagarlamudi, J., H. Daumé III, R. Udupa. 2012. Incorporating Lexical Priors into Topic Models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 204-213.

Jalonick, M. C. 2013. FDA Head Says Menu Labeling 'Thorny' Issue. *Associated Press*, March 12.

Kullback, S., R. Leibler. 1951. On Information and Sufficiency. *Annals of Mathematical Statistics*. 22 (1), 79-86.

Kuo, T., C. J. Jarosz, P. Simon, J. E. Fielding. 2009. Menu Labeling as a Potential Strategy for Combating the Obesity Epidemic: a Health Impact Assessment. *American Journal of Public Health*. 99(9), 1680--1686.

Lee, T. Y., E. T. Bradlow. 2011. Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*. 48(5), 881--894.

Lin, J., 1991. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*. 37(1), 145--151.

Lu, B., M. Ott, C. Cardie, B. Tsou. 2011. Multi-aspect Sentiment Analysis with Topic Models. Data Mining Workshops (ICDMW), *2011 IEEE 11th International Conference*. 81--88.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute*, May.

Mimno, David,A. McCallum. 2008,Topic models conditioned on arbitrary features with Dirichlet-Multinomial Regression , *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08), AUAI Press.*

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, A. McCallum. 2011. Optimizing Semantic Coherence in Topic Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262--272.

Minka, T. P. 2000. Estimating a Dirichlet Distribution. http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf.

Netzer, O., R. Feldman, J. Goldenberg, M. Fresko. 2012. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*. 31(3), 521--543.

Roe, B., A. S. Levy, B. M. Derby. 1999. The Impact of Health Claims on Consumer Search and Product Evaluation Outcomes: Results from FDA Experimental Data. *Journal of Public Policy and Marketing*. 89--105.

Rosen-Zvi, M., T. Griffiths, M. Steyvers, P.Smyth .2004. The Author-Topic Model for Authors and Documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. 487-494.

Steyvers, M., T. Griffiths. 2007. Probabilistic Topic Models. *Handbook of Latent Semantic Analysis*. 427(7), 424--440.

Thistlethwaite, D. L., D. T. Campbell. 1960. Regression-discontinuity Analysis: An Alternative to the Ex Post Facto Experiment. *Journal of Educational Psychology*. 51(6), 309.

Tirunillai, S., G. Tellis. 2014. Mining Marketing Meaning from Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*.

Wierenga, B. 2006. Motion pictures: Consumers, Channels, and Intuition. *Marketing Science*. 25(6), 674--677.

Yaniv, I., R. M. Hogarth. 1993. Judgmental Versus Statistical Prediction: Information Asymmetry and Combination Rules. *Psychological Science*. 4(1), 58--62.

CHAPTER 2

THE IMPACT OF MARKET DISRUPTION ON CONSUMER EXPERIENCE: WHEN UBER COMES TO TOWN

Section 1: Introduction

Taxi service is a regulated industry in the United Sates. Taxi service providers bid for a "medallion" or a license to operate a taxi. The number of medallions is regulated by the local government. Uber Technologies disrupted[32] the traditional taxi service industry by offering a software service on mobile phones that allows any individual with a car and a driver's license to offer transportation services. This resulted in an explosion in the supply of taxi-like service in several cities across the globe. For example, recent reports indicate that there are approximately 8,500 licensed taxi drivers in San Francisco but over 11,000 active Uber drivers[33].

Consider two perspectives - managerial and regulatory. Managers (of traditional taxis and Uber) might value understanding how consumers view the disruption; these customer views can provide insights to these companies on how to change their product offering. From a regulatory perspective, the normative implications of this market disruption are uncertain. Pre-Uber entry, while high medallion prices might signal high producer surplus, regulations on fares and licensing of drivers might ensure reasonable consumer satisfaction and hence consumer welfare. Post-entry, the introduction of more taxi supply has the potential to enhance consumer satisfaction and hence consumer welfare, assuming quality of taxi service (both traditional and

[32] "After Uber, San Francisco Has Seen a 65% Decline in Cab Use" http://www.theatlantic.com/technology/archive/2014/09/what-uber-is-doing-to-cabs-in-san-francisco-in-1-crazy-chart/380378/

[33] http://www.sfexaminer.com/sanfrancisco/uber-releases-driver-data-for-first-time-and-its-not-pretty-for-taxi-industry/Content?oid=2917635, http://www.bizjournals.com/sanfrancisco/blog/2014/01/city-sees-taxi-driver-shortage.html

Uber taxis) does not deteriorate with increased competition. Uber (being unregulated)[34] introduced dynamic pricing or "Surge Pricing" and simplified driver recruitment procedures. These practices might generate concerns about consumer welfare. Overall, whether the potential for enhanced consumer welfare (via increased competition) is realized is an empirical question.

A simple (if incomplete) measure of consumer welfare and whether their preferences are being served is their own view of their consumption experience- i.e. ratings and reviews for these services. Numerical ratings are a good summary measure, with review text offering further detail (we discuss in section 2 why these measures are complementary). Consequently we investigate consumer experience by jointly analyzing ratings and text based measures. Our dataset comprises of 21,232 online reviews posted between January, 2006 and November, 2014 on a leading review site for taxi services. These reviews include reviews for San Francisco (the first city where Uber entered) and San Jose (a city that serves as a potential control group in our analysis, see section 2 for details).

Overall customer experience (measured by rating) for traditional services in San Francisco significantly declines post-Uber entry (see section 2). The decline in traditionals' ratings is a cause for concern for managers of traditional taxis; Uber managers cannot afford to be sanguine about higher ratings relative to traditionals since their ratings also have a downward time trend. From a regulatory perspective, there appears to be potential for consumer welfare improvement based on consumers' positive experiences with Uber.

However, the changes in aggregate rating do not offer insight on specific attributes that maybe of interest to firms and regulators. We focus on product attributes mentioned in various

[34] Uber positioned itself as a software platform to match consumers and drivers, and is not legally a taxi company.

business press articles and are likely to be of interest to regulators and managers- safety, ease of reservation, payment mode, tipping, fares, cleanliness, and wait time.

We examine the more nuanced managerial and regulatory insights from this analysis of attribute evaluation for traditionals and Uber. We conduct the following broad analyses - do consumer views of positive and negative attributes for traditional taxis (before and after Uber's entry) differ from those for Uber, and how do these attributes matter to overall ratings of these two services? Are there specific attributes that are more salient in the discussion for traditionals and for Uber? To the best of our knowledge, this is the first attempt in marketing to study changes in several aspects of consumer experience of traditional players due to a disruptive firm. In order to extract measures of discussion for each of the above mentioned attributes, we calibrate a Bayesian model for analyzing textual data, specifically designed to measure multiple focal topics of interest. Our model belongs to a class of probabilistic topic models termed Latent Dirichlet Allocation (LDA) models (Blei et al, 2003). LDA models have been developed by computer science (specifically machine learning) researchers and marketing to analyze words of large sets of original texts in order to discover the themes or topics that run through them. We incorporate extensions to the model in order to specifically address aspects of analytical interest. The final model specification enables researchers and managers to focus on specific aspects of consumer experience by utilizing informed priors. The model also incorporates valence by exploiting both rating information and informed priors.

In order to test the robustness of our results, we estimate a differences in differences regression model with San Jose as control. Specifically, the physically proximate city of San Jose

did not witness the entry of Uber [35] till July 2013. Consequently, we can use reviews posted for taxi services in San Jose as a control group for the purpose of inferring the causal effect of Uber's entry in San Francisco.

Our research is related to two stream of literature in marketing. First, our substantive area of interest, disruptive innovation (Christensen, 1997) that has received considerable research attention. However, the idea has also been the subject of some debate. For example, disruption is not often not evident till after it has occurred (Danneels, 2004; Sood & Tellis, 2011). Additionally, there is no single price-quality winning strategy for incumbents or entrants. Given this diversity of entrant strategies, incumbents might survive disruption by responding appropriately (Markides, 2006). In our context, the entrant, Uber, possibly disrupts an existing market significantly by offering a service that appeals to major consumer segments right at the outset. This is in contrast to the idea that a potential disruptor first attracts a niche segment of customers by offering a poorer quality though cheaper service. Managers of traditional taxis and Uber may be interested in understanding a) how does the presence of a new alternative (from the entrant) affect how the services of the incumbents are viewed? b)what aspects of the service experience serve to differentiate the incumbents and entrants? c) what do a) and b) imply for the overall consumer experience? d) does the entrant introduce a new set of aspects that consumers now evaluate their experience on? e) if yes how do incumbent perform on these dimensions? and finally e) what might regulators do to continue to protect consumer interests? In order to address these questions, we would need to measure each aspect of consumer experience – both before and after the entry of the potentially disruptive firm.

[35] http://www.mercurynews.com/ci_23722259/uber-car-service-launching-silicon-valley

The second stream of literature to which our work is related is on examining user generated content such as blogs and reviews. User generated content can influence subsequent consumer choice and firm profits. For example, (Chevalier & Mayzlin, 2006) examine the effect of word of mouth on book sales using ratings data and measure of text such as counts of different types of words. (Tirunillai & Tellis, 2012) examine the effect of positive and negative word of mouth on stock returns using ratings data and keywords (see also Berger, et al., 2010, Chen & Xie, 2008 Chintagunta, et al., 2010, Gopinath, et al., 2014 and Luo, 2009). User-generated content is also useful for prediction tasks and for extracting market intelligence. For example (Netzer, et al., 2012) develop a brand map for the automobile industry using data from consumer forums. (Lee and Bradlow, 2011) automatically extract product features from consumer reviews (see also Archak, et al., 2011,Bao & Datta, 2014Das & Chen, 2007,Ghose, et al., 2012, Huber, et al., 2014Ma, et al., 2015, and Mela, et al., 2013). Our work is related to this latter stream, since our focal interest in extracting measures of consumer experience to address questions related to market disruption.

Next we discuss the data. Section 3 then presents the model specification, and discusses specific estimation challenges. Section 4 presents the results from the model, and their implications. Section 5 compares model performance with models which do not incorporate the unique features discussed above. Section 6 concludes.

Section 2: Data

We collected all (21,232) reviews posted on a leading online reviews portal for taxi services as of November, 2014[36]. An examination of the associated review rating data for

[36] We note that Uber offers in-cab survey and feedback mechanisms that serve as platforms to capture consumer opinions. Consequently, Uber consumers may not post reviews on this leading review website as frequently as non-Uber customers.

traditional cab services in San Francisco suggests that the average experience (as approximated by the rating) declined from (M=2.75, SD =1.71) pre-Uber entry to (M=2.59.SD=1.83) post-Uber entry. In contrast, Uber's rating is higher than traditional taxis' ratings (both prior to and post entry of Uber) (M=3.28, SD= 1.73). However, Uber's ratings also decline over time (see figure 2.1).

Figure 2.1

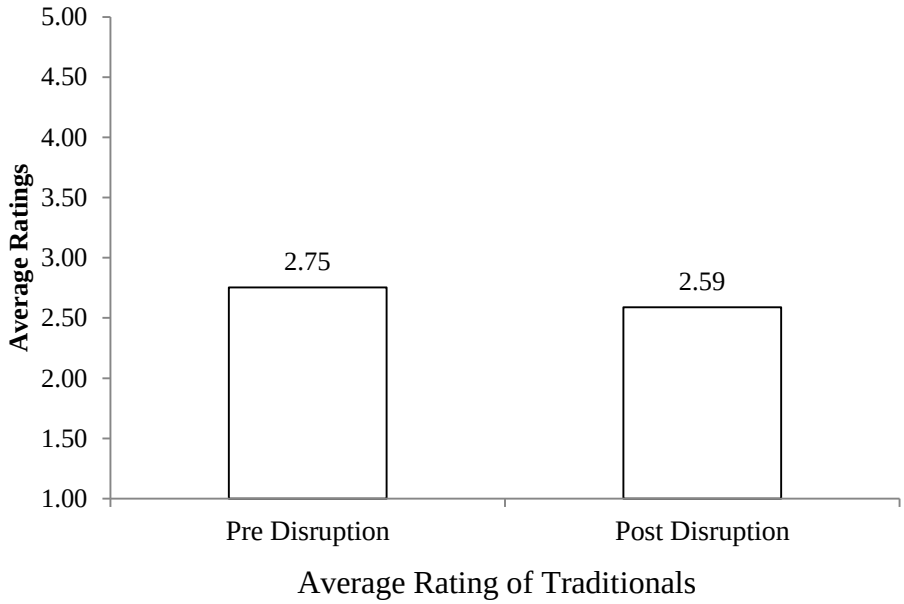
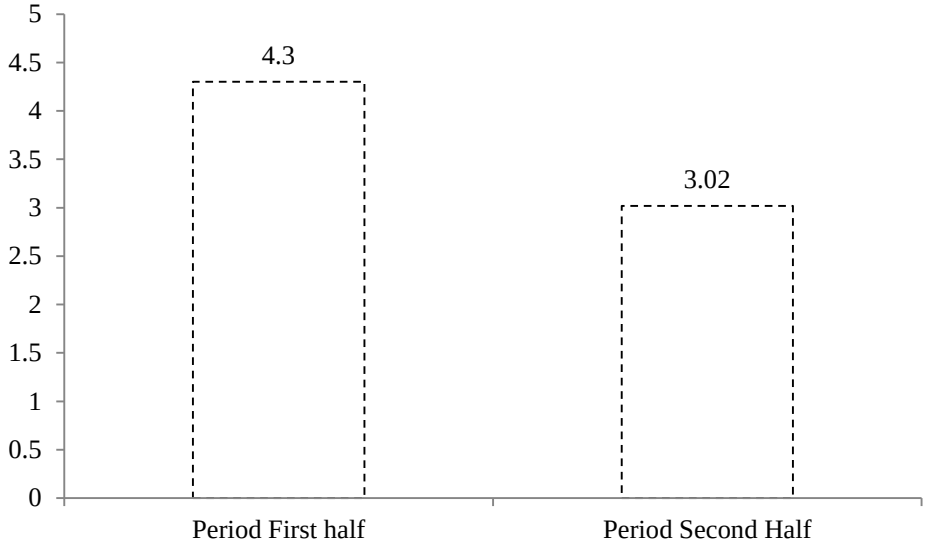


Figure 2.1 (Continued)



Average Rating for Uber

Rating Trends

There are several interesting features of this category that make it distinct from other studies of user generated content. The average rating does not reflect the positivity bias (Moe and Schweidel, 2012) observed in reviews of other categories. (Note this dissatisfaction with incumbent offering makes this market ripe for disruption). Also, the average number of reviews posted per reviewer is small (1.13). These features suggest that this is a category with both low levels of consumer experience and low engagement levels. As with any study using user generated content, selection bias is a potential limitation of this study. Specifically the selection bias may lead to the reviewers who are possibly more vocal and critical than the general population posting on the review site.

Further, one might be concerned whether these reviews can actually influence consumer choice for taxi services unlike other products and services studied by marketing researchers (Archak et al, 2011; Chintagunta et al, 2010). In addition to the influence of user generated

content on choice, consumer opinions as captured in surveys or via user generated content also offer a measure of performance for the different aspects of service/ product experience (Netzer et al, 2012; Lee & Bradlow, 2011). It is in this latter sense that we proceed with our analysis.

We now briefly describe our textual data to enable a better understanding of the model. Our data set includes reviews for taxi services in San Francisco (our city of focus) and San Jose. In order to test the robustness of our results, we estimate a differences in differences regression model with San Jose as control. Specifically, the physically proximate city of San Jose did not witness the entry of Uber till July 2013. Hence, we can use reviews posted for taxi services in San Jose as a control group for the purpose of inferring the causal effect of Uber's entry in San Francisco. We provide evidence for the appropriateness in the choice of San Jose as a control in section 5 below. The mean length of all 21,232 reviews is 122.88 words (SD=114.30). Each sentence is split into its component words using the Natural Language Toolkit's Tokenizer (Bird 2009). After eliminating stop words ("a", "the" etc.)[37] and words that occurred less than 5 times in the entire corpus (Griffiths and Steyvers 2004, Lu et al. 2011) the number of unique words in the corpus is 7,670.

We start by reporting the top 10 words and bottom 10 words across all reviews for non-Uber reviews in San Francisco in our data in Table 2.1. We observe words that are interesting but not very surprising. Given that the data set comprises of reviews about taxi and cab services, it is hardly surprising that "Cab" is the most frequently used word. However, both "Time" and "Minutes" are in the top 10, suggesting that possibly punctuality or speed or response is of particular interest in the context of cabs.

[37] We used a standard list of stop words available from the Mallet toolkit , http://mallet.cs.umass.edu/.

Table 2.1: Top 10 and Bottom 10 Words by frequency

Top 10		Bottom 10	
Word	**Frequency**	**Word**	**Frequency**
Cab	7,382	Saver	6
Driver	5,484	Stinks	6
Time	5,005	Offenses	6
Service	4,343	Paris	6
Called	3,303	Unanswered	6
Minutes	3,107	Stolen	6
Airport	3,054	Notifying	6
Call	2,930	Announced	6
company	2,474	Surrounding	6
Ride	2,328	Preferences	6

While this analysis of word counts is useful to obtain a preliminary sense of the data, it is difficult to glean further insight without relying on a definitive list of words to represent each topic of interest. However, any word count method has to rely on specifying ex-ante what words to count. In a context of evolving consumer preferences and market structure, the choice of words to count is likely to be subjective and inaccurate (see Tirunillai and Tellis (2014) for an effective and appropriate usage of separately validated words across multiple studies).

To address this challenge, we propose an LDA style model with suitable extensions that address the specific research questions. In our case, each review is a "document" as defined in the LDA terminology. The collection of all reviews is called the "corpus". A key advantage of LDA models is that the documents do not need any prior annotation (e.g. deciding which words represent "safety" across all reviews). Such modeling enables us to summarize the corpus at a scale that would be impossible by human annotation (Blei 2012). LDA style models offer a data-based, replicable, objective and principled methodology of inferring topics from text corpora.

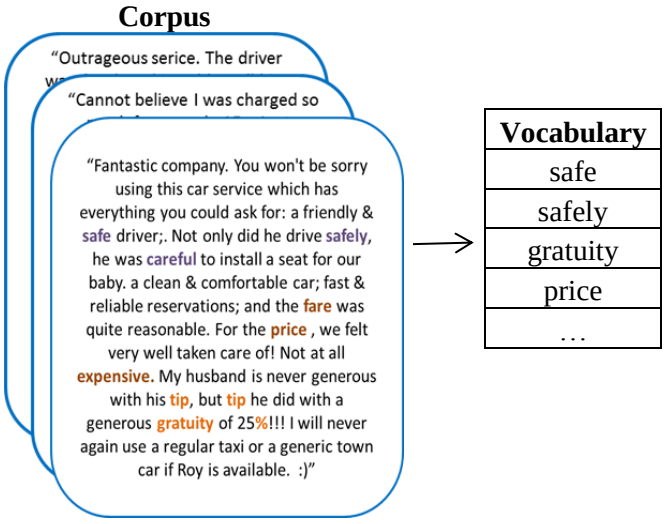
Section 3: The Model

In order to address the questions discussed earlier, our analyses proceeds as follows. We will describe the model specification that will enable us to extract measures for multiple aspects of consumer experience. Next we will describe the estimation procedure to obtain the measures. Finally, we will propose a framework to analyze these measures.

3.1 Extracting measures for multiple aspects of consumer experience

We start with some standard definitions and notation. Each review r, (r=1,…, R) is composed of n_r words. The number of total word instances in the corpus is N (e.g. for a corpus of 10,000 reviews with 100 words each, N is 1 million). The corpus is therefore defined by an N-dimensional vector $\boldsymbol{w} = \{w_{11}, w_{12}, w_{ri}, .., w_{Rn_R}\}$, where w_{ri} is the i^{th} word of review r. A vocabulary is the set of all unique words across all reviews (e.g. all unique words in all reviews posted on a website). It is defined by V unique words, where a word is denoted by v. For example, Figure 2.2 shows a list of words that are a subset of the vocabulary.

Figure 2.2



Corpus and Vocabulary

Our objective is to identify and measure latent topics in our corpus as in standard topic modeling (Blei et al, 2003).In general, this approach involves the following ideas. Each word in the vocabulary belongs to every topic, such that topic ϕ_t (t=1,..,*T*) is a probability distribution over all *v* in $\{1,2,..,V\}$. Element ϕ_{tv} denotes the probability of word *v* given topic *t*. Review *d* is a mixture of the *T* topic. θ_r is a *T*-dimensional vector that represents the proportions of each topic in review *r*. For a 3- topic model for example, review *r* might be summarized as $\theta_r = [0.25, 0.5, 0.25]$.

However, this class of models tends to generate topics that reflect words that occur with high probability. Topics that consist of low- probability words do not get identified as distinct topics. This phenomenon is intrinsic to the estimation algorithm of the standard topic modeling framework. This poses a challenge to researchers who are interested in the less discussed, smaller topics (or "rare" topics") in a corpus. For example, managers and regulators might be interested in identifying emergent or nascent themes that are of specific interest to them. Prominent themes today and may not have been prominent in the past. This feature is likely to be important in examining markets undergoing changes in consumer tastes and competitive landscapes (such as ours and others). Recognizing this specific issue, recent research has focused on incorporating domain knowledge explicitly within these models to uncover rare topics. (Andrzejewski et al 2009) advocates specifying constraints that ensure certain words appear either together or not at all. (Hu et al, 2014) develop on this theme and allow researchers to interactively incorporate these constraints during the estimation process. While the former approach requires information on which words should and should not occur together in a topic the latter approach requires multiple researcher interventions.

In addition to the challenge of measuring rare topics, one might view topics as resulting from a combination of "factors". For example, a review for Uber maybe a combination of sentiment (positive or negative) and a specific attribute such as driving safety ("safety"). Managers, researchers and regulators may not only want to measure how much a theme was discussed, but also whether this discussion was favorable or not. This allows us to capture nuances on whether, for example, the increase in discussion of safety was due to an increase in favorable or unfavorable discussions. Valence in text has been used previously by (Lee and Bradlow, 2011), where the review content contained distinctly categorized pros and cons. We do not have a similar categorization in our data set and must explicitly model valence from within the model.

An alternative approach that jointly addresses the issue of rare topics and topics as a combination of factors involves using seed words (words that represent a topic with high probability). For example, (Lin et al, 2009) and (Tirunillai and Tellis, 2014) incorporate seed words to explicitly incorporate valence within the model and estimate each topic separately to represent positive and negative valence. These seed words are determined prior to estimation and are specified deterministically. This deterministic approach works very well for valence, since there are established lists of positive and negative words (Tausczik and Pennebaker, 2010). However, in our application we want to measure topics such as "safety", for which no such domain specific list of words is readily accessible. In contrast, (Jagarlamudi et al, 2012) explicitly model seed words in a probabilistic fashion and effectively estimate two topics for each seeded topic – one topic that models a distribution over the entire vocabulary and another topic that models a distribution over the seed words only. The challenge that emerges from these seeding approaches is the rapid increase in parameters. We must effectively estimate two

distributions over the vocabulary for each seeded topic. We address this in our approach (outlined below) by using a probabilistic specification of the priors for each of the distributions.

For instance, to incorporate two sentiments (positive and negative) for a set of *T* different attributes, one must effectively estimate $2 \times T$ topics, the Cartesian product of the two-level factor sentiment and a *T*-level factor, where each level is a topic(a $1 \times V$ vector with V parameters) of interest. Consequently we are required to estimate $2 \times T \times \mathrm{V}$ parameters, in addition to the usual document level parameters. As the number of factors increase, the associated model complexity increases exponentially. In order to address this problem, we adopt a recent variant of (Eisenstein et al, 2011) called Factorial LDA (Paul and Dredze, 2012). Rather than introduce parameters as a product of factors, FLDA incorporates all possible factor level combinations additively.

Another concern with standard LDA for empirical researchers is the issue of estimating duplicate topics. The model structure does not preclude effectively the same topic being estimated multiple times. Consequently, topics may need to be consolidated in a post processing step (Griffiths and Steyvers, 2004) using a clustering algorithm. In order to avoid a deterministic allocation to clusters in the post processing step, we exploit the additive structure of the Factorial LDA model to focus the model on the topics of our interest and to avoid the duplication of topics (described along with estimation in section 3.2.1).

Finally, it is likely that the observed review and rating are jointly determined at the time of writing. We model this intuition explicitly. Our approach is similar to (Mcauliffe and Blei, 2006). i.e., we assume that the same latent process generates both observed review text and the associated rating). However we use Bayesian methods in place of variational inference for

estimation in order to integrate two different modeling approaches – incorporation of ratings and modeling positive and negative aspects.

There are at least two alternative specifications to incorporate rating information. One option would be to incorporate rating information within the priors for the topic proportions for each review (Mimno and McCallum, 2008). Alternatively one could model the rating directly as a function of the text (Wang et al, 2010). Both approaches make a causal structural assumption - either text causes ratings or ratings cause text. Since the direction of causality is not ex-ante obvious to us, we jointly model text and rating as a function of latent process – topic proportions. This specification also potentially addresses the econometric issues of regressing ratings on topic proportions (see section 4 below).

We now turn to the model specification in our application. Specifically, we have two factors namely "Sentiment" and "Attributes" as shown in Table 2.2. These 8 attributes were identified from 12,440 media reports (including blogs, publications, online news and message boards) obtained from Factiva.com that contained the words "Uber technologies" as of March, 2015. We also estimate an additional topic with no priors that serves as a catch-all for other topics. Our approach allows us to specify virtually any attribute of interest subject to its presence in the corpus.

Table 2.2: Factors and Factor Components

Factor	Components
Sentiment	Positive, Negative
Attributes	Safety Fare Driving Payment Mode Cleanliness Tipping Wait Time Reservations

Each sentiment and attribute combination (or tuple) is treated as a topic. These tuples are mathematically the same as topics in the traditional LDA model with the additional feature that each tuple is a specific sentiment and attribute combination, for example ("positive", "reservations"). Each topic is modeled as a Dirichlet distribution (ϕ_t) over the vocabulary. In the rest of this paper, a topic refers to a combination of a valence and an attribute.

We assume the following generative process by which the textual data in each review and the associated rating is generated. The first step for generating word *i* in review *r* is to draw topic proportion θ_r from a Dirichlet distribution with an asymmetric *T*-dimensional parameter vector α_r. The second step is to choose z_{ri}, the topic assignment for the word *i* in review *r*. This is drawn from a categorical distribution with parameter θ_r. This is a particularly convenient choice of distributions as the Dirichlet distribution is conjugate to the categorical distribution, i.e. the posterior distribution of θ_r is also Dirichlet. Given the topic assignment z_{ri}, the word *i* in review *r* is drawn from a categorical distribution associated with the assigned topic. To exploit conjugacy, each topic distribution is also specified Dirichlet, that is $\phi_t \sim Dirichlet(\omega_t)$ where ω_t

is an asymmetric *V*-dimensional parameter vector. This process is repeated for each word *i* in review *r*. It ignores the order of words within a review, i.e. LDA is a "bag-of-words" model (Eliashberg, Hui and Zhang 2007; Netzer et al. 2012; Tirunillai and Tellis 2014) [38]. We define θ_r as the empirical topic proportions i.e. the number of words in the review assigned to topic *t* divided by the total number of words in the review. Given θ_r, the rating (y_r) assigned is modeled as a normal distribution with mean specified as a weighted combination of the topic proportions and a vector of weights β. The generative process for review *r* can be summarized as follows.

a) $\theta_r \mid \alpha_r \sim Dirichlet(\alpha_r)$

b) $z_{ri} \mid \theta_r \sim Categorical(\theta_r)$

c) $\phi = [\phi_1, \phi_2, \phi_3, ..., \phi_T]$

d) $\phi_t \mid \omega_t \sim Dirichlet(\omega_t)$

e) $w_{ri} \mid (z_{ri} = t), \phi \sim Categorical(\phi_t)$

f) $y_r \mid \theta_r, \beta \sim N(\theta_r^T \beta, \sigma)$

The researcher does not observe the topics, the (probabilistic) membership of words in each topic, the distribution of topics for each review, or the choice of topic that led to a specific choice of a word. The computational approach is to use the observed reviews and ratings to infer these distributions, i.e. to uncover the hidden topic structure that generated the observed set of reviews. The process described above defines a joint probability distribution over both the

[38] Modeling word order is computationally intensive and therefore rare in the computer science literature and has not yet been implemented in marketing literature. Such models have usually been limited to incorporating bi-grams (word pairs) or tri-grams (a triplet of words). Given the computational burden posed by estimating the hyperparameters, we chose to retain the standard "bag-of-words" assumption.

observed and hidden random variables. We use this joint distribution to compute the conditional (or posterior) distribution of the hidden variables given the observed reviews and words.

In contrast to the traditional LDA model where the Dirichlet prior for each ϕ_t is a symmetric hyper-parameter vector of size *V*, in Factorial LDA, for each ϕ_t we have a prior for each word *v* specified as follows:

$$\omega_v^t = \exp(\omega_b + \omega_0^v + \sum_k \omega_{t(k)}^{v,k}) \quad (1)$$

The terms in the exponential are as follows. ω_b is a scalar bias term common for all words in the vocabulary. ω_0^v is a weight specific to the word over the entire corpus and can be thought of as an index of how likely it is to observe *v* in the corpus. As this weight is not topic specific, this term may also be thought of as the "background" weight. The sign and magnitude of the weight affects the likelihood of observing word *v*, irrespective of topic. The vector describing these weights for each word in the vocabulary is termed the "background distribution". As we will see in section 5, excluding these background weights leads to a significantly worse fit for the model. In the last term in (1), *t(k)* indicates the level associated with factor *k* in topic *t*. Hence, $\Sigma_k \omega_{t(k)}^{v,k}$ is a sum over the weight of the word in the levels present (from the two factors) in the topic. To illustrate this consider the word "driver" with weight 0.2 in the sentiment "positive" and weight 0.3 in the attribute "reservations" . In the topic ("positive", "reservations"), the term $\Sigma_k \omega_{t(k)}^{v,k}$ for the word "driver" is equal to 0.2 +0.3=0.5.

Introducing this additive term ensures that both sentiments for each attribute will be estimated. Unlike the applications in Lin et al (2009) and Tirunillai and Tellis (2014), where not all attribute and valence combinations emerge as topics (and that is not the focus), our analysis

requires us to measure both sentiments for every attribute (even if one sentiment is discussed very little). Topics sharing the same factor level also share the same weights for words in that factor level. For example, the topic (positive, reservations) and (negative, reservations) both share reservations as a common attribute. The topic weights for these two topics as computed in (1) will only differ on the sentiment aspect, but not on the "reservations" attribute. This ensures that the =positive topic for "reservations" and the "negative" topic for "reservations" both ascribe the same weights to reservation related words. This is important to ensure that either valence is associated with a consistent definition of the aspect reservation.

Additionally, in place of a separate weight for a word in all *T* topics (i.e. $T \times V$ weights), we need only specify *A* (number of attribute levels) +2(sentiment levels) + 1(background) = $(A+3) \times V$ weights to generate topic specific weights. A+3 is generally much smaller than *T*.

Similarly, the topic distribution within a review is modeled as a categorical distribution with a Dirichlet prior such that for each θ_r we have a *T*-dimensional asymmetric prior for each review *d* specified as follows:

$$\alpha_r^t = \exp(\alpha_b + \sum_k \alpha_{t(k)}^{R,k} + \sum_k \alpha_{t(k)}^{r,k}) \quad (2)$$

Similar to ω_b in (1), α_b is a scalar bias term common for all reviews in the corpus. $\Sigma_k \alpha_{t(k)}^{R,k}$ is a weight term specific to the topic shared with the entire corpus *D* obtained by summing over the weights of the levels present in the topic. Extending the previous example, suppose that the topic ("positive", "reservations") has corpus level weights (shared by all reviews) of 0.6 and 0.2 for factor levels "positive" and "reservations" respectively. Consequently, $\Sigma_k \alpha_{t(k)}^{R,k}$ is equal to 0.8. Finally, unlike the second term in (2), $\Sigma_k \alpha_{t(k)}^{r,k}$ is a sum over the *review specific* weights for the levels present (from the two factors) in the topic. Comparing our model to existing work,

McAuliffe and Blei, (2006) incorporate rating in the LDA model using variational inference as the estimation procedure, whereas Paul and Dredze (2012) employ Bayesian estimation. The joint specification of factors and accounting for review rating in a unified Bayesian estimation framework is novel to the current literature in Marketing and Computer Science.

3.2 Estimating measures for multiple aspects of consumer experience

We first describe how we estimate each parameter, and then our method of seeding. We estimate the hyperparameter vectors α_r and ω_t, the review level topic proportions θ_r, the vector of word level assignments of topics z_r and the topic level parameter ϕ_t. Assuming reviews are conditionally independent and identically distributed, the likelihood of the data conditional on the hyperparameters is calculated as follows:

$$(L \mid \alpha, \omega) = \prod_{r=1}^{R} \int_{\phi} \int_{\theta} p(y_r \mid \theta_r, \beta)\, p(\theta_r \mid \alpha_r) \times p(\phi \mid \omega) \prod_{i=1}^{n_r} \sum_{t=1}^{T} \prod_{v=1}^{V} [\phi_{t,v} \times \theta_r^t]^{I(w_{ir}=v)}\, d\phi d\theta \qquad (3)$$

Where *I* is the indicator function, and α_r is a *T*-dimensional vector with element defined in (2) and ω is a $T \times V$ matrix with each element in row t defined in (1). We face two estimation challenges: this function does not have a closed-formed analytical solution due to the product term involving $\phi_{t,v}$ and θ_r^t (Dickey 1983), and the dimensionality of our parameter space is very high (a common feature of problems associated with "big data"). The dimensionality problem is owing to the large number of unique words in the corpus (*V*), the potentially large number of tuples and the large number of reviews (note that θ_r is review specific). Following the computer science literature (Griffiths and Steyvers 2004), instead of estimating ϕ_t or θ_r as parameters we first estimate the posterior distribution of the assignment of words to topics,

$P(z \mid w, y)$ based on the equation $P(z \mid w, y) = P(z, w, y) / \sum_z P(z, w, y)$ .The numerator of the right hand side of this equation can be factorized and simplified as $P(z, w, y) = P(w, y \mid z)P(z)$. We now turn our attention to $P(w, y \mid z)$ and $P(z)$. Given the conjugacy between the distribution of observing word *v* given tuple *t* (assumed to be a Categorical Distribution as described in section 2.1) and the Dirichlet prior ($\phi_t \mid \omega_t \sim Dirichlet(\omega_t), \forall t$), the posterior distribution of $P(w, y \mid z)$ is as follows (Griffiths and Steyvers 2004):

$$P(w, y \mid z) = \prod_{t \in T} \left[\left[\frac{\Gamma(\Sigma_v \omega_v^t)}{\Pi_v \Gamma(\omega_v^t)} \right] \left[\frac{\prod_{v \in V} \Gamma(\omega_v^t + n_t^v)}{\Gamma(\Sigma_v \omega_v^t + n_t)} \right] \right] \times N(\hat{\theta}_r^{\top} \beta, \sigma) \quad (4)$$

$\Gamma(.)$ is the standard gamma function. Here, n_t^v is the number of times the word *v* in the vocabulary is assigned to topic *t* in the corpus; n_t is the number of words in the corpus which are assigned to topic *t.* Similarly, the conjugacy between the topic assigned to each word in a review (assumed to be a Categorical Distribution) and the Dirichlet prior ($\theta_r \mid \alpha_r \sim Dirichlet(\alpha_r)$) yields review specific topic assignments. *N* indicates a normal distribution for ratings with mean $\hat{\theta}_r^{\top} \beta$ and variance σ . These tuple assignments are conditionally independent across reviews and can be multiplied to yield:

$$P(z) = \prod_{r \in R} \left[\left[\frac{\Gamma(\Sigma_t \alpha_r^t)}{\Pi_t \Gamma(\alpha_r^t)} \right] \left[\frac{\prod_{t \in T} \Gamma(\alpha_r^t + n_r^t)}{\Gamma(\Sigma_t \alpha_r^t + n_r)} \right] \right] \quad (5)$$

n_r^t is the number of words in review *d* assigned to topic *t* and n_r is the length in words of review *r*. Note that while $P(z, w, y)$ may be factored and computed as described above, $\sum_z P(z, w, y)$ cannot be computed directly because it does not factorize and involves T^N terms, which is again computationally challenging. We thus adopt an MCMC approach which relies on Gibbs sampling of the latent tuple assignment variable $\boldsymbol{Z}$ (Griffiths and Steyvers 2004).[39] The full conditional distribution of $\boldsymbol{Z}$ is free of ϕ_t and θ_r, enabling us to estimate ϕ_k and θ_r by averaging the means of the posterior Dirichlet distributions across iterations from a single MCMC chain[40]. The Gibb's sampling scheme for (3) can be derived from (4) and (5) as follows:

$$P(z_i = j \mid z_{-i}, w, y)_r \propto \frac{n_{j,-i}^w + \omega_w^j}{\Sigma_v \omega_v^j + n_j - 1} \frac{n_{j,-i}^r + \alpha_r^j}{\Sigma_t \alpha_r^t + n^d - 1} \quad \frac{1}{\sigma} exp - \frac{(y_r - \hat{\theta}_r \beta)^2}{2\sigma^2} \tag{6}$$

In each Gibbs iteration, the probability of assigning topic *j* to the word *w* in position *i* in review *r* is proportional to the product of three terms. The first term is the number of times word *w* was assigned topic *j* as a proportion of the total words assigned to topic *j* (adjusted for smoothing) in the entire corpus. The second term is the number of words in review *r* assigned to topic *t* as a proportion of the total number of words in the review *r* (adjusted for smoothing). Any topic assignment is thus a function of both the corpus and the review. The last term is the error in predicting the correct rating for a review given the empirical topic proportions β, and σ. As topic proportions change with each iteration, the error estimate is revised and places more or less

[39] Variational Inference (VI) methods are also commonly employed in computer science and statistics for large-scale problems with intractable integrals. Whereas Monte Carlo methods provide numerical approximations of the exact posterior by sampling, VI methods provide a locally optimal but precise analytical solution to an approximation of the posterior. We estimated the model using a VI method and obtained almost identical results with comparable computational speed. We chose Monte Carlo methods since they are more common in the marketing literature.

[40] For example, $P(\phi_t \mid z, w) = Dirichlet(\omega_1^t + n_t^1, ..., \omega_v^t + n_t^v)$ and the estimated mean vector of this distribution from a single MCMC iteration is $(\frac{\omega_1^t + n_k^1}{\Sigma_v \omega_v^t + n_t}, ..., \frac{\omega_v^t + n_k^v}{\Sigma_v \omega_v^t + n_t})$.

confidence on the next tuple assignment in a Bayesian estimation framework. While Mcauliffe and Blei (2008) specify a similar approach using Variational Inference methods, this particular formulation and interpretation is novel to the best of our knowledge.

We now describe how the terms comprising ω_v^t namely ω_b, ω_0^v and $\omega_{t(k)}^{v,k}$ are estimated[41]. Each of these parameters is assumed to have a prior distribution $N(0, \sigma^2 I)$ [42]. For instance, the joint probability of *w*, *z* and ω_b retaining only terms containing ω_b can be written as follows:

$$L_{\omega_b} \propto P(w, y, z \mid \{\alpha_d^t, \omega_w^t : \forall t, d, w\}) \times \quad N(\omega_b \mid 0, \sigma^2) \tag{7a}$$

$$L_{\omega_b} \propto \prod_{t \in T} \left[\left[\frac{\Gamma(\Sigma_v \omega_v^t)}{\Pi_v \Gamma(\omega_v^t)} \right] \left[\frac{\prod_{v \in V} \Gamma(\omega_v^t + n_t^v)}{\Gamma(\Sigma_v \omega_v^t + n_t)} \right] \right] \times N(\omega_b \mid 0, \sigma^2) \tag{7b}$$

We maximize the likelihood of the topic assignments for each word in the corpus with respect to the parameter ω_b using gradient ascent. Consequently, the estimation algorithm consists of alternating between the MCMC iterations using (6) and the stochastic EM step where (7b) is maximized with respect to the parameter ω_b. A similar principle applies for ω_0^v and $\omega_{t(k)}^{v,k}$.Details for the estimation of this parameter appears in Online Appendix 1. Parameters β and σ are estimated in the EM step using maximum likelihood.

The estimation algorithm is implemented in Java[43] and was modified to infer estimates on new data (San Jose reviews) and to generate output relevant for our analysis. The MCMC chain ran for 30,000 iterations, with the first 2,000 iterations for "burn-in". We estimate all

[41] A similar approach is used to estimate α_d^t. We focus on ω_v^t as it helps outline both the parameter estimation procedure and the seeding procedure.
[42] Directionally, results are not sensitive to the choice of variance for this prior. We used different variance values (0.5, 5, 25), but our findings remain directionally the same.
[43] http://cs.jhu.edu/~mpaul/downloads/flda.php

hyperparameters every 100th iteration using the gradient ascent algorithm. The last 5,000 iterations (using a sampling lag of 10) yielded 500 samples that were used to compute the moments of the posterior parameter distributions.

3.2.1 Seeding

Seeding in this framework requires the researcher to specify the attributes of interest, the seed words for each attribute and the mean for the prior distribution. In our application, we considered the following topics (see Table 2.2) based on the number of news articles mentioning an attribute in the context of Uber. The seed words for each attribute are synonyms or plural/tense variants of the root word. Positive and negative words listed in (Tausczik and Pennebaker, 2010) are used as seeds for positive and negative aspects respectively.

Table 2.3: Attributes and Seed Words

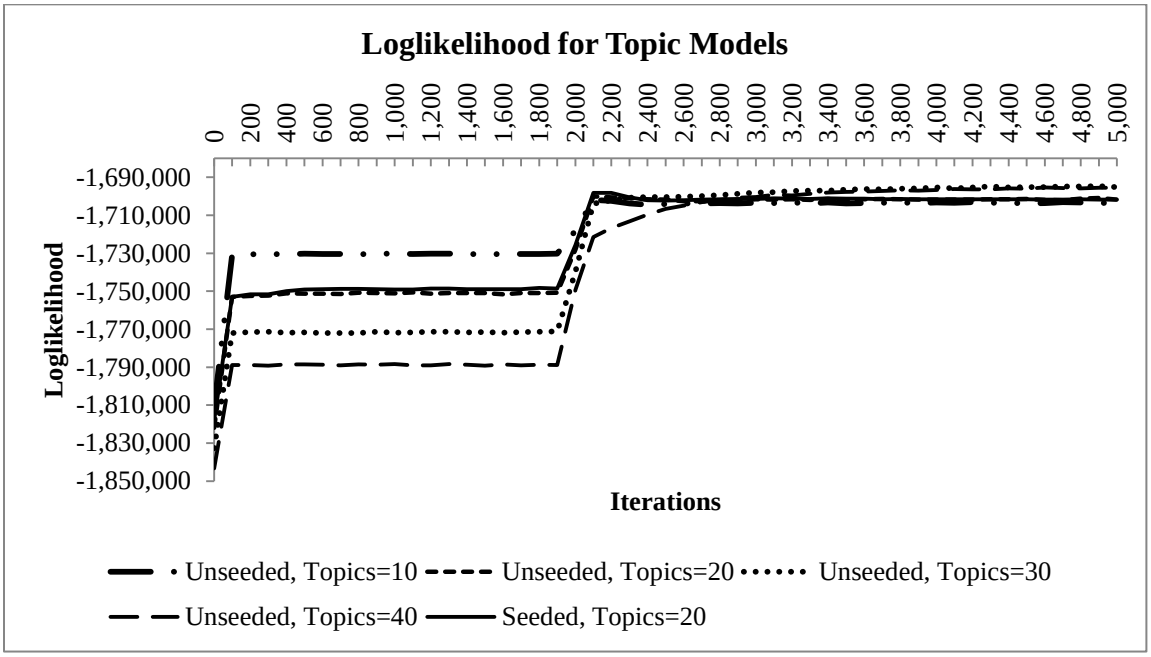
Index	Attributes	Publications mentioning both "Uber" and attribute[44]	Seed Words
	Uber	11,906	
1	Safety	3,428	Safe, safely, safety
2	Fare	1,814	fee, fare, fares
3	Driving	1,431	driving, drive, driver
4	Payment Mode	1,003	card, credit, cash
5	Cleanliness	366	clean
6	Tipping	294	gratuity, tip,%
7	Wait Time	47	wait ,waiting, waited
8	Reservations	17	reservation ,reservations, call

The prior mean weight specified is common for all seed words and is specified by comparing the loss in fit. The prior mean specified for a seed word associated with a particular topic is 5.0. All other seed words that we seek to separate have a prior mean of -5.0 specified. This addresses the scenario where, for example, a topic seeded with the word "fare" is more

[44] Source: Factiva.com

likely to contain "tip" with a high probability than other words. By assigning a negative weight that to "tip" in the "fare" topic, we make "tip" much less likely to appear in that topic. All other words have a zero mean prior. To ensure that these priors are very diffuse, the variance for all priors is set to 10. As can be observed from Figure 2.2, our specification has a very minor loss in fit in comparison to the completely unseeded model, where all priors are zero mean priors. Interestingly, increasing the number of topics in the model offers only a marginal improvement in fit, as measured by log likelihood of the observed data. Specifically, Figure 2.3 shows that the improvement in fit by increasing the number of topics to 30 leads to a small improvement in fit.

Figure 2.3



Fit comparisons

3.3 Examining measures of consumer experience

For each of the sentiment-attribute topics, we address the following questions in an econometric framework. First, we examine the effect of the sentiment-attribute topic proportions on the ratings. This serves as a face validity test to check whether the latent topics correspond with the variation in overall review ratings. Given that the numerical rating is an indicator of the user experience, we expect the review content to be related to this score. We estimate the following regression model:

$$rating_r = \sum_{t=1}^{T} \gamma_t \theta_{tr} + controls + \epsilon_r \quad (8)$$

Here, θ_{tr} is the proportion of topic *t* in review *r*. Note that in the LDA specification, θ_{tr} is assumed to explain overall review rating. The data are the reviews for Non-Uber taxi services in San Francisco. $rating_r$ is the rating of the review. In order to control for unobserved heterogeneity due to firm specific variation, we incorporate firm fixed effects. However, since the volume of firms is very large, we categorize firms into quartiles based on the volume of reviews for each firm. This serves as an approximation for firm size and the fixed effect for each quartile is included in the model. We also wish to account for time varying effects that may explain variation of ratings over time. For example, ratings may generally decline over time because of increased product familiarity. We incorporate year-quarter fixed effects to control for these time varying effects. Finally, review length maybe correlated with over-all rating. Longer or shorter reviews maybe more favorably rated. To account for this we include review length as a covariate. Our coefficients of interest are the topic proportion coefficients. Specifically, we will want to see whether these coefficients have intuitive appeal. For example, positive topics are

expected to contribute to higher ratings and consequently have larger coefficients than negative topics in general.

Our substantive interest lies in whether Uber's entry made certain topics more or less salient in reviews of traditional operators. In order to examine this question, we employ the following regression specification:

$$\theta_{tr} = \gamma_{t,const} + \gamma_{t,post}\, post_r + controls + \epsilon_{rt} \quad (9)$$

$post_r$ is the dummy variable which takes value 1 for reviews posted on or after July, 2010 , the date of Uber's entry in the city of San Francisco[45], and 0 otherwise. As discussed above, controls include review length and fixed effects for firms and year-quarters. We estimate the model on non-Uber reviews posted over an 8-year period (i.e. 32 quarters) starting July 1, 2006 and ending June 1, 2014. This period was chosen so that the duration of time (in which reviews were posted) before and after July, 2010 (date of Uber's entry), is the same. Error terms are assumed IID and normally distributed. All parameters are tuple-specific.

We would also like to understand whether Uber's entry changed (for traditionals) the sensitivity of the overall consumer experience to each of the attributes. We use the following regression specification to examine this question.

$$rating_r = \gamma_{t,const} + \gamma_{t,post}\, post_r + \sum_{t=1}^{T-1} \gamma_{t,3}\theta_{tr} + \sum_{t=1}^{T-1} \gamma_{t,4} post_r \quad \times \quad \theta_{tr} + controls + \epsilon_{rt} \quad (10)$$

The data used for (9) is used to estimate (10). Controls review length and fixed effects for firms and year-quarters for similar reasons as argued above for equation(8). The coefficients of interest are $\gamma_{t,4}$, for all t, as they indicate whether there was change in sensitivity to attributes post Uber's entry.

[45] http://www.tc.umn.edu/~ssen/IDSC6050/Case15/Group15_index.html

While changes in amount of discussion of topics and sensitivity of the overall experience to these topics for traditionals is our primary objective, we are also interested in whether these experience attributes set Uber apart from traditionals. We next examine whether there are specific sentiment-attribute combinations that are discussed differently in reviews for Uber versus reviews for traditionals. The model we estimate is:

$$\theta_{tr} = \gamma_{t,const} + \gamma_{t,uber} uber_r + controls + \epsilon_{rt} \quad (11)$$

In this specification, $uber_r$ is a dummy variable that takes value 1 when the review is for Uber and 0 otherwise. As controls, we also include year and quarter specific dummy variables to account for time period specific shocks, fixed effects for firm size[46], and review length. In order to compare Uber and non-Uber reviews we restrict the data set to reviews posted from June, 2010 (date of Uber's entry) to November, 2014 (the date till which we have data collected).

A related question is whether the overall consumption experience (as approximated by the review rating) for Uber is more (or less) sensitive to certain topics in comparison to traditionals.

$$rating_r = \gamma_{t,const} + \gamma_{t,post} uber_r + \sum_{t=1}^{T-1} \gamma_{t,3} \theta_{tr} + \sum_{t=1}^{T-1} \gamma_{t,4} uber_r \times \theta_{tr} + controls + \epsilon_{rt} \quad (12)$$

Using the same data set as (11), we focus on the coefficients $\gamma_{t,4}$ to understand the differences in sensitivity between Uber and traditionals for each attribute.

We estimate Equations 8 to 12 within the standard regression framework and report the results below.

[46] Firms are categorized into 4 quartiles based on the volume of reviews for each firm. This serves as a proxy for firm size.

Section 4: Analysis

Table 2.4 lists the top words by weight for the background distribution, the sentiment distributions and the attribute distributions. The background distribution appears to serve its purpose by focusing on the most commonly used words across reviews such as time, driver, experience, back, make, service, find, good, ride and day. A visual inspection indicates that our seeding approach has encouraged the attributes to remain quite distinct. The cosine distance averaged over all pairs of background, attribute and sentiment distributions is 0.12, suggesting that the attributes are indeed quite distinct.

Table 2.4: Attributes and Sentiments Estimated

ID	Name	Top 10 words in decreasing order of weights belonging to each factor level and for the Background Distribtuion.
1	Background	hope,time,experience,driver,make,tip,company,service,back, phone
		Sentiment
2	Positive	Good,nice,great,awesome,excellent,time,drivers,times,driver make
3	Negative	Bad,worst,horrible,sucks,horrendous,atrocious,back,driver,time,didn't
		Attributes
5	Safety	Safely,safe,careful,safety,taxi,city,night,ride,sf back
6	Fare	Fares,fare,fee,im,people,time,work,ive,guys youre
7	Driver Driving	driving,drive,stop,turn,drove,road,freeway,front,pulled,lane
8	Credit Card	cash,card,receipt,office,cards,found,wallet,left,cell,machine
9	Clean	clean,friendly,professional,recommend,reliable,limo,cleanliness,courteous,highly ,cars
10	Tipping	tip,$,%,cost,extra,fee,pay,miles,included,gratuity
11	Wait Time	Wait,waiting,waited,time,airport,driver,sfo,pick,flight minutes
13	Reservations	shuttle,reserve,flight,supershuttle,reservation,online,reservations,van,airport,bart

Table 2.5 suggests that the relationship between the estimated topics and overall ratings (equation 8) is as expected. The sign of the positive and negative attributes are in the expected

directions[47]. For example, (positive, safe) (M=1.86, S.E. = 0.11) has a greater incremental effect on rating than (negative, safe) (M=0.71, S.E.=0.27). In general positive topics have a greater positive effect than negative topics on the over-all rating.

We now present the analysis for each aspect of experience for traditionals (using equations 9 and 10) and for Uber (using equations 11 and 12). In the interest of clarity, we collect results from tables 2.6 through 2.9 for each aspect of experience below.

Table 2.5: Regression of Rating on Topics[48]

Dependent variable is rating

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	0.07	0.44	0.15	0.88	
rev_len	0.00	0.00	1.25	0.21	
p_fare	**2.30**	**0.10**	**23.82**	**0.00**	***
n_fare	**2.06**	**0.17**	**12.96**	**0.00**	***
p_tip	0.10	0.10	0.98	0.33	
n_tip	**-0.74**	**0.09**	**-8.29**	**0.00**	***
p_wait	**1.93**	**0.10**	**19.67**	**0.00**	***
n_wait	**0.24**	**0.09**	**2.74**	**0.01**	**
p_reserve	**1.87**	**0.09**	**21.74**	**0.00**	***
n_reserve	-0.01	0.08	-0.16	0.88	
p_clean	**2.05**	**0.09**	**23.74**	**0.00**	***
n_clean	**1.26**	**0.18**	**7.17**	**0.00**	***
p_safe	**1.86**	**0.11**	**16.82**	**0.00**	***
n_safe	**0.71**	**0.27**	**2.66**	**0.01**	**
p_card	0.24	0.21	1.19	0.23	
n_card	0.06	0.09	0.69	0.49	
p_drive	**1.77**	**0.11**	**16.56**	**0.00**	***
n_drive	**-0.41**	**0.06**	**-6.37**	**0.00**	***

*** p<0.01, ** p<0.05, *p<0.10

Adjusted R Square : 0.71

[47] The naming convention for the variables is as follows. 'p_' indicates the positive valence of an aspect and 'n_' the negative valence of the aspect.

[48] Coefficients for controls are omitted in all tables in order to conserve space but are available on request.

Table 2.6: Change in discussion of attributes for Traditionals(Equation 9)

Dependent variable is topic proportions

Dependent Variable	Estimate of 'Post' Coefficient	Std. Error	t value	Pr(>|t|)	Significance
p_fare	**1.06**	**0.28**	**3.71**	**0.00**	***
n_fare	**0.38**	**0.15**	**2.57**	**0.01**	**
p_tip	0.44	0.26	1.66	0.10	*
n_tip	**0.98**	**0.30**	**3.27**	**0.00**	***
p_wait	**1.01**	**0.26**	**3.92**	**0.00**	***
n_wait	**-0.72**	**0.30**	**-2.42**	**0.02**	**
p_reserve	**-1.15**	**0.30**	**-3.80**	**0.00**	***
n_reserve	**-1.62**	**0.36**	**-4.45**	**0.00**	***
p_clean	0.03	0.32	0.10	0.92	
n_clean	-0.05	0.14	-0.32	0.75	
p_safe	-0.05	0.23	-0.23	0.82	
n_safe	-0.14	0.09	-1.57	0.12	
p_card	**0.50**	**0.12**	**4.04**	**0.00**	***
n_card	0.32	0.29	1.11	0.27	
p_drive	**-1.61**	**0.25**	**-6.53**	**0.00**	***
n_drive	-0.89	0.46	-1.95	0.05	**
p_other	**0.97**	**0.25**	**3.82**	**0.00**	***
n_other	0.53	0.39	1.36	0.18	

***** p<0.01, ** p<0.05, *p<0.10**

All regressions are significant, with adjusted r-squares ranging between 0.04 and 0.51

Table 2.7: Change in sensitivity of overall-ratings to attributes for Traditionals

(Equation 10)

Dependent variable is rating

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-0.14	0.45	-0.32	0.75	
Post	-0.19	0.17	-1.16	0.25	
rev_len	0.00	0.00	0.41	0.68	
p_fare	**2.47**	**0.25**	**9.85**	**0.00**	***
n_fare	**3.03**	**0.36**	**8.33**	**0.00**	***
p_tip	0.18	0.26	0.70	0.48	
n_tip	**-0.47**	**0.19**	**-2.54**	**0.01**	**
p_wait	**1.90**	**0.24**	**7.95**	**0.00**	***
n_wait	0.26	0.15	1.71	0.09	*
p_reserve	**2.02**	**0.14**	**14.74**	**0.00**	***
n_reserve	0.14	0.14	1.01	0.31	
p_clean	**1.82**	**0.17**	**10.39**	**0.00**	***
n_clean	**2.74**	**0.34**	**8.01**	**0.00**	***
p_safe	**1.93**	**0.21**	**9.09**	**0.00**	***
n_safe	**1.15**	**0.51**	**2.24**	**0.03**	**
p_card	0.24	0.62	0.39	0.70	
n_card	**0.49**	**0.20**	**2.41**	**0.02**	**
p_drive	**2.00**	**0.16**	**12.38**	**0.00**	***
n_drive	**-0.30**	**0.12**	**-2.44**	**0.01**	**
Post:rev_len	0.00	0.00	0.31	0.75	
Post:p_fare	-0.20	0.26	-0.78	0.44	
Post:n_fare	**-0.98**	**0.41**	**-2.37**	**0.02**	**
Post:p_tip	-0.12	0.28	-0.44	0.66	
Post:n_tip	-0.32	0.21	-1.53	0.13	
Post:p_wait	0.02	0.26	0.09	0.93	
Post:n_wait	0.00	0.16	-0.01	0.99	
Post:p_reserve	-0.19	0.17	-1.11	0.27	
Post:n_reserve	-0.21	0.16	-1.34	0.18	
Post:p_clean	0.28	0.19	1.49	0.14	
Post:n_clean	**-1.97**	**0.40**	**-4.94**	**0.00**	***
Post:p_safe	-0.10	0.23	-0.42	0.67	
Post:n_safe	-0.56	0.60	-0.93	0.35	
Post:p_card	-0.01	0.65	-0.01	0.99	
Post:n_card	**-0.54**	**0.23**	**-2.37**	**0.02**	**
Post:p_drive	-0.33	0.20	-1.68	0.09	*
Post:n_drive	-0.13	0.14	-0.89	0.37	

***** p<0.01, ** p<0.05, *p<0.10**

Adjusted R Square: 0.72

Table 2.8: Difference in discussion of attributes between Uber and Traditionals

(Equation 11)

Dependent variable is topic proportions

Dependent Variable	Estimate of 'Uber' Coefficient	Std. Error	t value	Pr(>|t|)	Significance
p_fare	**-4.51**	**0.49**	**-9.17**	**0.00**	***
n_fare	**-0.95**	**0.21**	**-4.57**	**0.00**	***
p_tip	**1.41**	**0.49**	**2.90**	**0.00**	***
n_tip	**-1.41**	**0.41**	**-3.46**	**0.00**	***
p_wait	**-4.20**	**0.39**	**-10.63**	**0.00**	***
n_wait	**-8.91**	**0.63**	**-14.16**	**0.00**	***
p_reserve	-0.38	0.44	-0.87	0.38	
n_reserve	**-3.66**	**0.59**	**-6.21**	**0.00**	***
p_clean	0.38	0.49	0.78	0.44	
n_clean	**0.63**	**0.20**	**3.12**	**0.00**	***
p_safe	**-3.33**	**0.46**	**-7.19**	**0.00**	***
n_safe	-0.24	0.12	-1.93	0.05	**
p_card	**24.70**	**0.35**	**69.88**	**0.00**	***
n_card	**6.78**	**0.44**	**15.59**	**0.00**	***
p_drive	**0.92**	**0.41**	**2.26**	**0.02**	**
n_drive	**-1.48**	**0.65**	**-2.26**	**0.02**	**
p_other	**-1.65**	**0.38**	**-4.41**	**0.00**	***
n_other	**-4.10**	**0.53**	**-7.75**	**0.00**	***

***** p<0.01, ** p<0.05, *p<0.10**

All regressions are significant, with adjusted r-squares ranging between 0.02 and 0.29

Table 2.9: Difference in sensitivity of overall-ratings to attributes between Uber and Traditionals
(Equation 12)
Dependent variable is rating

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-0.30	0.42	-0.72	0.47	
uber_c	-0.51	0.48	-1.05	0.29	
rev_len	0.00	0.00	1.25	0.21	
p_fare	**2.26**	**0.10**	**21.77**	**0.00**	***
n_fare	**2.08**	**0.20**	**10.43**	**0.00**	***
p_tip	0.07	0.11	0.67	0.50	
n_tip	**-0.80**	**0.10**	**-7.82**	**0.00**	***
p_wait	**1.93**	**0.11**	**17.94**	**0.00**	***
n_wait	**0.28**	**0.10**	**2.72**	**0.01**	**
p_reserve	**1.75**	**0.11**	**15.96**	**0.00**	***
n_reserve	-0.08	0.10	-0.81	0.42	
p_clean	**2.09**	**0.10**	**21.74**	**0.00**	***
n_clean	**0.75**	**0.21**	**3.66**	**0.00**	***
p_safe	**1.82**	**0.12**	**14.89**	**0.00**	***
n_safe	0.61	0.31	1.94	0.05	**
p_card	0.35	0.22	1.63	0.10	*
n_card	-0.08	0.10	-0.77	0.44	
p_drive	**1.65**	**0.14**	**11.72**	**0.00**	***
n_drive	**-0.42**	**0.07**	**-5.69**	**0.00**	***
uber_c:rev_len	0.00	0.00	1.22	0.22	
uber_c:p_fare	**1.86**	**0.48**	**3.91**	**0.00**	***
uber_c:n_fare	0.71	0.93	0.76	0.45	
uber_c:p_tip	0.37	0.37	1.00	0.32	
uber_c:n_tip	0.31	0.46	0.66	0.51	
uber_c:p_wait	-0.26	0.46	-0.57	0.57	
uber_c:n_wait	0.36	0.80	0.45	0.65	
uber_c:p_reserve	**1.22**	**0.43**	**2.81**	**0.00**	***
uber_c:n_reserve	**3.53**	**0.61**	**5.76**	**0.00**	***
uber_c:p_clean	**1.43**	**0.34**	**4.21**	**0.00**	***
uber_c:n_clean	0.95	0.57	1.67	0.09	*
uber_c:p_safe	0.15	0.81	0.18	0.85	
uber_c:n_safe	**3.53**	**1.20**	**2.93**	**0.00**	***
uber_c:p_card	**1.53**	**0.35**	**4.39**	**0.00**	***
uber_c:n_card	0.58	0.32	1.81	0.07	*
uber_c:p_drive	**0.90**	**0.38**	**2.33**	**0.02**	**
uber_c:n_drive	0.66	0.36	1.85	0.06	*

*** p<0.01, ** p<0.05, *p<0.10
Adjusted R Square: 0.69

While results in tables 2.6 and 2.7 have a causal interpretation in additional analyses (reported below), results in tables 2.8 and 2.9 are descriptive.

4.1 Driving Experience

Traditional Taxi Reviews Favorable discussion of driving experience declines post Uber's entry in traditional taxi reviews (M = -1.61, t = -6.53), while the amount of unfavorable discussion of this aspect remains unchanged. Interestingly, sensitivity of the overall experience (as measured by the overall review rating) to driving experience (both favorable and unfavorable) remains unchanged with Uber's entry.

Uber Reviews In contrast to traditional taxi reviews, favorable driving experiences are discussed significantly more in Uber reviews (M= 0.92, t =2.26). Unfavorable driving experiences are discussed less in Uber reviews (M = -1.48, t = -2.27) as well. Further, favorable discussion of driving experience benefits Uber more (in terms of overall experience) than traditionals (M =0.90, t=2.33). Unfavorable discussion of this aspect appears to affect both Uber and traditionals similarly (in terms of overall experience).

Implications Traditional cab services have claimed that driver training improves safety[49] and the absence of similar procedures at Uber may be potentially harmful to consumers. This is not quite reflected in the reviews of ride experience. Traditional cab services may wish to adopt other means of increasing salience of this aspect or choose other aspects to emphasize differences from Uber.

4.2 Safety

Traditional Taxi Reviews Both favorable and unfavorable discussions of safety remain unchanged post Uber's entry. Interestingly, sensitivity of the overall experience- to safety (both favorable and unfavorable) remains unchanged with Uber's entry.

[49] http://www.sftwa.org/white_paper

Uber Reviews In contrast to traditional taxi reviews, favorable discussion of safety is significantly less in Uber reviews (M= -3.33, t =-7.79). Unfavorable discussion of safety is also less in Uber reviews (M = -0.24, t = -1.93) as well. Further, overall experience is equally sensitive to safety for both Uber and traditionals. Yet unfavorable discussion of safety hurts overall experience for Uber less than it does for traditionals (M =3.53, t=2.93).

Implications Safety is related to driving experience and is presented as the number one concern by traditional cab services[50]. Uber and traditionals may both be similarly affected by negative incidents of safety.

4.3 Reservations

Traditional Taxi Reviews Favorable discussion of reservations declines post Uber's entry (M = -1.15, t = -3.80), as does unfavorable discussion of reservations (M = -1.62, t= -4.45). Interestingly, sensitivity of the overall experience to reservations (both favorable and unfavorable) remains unchanged with Uber's entry.

Uber Reviews It is interesting to observe that favorable discussion of reservations is similar across Uber and traditional reviews. However, unfavorable reservation experiences are discussed more in traditional reviews (M = -3.66, t = -6.21). Further, overall experience for traditionals is hurt more than overall experience for Uber by both favorable (M=1.22, t=2.81) and unfavorable (M=3.53, t=5.76) reservation experiences.

Implications Despite adoption of software applications such as Flywheel and Curb[51] to offer reservation convenience, the reservation experience with Uber is viewed more favorably than with traditional cab services. This might be a potential area of further investigation for managers of traditional cab services.

[50] http://www.sftwa.org/white_paper
[51] http://www.sftwa.org/white_paper

4.4 Fares

Traditional Taxi Reviews Favorable discussion of fares increases post Uber's entry (M = 1.06, t = 3.71), as does unfavorable discussion of fares (M = 0.38, t= 2.57). Sensitivity of the overall experience to favorable fares remains unchanged with Uber's entry, whereas unfavorable fares hurts the overall experience less (M = -0.98, t= -2.37) post Uber's entry.

Uber Reviews Fares (both favorable and unfavorable) are discussed less (M = -4.51, t = -9.17 and M = -0.95, t = -4.57 respectively) in Uber reviews in comparison to reviews for traditionals. Further, overall experience for Uber benefits from the favorable discussion of fares (M=1.86, t=3.91). Uber is as sensitive as traditionals for unfavorable fares.

Implications Uber claims to generally offer lower fares (despite its dynamic pricing ('Surge' pricing) model. Our analysis suggests that there are consumers who might favor a predictable fare in contrast to dynamic pricing. It should be noted that traditional cab services are regulated and do not get to officially set their own prices. So the rise in favorable experience may not necessarily have to do with tangible advantages in pricing. Further, Uber is as sensitive as traditionals for unfavorable fares – which may especially matter for periods when 'Surge' pricing is in effect.

4.5 Tipping

Traditional Taxi Reviews Favorable discussion of tips remains unchanged post Uber's entry. However, unfavorable discussion of tips increases (M = 0.98, t= 3.27) post Uber's entry. Sensitivity of the overall experience to tipping (both favorable and unfavorable) remains unchanged with Uber's entry.

Uber Reviews Favorable tipping experiences are discussed more in Uber reviews (M = 1.41, t=2.90) than in reviews for traditional services. The exact opposite is true for unfavorable

tipping experiences (M= -1.41, t=-3.46). Overall experience is equally sensitive to tipping (both favorable and unfavorable) across Uber and traditionals.

Implications Tipping is not separately required in a typical Uber transaction (as it is included in the total charge). Yet both overall rating for Uber and traditionals are equally sensitive to tipping.

4.6 Cleanliness

Traditional Taxi Reviews Favorable and unfavorable discussion of cleanliness remains unchanged post Uber's entry. Sensitivity of the overall experience to favorable discussion of cleanliness remains unchanged post Uber's entry, whereas there is a decrease in sensitivity to unfavorable discussion of cleanliness (M = -1.97, t = -4.94).

Uber Reviews While favorable discussion of cleanliness is about the same in both Uber and traditional cab reviews, unfavorable discussion of cleanliness is higher in Uber reviews (M = 0.63, t = 3.12). Overall experience is equally sensitive across Uber and traditionals to unfavorable discussion of cleanliness. However, overall experience for Uber benefits more from favorable discussion of cleanliness (M = 1.43, t = 4.21).

Implications Uber remains as sensitive as traditionals to unfavorable discussion of cleanliness and this maybe an aspect for potential improvement for Uber.

4.7 Payment Mode

Traditional Taxi Reviews Favorable discussion of payment mode increases (M = 0.50, t = 4.04) for traditionals post Uber's entry whereas unfavorable discussion of payment modes remains unchanged. Sensitivity of the overall experience to favorable discussion of payment modes remains unchanged post Uber's entry, whereas there is a decrease in sensitivity to unfavorable discussion of this aspect (M = -0.54, t = -2.37).

Uber Reviews Interestingly, both favorable and unfavorable discussion of payment modes is higher in Uber reviews than in traditional cab reviews (M = 24.70, t = 69.88 and M= 6.78 , t = 15.59 respectively) . Overall experience for Uber benefits more from favorable discussion of payment modes than traditionals. But both Uber and traditionals remain equally sensitive to unfavorable discussion of payment modes.

Implications It is interesting that Uber benefits more from favorable discussion of payment modes and generally payment mode is more salient in Uber reviews. Traditionals may want to focus communication on the improvements they have introduced in accepting multiple payment modes.

4.8 Wait Times

Traditional Taxi Reviews Favorable discussion of wait times increases (M = 1.01, t = 3.92) for traditionals post Uber's entry whereas unfavorable discussion of this aspect declined (M=-0.72, t=-2.42). Sensitivity of the overall experience to unfavorable discussion of wait times remains unchanged post Uber's entry, whereas there is a decrease in sensitivity to favorable discussion of this aspect (M = 1.90, t = 7.95).

Uber Reviews Interestingly, both favorable and unfavorable discussion of wait times is lower in Uber reviews than in traditional cab reviews (M = -4.20, t = -10.63 and M= -8.90,

t = -14.16 respectively). But both Uber and traditionals remain equally sensitive to discussion of wait times.

Implications The increase in favorable discussion of wait times for traditionals may simply reflect the reduced demand for traditional cabs (as customer switch to Uber). While wait times may not be as salient in Uber reviews, both Uber and traditionals remain equally sensitive

to discussion of wait times. It is possible that Uber's ability to match drivers is leading to small wait times, to the extent it is not discussed as much in reviews.

4.9 Trends in Uber Reviews

We also examined Uber reviews over time from July, 2010 to June, 2014. We used July, 2012 as a mid-point to examine changes in service experience with time for Uber. Tables 2.10 and 2.11 present the results from equations 9 and 10 respectively, adjusted for Uber reviews. The definition of the post period here is the period after June , 2012.

Table 2.10 : Change in discussion of attributes for Uber

Dependent variable is rating

Post is a dummy with indicating Uber reviews posted after June, 2012.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	1.71	1.63	1.05	0.29	
Post	-3.32	1.83	-1.82	0.07	*
rev_len	0.00	0.00	-0.52	0.60	
p_fare	**8.01**	**2.01**	**3.99**	**0.00**	***
n_fare	7.75	4.94	1.57	0.12	
p_tip	2.05	2.78	0.74	0.46	
n_tip	-1.19	4.30	-0.28	0.78	
p_wait	0.71	2.97	0.24	0.81	
n_wait	7.67	5.12	1.50	0.13	
p_reserve	**6.57**	**1.71**	**3.85**	**0.00**	***
n_reserve	**9.85**	**2.67**	**3.68**	**0.00**	***
p_clean	**8.04**	**1.72**	**4.67**	**0.00**	***
n_clean	3.49	2.38	1.47	0.14	
p_safe	5.37	2.90	1.85	0.06	*
n_safe	10.65	5.37	1.98	0.05	**
p_card	**5.97**	**1.45**	**4.12**	**0.00**	***
n_card	2.31	1.67	1.39	0.17	
p_drive	**8.98**	**1.81**	**4.97**	**0.00**	***
n_drive	2.29	2.27	1.01	0.31	
Post:rev_len	0.00	0.00	1.53	0.13	
Post:p_fare	3.89	2.86	1.36	0.17	
Post:n_fare	-1.25	5.85	-0.21	0.83	
Post:p_tip	-0.98	3.00	-0.33	0.74	
Post:n_tip	1.06	4.52	0.24	0.81	
Post:p_wait	3.63	3.31	1.10	0.27	
Post:n_wait	-6.36	5.76	-1.10	0.27	
Post:p_reserve	0.88	2.82	0.31	0.76	
Post:n_reserve	-2.86	3.67	-0.78	0.44	
Post:p_clean	1.60	2.06	0.77	0.44	
Post:n_clean	1.62	3.14	0.52	0.61	
Post:p_safe	0.62	6.18	0.10	0.92	
Post:n_safe	-2.79	6.99	-0.40	0.69	
Post:p_card	-1.80	1.73	-1.04	0.30	
Post:n_card	-1.66	1.97	-0.84	0.40	
Post:p_drive	-4.28	2.24	-1.91	0.06	*
Post:n_drive	-1.14	2.56	-0.44	0.66	

5 *** p<0.01, ** p<0.05, *p<0.10

6 Adjusted R-Square: 0.55

Table 2.11 : Change in sensitivity of overall-ratings to attributes in Uber Reviews

Variable	Estimate Of 'Uber'	Std. Error	t value	Pr(>|t|)	Significance
p_fare	**-1.21**	**0.53**	**-2.30**	**0.02**	***
n_fare	**-0.25**	**0.27**	**-0.93**	**0.35**	***
p_tip	1.30	0.80	1.63	0.10	
n_tip	0.56	0.57	0.98	0.33	
p_wait	1.19	0.55	2.17	0.03	
n_wait	0.72	0.30	2.41	0.02	
p_reserve	**-3.31**	**0.60**	**-5.55**	**0.00**	***
n_reserve	**-1.76**	**0.42**	**-4.20**	**0.00**	***
p_clean	1.25	0.82	1.53	0.13	
n_clean	**-0.49**	**0.45**	**-1.09**	**0.28**	***
p_safe	**-1.01**	**0.28**	**-3.60**	**0.00**	***
n_safe	**-0.10**	**0.19**	**-0.52**	**0.60**	***
p_card	5.36	1.54	3.49	0.00	
n_card	**-2.34**	**0.97**	**-2.41**	**0.02**	***
p_drive	0.20	0.75	0.27	0.79	
n_drive	0.52	0.84	0.62	0.53	
p_other	**-0.43**	**0.47**	**-0.91**	**0.37**	***
n_other	**-0.20**	**0.44**	**-0.46**	**0.65**	***

***** p<0.01, ** p<0.05, *p<0.10**

All regressions are significant, with adjusted r-squares ranging between 0.02 and 0.16

It is clear that there is no change on the various aspects of the service experience in terms of how much each aspect is discussed (see Table 2.10). What is interesting is that all service attributes also contribute less to overall satisfaction (see table 2.11). This suggests that consumers maybe revising their expectations from Uber.

4.10 Summary

The analysis reveals the following interesting findings. First, there is some good news for traditionals despite the loss of market share to Uber. Unexpectedly, there is an increase in favorable discussion of fares in reviews for traditionals. Despite the much touted advantages and rigor of the driver recruitment processes of the traditionals, driving experiences for traditionals declines pot Uber entry. In fact, Uber reviews register a higher amount of favorable driving

experiences. Uber is as sensitive as traditionals to discussion of safety, though the salience of safety itself appears to be lower for Uber. In terms of sensitivity, Uber is as sensitive as traditionals to unfavorable discussions of wait times and fares.

Finally, consumers appear to be resetting expectations in terms of Uber service experience, for example discussion of favorable fares declines with time for Uber.

This presents a thorny problem for regulators. On one hand, Uber, an unregulated entity, violating norms that are presumably in place to benefit consumers, is more favorably viewed on several experience dimensions than traditionals. On the other hand some service dimensions for traditionals are actually improving. Yet overall experience for both Uber and traditionals are declining. Finally, regulators may also need to consider aspects that consumers may not consider important or even be aware of. For example, safety may not be purely a concern generated by the lobbying efforts of traditional cab companies. Protection of consumer data collected by Uber may represent another dimension of concern that consumers (and press) may not be articulating loudly enough. It is possible that these changes are driven by either a) new customers who have usually never used traditional taxi services or b) self-selection of passengers (and thus reviewers) between traditional cab services and Uber post entry. However, we do not have data to either confirm or reject the same.

Section 5: Model Validation and Robustness Checks

In order to test the robustness of our results, we estimate a difference in differences regression model with San Jose as control. Specifically, the physically proximate city of San Jose did not witness the entry of Uber [52] till July 2013. Hence, we can use reviews posted for taxi services in San Jose as a control group for the purpose of inferring the causal effect of Uber's

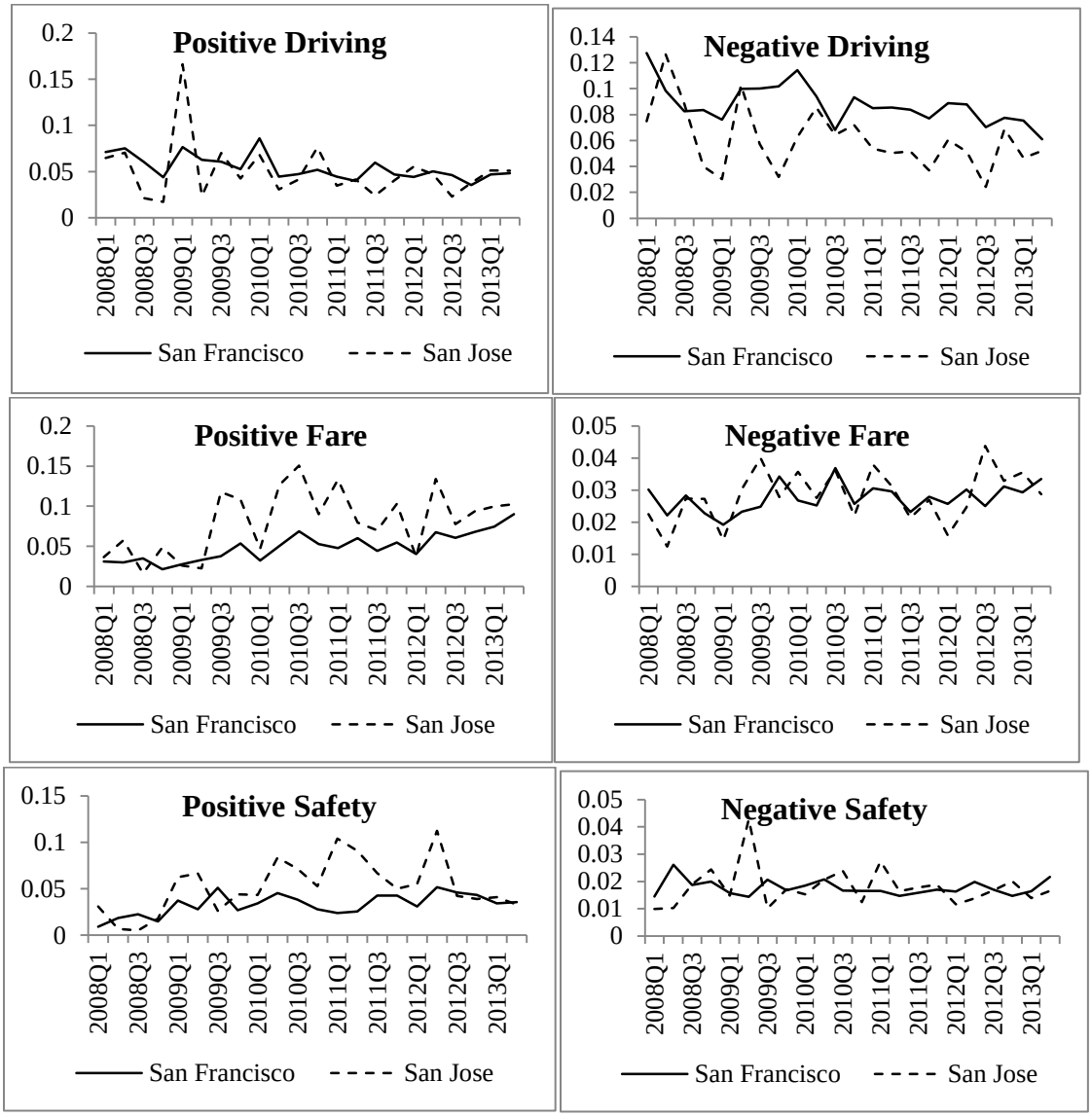
[52] http://www.mercurynews.com/ci_23722259/uber-car-service-launching-silicon-valley

entry in San Francisco. Both San Jose and San Francisco have a similar regulated medallion based system that licenses taxi services to operate in their respective cities. Further, on several aspects of consumer experience, the trends for topics are similar in the pre period. This gives us some confidence that San Jose is an appropriate control for our analysis (see Figure 2.4).

For us to draw causal inference, it is imperative that we compare changes (before vis-a-vis after the regulation) in the *same* construct across the groups. This construct is the topic proportion estimated on the treatment group—i.e. San Francisco. We first estimate changes in topic proportions within the treatment group, and then compare this to changes in the *exact same topic* within the control group. To identify the causal effect of the regulation on the proportion of each topic's discussion, we implement a "difference-in-differences" methodology (Angrist and Pischke, 2009; Goldfarb and Tucker, 2014).

We calculate the causal effect of a treatment (i.e. the entry of Uber) on the outcome variable (proportion of a given topic) by comparing the average change in the outcome variable for the treatment group (San Francisco) to the average change for the control group (San Jose).

Figure 2.4



Topic Trends for San Francisco and San Jose

We regress the outcome variable on two main effects (the effect of belonging to the treatment group on the outcome, and the effect of the treatment on the outcome), the interaction of these two effects, and several control variables, as follows:

$$\theta_{tr} = \gamma_{constant} + \gamma_{sfo} \times sfo_r + \gamma_{post} \times post + \gamma_{interaction} \times sfo_r \times post + controls + \epsilon_{tr} \quad (13)$$

$$rating_r = \gamma_{constant} + \gamma_t \times sfo_r + \gamma_{post} \times post + \Sigma_t \gamma_t \times \theta_{tr} + \Sigma_t \gamma_{t,sfo,post} sfo_r \times post \times \theta_{tr} + controls + \epsilon_{tr} \quad (14)$$

The notation used is the same as in the previous model specifications with two exceptions. Sfo_r is the dummy variable which accounts for the effect of belonging to the treatment group. It controls for unobserved time invariant factors which might affect topic proportions of San Francisco and San Jose differently. $post_r$ is the dummy variable which accounts for the effect of the treatment. It is possible that at the time of the implementation of the regulation, there were unobserved events which affected topic proportions across both San Jose and San Francisco. The main effect of $post_r$ controls for how topic proportions for all non-Uber services changed after the entry of Uber. The coefficient of the interaction term ($\gamma_{interaction}$) captures the crucial effect of the treatment on the treatment group. Further, we control for review specific attributes namely review length and rating. We also control for time fixed effects, firm fixed effects. We estimate the model on reviews posted between January1, 2008(date from which San Jose cab reviews are available) and July, 2013 (when Uber entered San Jose).

Table 2.12 : Differences in Differences Model for traditionals with San Jose as control

Dependent variable is topic proportions

Dependent Variable	Estimate of City x'Post' Interaction	Std. Error	t value	Pr(>|t|)	Significance
p_fare	**3.01**	**1.05**	**2.88**	**0.00**	***
n_fare	0.75	0.44	1.71	0.09	*
p_tip	-0.52	1.14	-0.45	0.65	
n_tip	**1.94**	**0.93**	**2.09**	**0.04**	**
p_wait	**2.48**	**0.84**	**2.97**	**0.00**	***
n_wait	**5.76**	**1.40**	**4.11**	**0.00**	***
p_reserve	**-2.80**	**1.13**	**-2.48**	**0.01**	**
n_reserve	**-4.00**	**1.46**	**-2.75**	**0.01**	**
p_clean	-1.58	1.19	-1.33	0.18	
n_clean	-0.34	0.43	-0.79	0.43	
p_safe	0.63	1.02	0.62	0.54	
n_safe	0.02	0.25	0.10	0.92	
p_card	-0.09	0.25	-0.35	0.73	
n_card	-1.15	0.92	-1.24	0.21	
p_drive	0.40	0.95	0.42	0.68	
n_drive	-2.32	1.39	-1.67	0.10	*
p_other	**-2.00**	**0.93**	**-2.14**	**0.03**	**
n_other	-0.20	1.18	-0.17	0.86	

***** p<0.01, ** p<0.05, *p<0.10**

All regressions are significant, with adjusted r-squares ranging between 0.11 and 0.27

The results for change in discussion of attributes, using San Jose as control, are largely consistent with the preceding analyses, with the exception of the sign on the coefficient for unfavorable waiting experience. Favorable discussion of fares and wait times increases, while the favorable discussion of reservations declines.

Table 2.13: Differences in Differences Model for traditionals with San Jose as control

Dependent variable is rating

Variable	Estimate of City x'Post' Interaction	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	**1.67**	**0.22**	**7.67**	**0.00**	***
rev_len	0.00	0.00	-0.04	0.97	
p_fare	**4.84**	**0.34**	**14.27**	**0.00**	***
n_fare	**5.58**	**0.79**	**7.08**	**0.00**	***
p_tip	**-0.90**	**0.39**	**-2.32**	**0.02**	**
n_tip	**-2.84**	**0.37**	**-7.74**	**0.00**	***
p_wait	**3.52**	**0.39**	**8.93**	**0.00**	***
n_wait	**-1.20**	**0.29**	**-4.11**	**0.00**	***
p_reserve	**4.37**	**0.34**	**13.01**	**0.00**	***
n_reserve	**-1.78**	**0.31**	**-5.68**	**0.00**	***
p_clean	**4.36**	**0.26**	**16.57**	**0.00**	***
n_clean	**2.95**	**0.78**	**3.79**	**0.00**	***
p_safe	**3.87**	**0.31**	**12.63**	**0.00**	***
n_safe	0.15	1.22	0.12	0.90	
p_card	2.78	1.46	1.91	0.06	*
n_card	**-0.94**	**0.38**	**-2.45**	**0.01**	**
p_drive	**4.37**	**0.34**	**13.01**	**0.00**	***
n_drive	**-1.93**	**0.26**	**-7.33**	**0.00**	***
city_indic	0.06	0.11	0.57	0.57	
city_post	**-0.97**	**0.23**	**-4.17**	**0.00**	***
p_fare:city_post	**1.21**	**0.43**	**2.80**	**0.01**	**
n_fare:city_post	0.58	0.99	0.58	0.56	
p_tip:city_post	0.73	0.46	1.58	0.11	
n_tip:city_post	**1.12**	**0.46**	**2.40**	**0.02**	**
p_wait:city_post	0.79	0.52	1.51	0.13	
n_wait:city_post	**1.32**	**0.38**	**3.44**	**0.00**	***
p_reserve:city_post	0.23	0.43	0.53	0.60	
n_reserve:city_post	**1.04**	**0.39**	**2.67**	**0.01**	**
p_clean:city_post	**1.43**	**0.36**	**3.95**	**0.00**	***
n_clean:city_post	-1.76	0.98	-1.81	0.07	*
p_safe:city_post	0.83	0.42	1.96	0.05	**
n_safe:city_post	-0.90	1.63	-0.56	0.58	
p_card:city_post	**-5.10**	**1.76**	**-2.89**	**0.00**	***
n_card:city_post	0.91	0.48	1.89	0.06	*
p_drive:city_post	0.78	0.45	1.74	0.08	*
n_drive:city_post	**0.88**	**0.33**	**2.64**	**0.01**	**

***** p<0.01, ** p<0.05, *p<0.10**

Adjusted R-Square: 0.69

The sensitivity of the overall experience to these attributes is also largely consistent with the preceding analyses. The signs for poor fares and unclean in cab experience are similar to those reported above. The sign for unfavorable payment experiences is however not consistent.

The identifying assumption of the difference-in-differences approach is that the topic proportions and ratings for tradition taxi services in San Jose are a valid counterfactual for the topic proportions and ratings that would have been obtained for traditional cabs services in San Francisco in the absence of Uber.

We also test the robustness of critical assumptions of our model, namely – inclusion of background weights, seed words, and ratings. We examine the fit of the model under each assumption. We will use the in sample log-likelihood as a measure of fit.

Inclusion of Background Weights The model with no background weights is nested within the full model. It is equivalent to the fully specified model where the vector of background weights is a zero vector. Figure 2.3 plots the log likelihood of the model with and without background weights. Clearly, estimating background weights leads to significantly better fit.

Inclusion of Seed words We compare the seeded and unseeded models on log likelihood and find that the difference in fit is negligible. Our choice of weights does not significantly worsen the fit of the model.

Inclusion of ratings Incorporating ratings allows us to jointly model reviews and ratings. A regression of rating on topics derived from the full model has an adjusted r-square of 0.61. In contrast a regression of ratings on topics derived from a model excluding topics has a poorer fit with an adjusted r-square of 0.57.

Section 6: Conclusion

The growth of the internet has led to the availability of very large quantities of data that are often less structured than data collected offline. Such data are often in the form of opinions of consumers (e.g. blogs, product reviews), are from an increasingly representative subset of the population, are in the public domain, and are available for long periods of time (e.g. 8 years in this research). This provides an unprecedented opportunity for marketers to not only understand what consumers are saying about their products at a given point in time, but also to continuously track changes in consumer opinion over time.

However, a major challenge for researchers is that much of these data are textual. It is perhaps for this reason that much of the research based on user-generated online content has focused on numerical descriptors of these data or simpler measures like word count. Techniques to analyze large volumes of text are at a nascent stage even in computer science. Yet, there is considerable interest from practitioners in using these data to gain usable knowledgeable. A recent report by the McKinsey Global Institute (Manyika et al. 2011) suggests that analyzing such data will become a "key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus."

Early research using online textual data in marketing has been focused on inferring market structure and product attributes in specific product categories; to ascertain the extent to which these correlate with consumer level data collected from more traditional experimental and survey based techniques; and to incorporate measures of such data in demand models. To our knowledge, no other extant research uses textual analyses to investigate the effect of a market entrant on consumer experience.

Substantively, we find that consumer opinion of traditional services improved along two key aspects after Uber's entry in San Francisco: pricing and wait times. In contrast to widely discussed safety concerns in the business press, we find that discussion of unsafe rides about Uber services in San Francisco hurts Ubers overall experience less than it hurts traditional services. On the other hand, Uber's overall rating is also declining, potentially because consumers have revised expectations on service quality. Both traditionals and Uber may want to examine these aspects of the service experience further. Our analysis also suggests that the effect on consumer experience is more complex than what a simple analysis of aggregate ratings might imply. This presents a special challenge for regulators seeking to evaluate the effect of Uber on traditional cab services.

Methodologically, we extend the Latent Dirichlet Allocation set of models in computer science. We look forward to several strategy- and policy-relevant applications as well as more sophisticated models in this area of topic detection and measurement.

References

Archak, N., Ghose, A., & Ipeirotis, P. G. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485-1509.

Andrzejewski, D., Zhu, X., & Craven, M.2009. Incorporating domain knowledge into topic modeling via Dirichlet forest priors. *In Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 25-32).

Angrist, J. D., & Pischke, J. S. 2008. Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Bao, Y. & Datta, A., 2014. Simultaneously discovering and quantifying risk types from textual risk disclosures. Management Science, 60(6), pp. 1371--1391.

Berger, J., Sorensen, A. T. & Rasmussen, S. J., 2010. Positive effects of negative publicity: When negative reviews increase sales. Marketing Science, 29(5), pp. 815--827.

Bird, S., E. Loper, E. Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D. M., 2012. Probabilistic topic models. Communications of the ACM, 55(4), pp. 77--84.

Blei, D. M., A. Ng, M. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. Volume 3, 993--1022.

Chen, Y. & Xie, J., 2008. Online consumer review: Word-of-mouth as a new element of marketing communication mix. Management Science, 54(3), pp. 477--491.

Chevalier, J. A. & Mayzlin, D., 2006. The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3), pp. 345--354.

Chintagunta, P. K., Gopinath, S., & Venkataraman, S. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, *29*(5), 944-957.

Danneels, E., 2004. Disruptive technology reconsidered: A critique and research agenda. Journal of product innovation management, 21(4), pp. 246--258.

Das, S. R. & Chen, M. Y., 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. Management Science, 53(9), pp. 1375--1388.

Dickey, J. M. 1983. Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. *Journal of the American Statistical Association*. 78(383), 628--637.

Eisenstein, J., Ahmed, A., and Xing, E. P. 2011. Sparse Additive Generative Models of Text. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1041-1048)

Eliashberg, J., S. K. Hui, J. Zhang. 2007. From Story Line to Box Office: A New Approach for Green-lighting Movie Scripts. *Management Science*. 53(6), 881--893.

Ghose, A., Ipeirotis, P. G. & Li, B., 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. Marketing Science, 31(3), pp. 493--520.

Goldfarb, A., & Tucker, C. 2014). Conducting Research with Quasi-Experiments: A Guide for Marketers. *Rotman School of Management Working Paper*, (2420920).

Gopinath, S., Thomas, J. S. & Krishnamurthi, L., 2014. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. Marketing Science, 33(2), pp. 241--258.

Griffiths, T., M. Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*. 101(Suppl 1), 5228--5235.
Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A. 2014. Interactive Topic Modeling. *Machine Learning*, 95(3), 423-469.

Huber, J., Kamakura, W. & Mela, C. F., 2014. A Topical History of JMR. Journal of Marketing Research, 51(1), pp. 84--91.

Jagarlamudi, J., H. Daumé III, R. Udupa. 2012. Incorporating Lexical Priors into Topic Models. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 204-213.

Lee, T. Y., E. T. Bradlow. 2011. Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research*. 48(5), 881--894.

Lin, C., and He, Y. 2009. Joint Sentiment/Topic Model for Sentiment Analysis. *In Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 375-384).

Lu, B., M. Ott, C. Cardie, B. Tsou. 2011. Multi-aspect Sentiment Analysis with Topic Models. Data Mining Workshops (ICDMW), *2011 IEEE 11th International Conference*. 81--88.Luo, X., 2009. Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. Marketing Science, 28(1), pp. 148--165.

Ma, L., Sun, B. & Kekre, S., 2015. The Squeaky Wheel Gets the Grease—An Empirical Analysis of Customer Voice and Firm Intervention on Twitter. Marketing Science.

Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. H. Byers. 2011. Big Data: the Next Frontier for Innovation, Competition, and Productivity. *McKinsey Global Institute*, May.

Markides, C., 2006. Disruptive innovation: In need of better theory*. Journal of product innovation management, 23(1), pp. 19--25.

Mcauliffe, Jon D., and David M. Blei. Supervised Topic Models.2008. *In Advances in Neural Information Processing Systems* 121-128.

Mela, C. F., Roos, J. & Deng, Y., 2013. Invited Paper-A Keyword History of Marketing Science. Marketing Science, 32(1), pp. 8--18.

Mimno, David,A. McCallum. 2008,Topic models conditioned on arbitrary features with Dirichlet-Multinomial Regression , *Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08), AUAI Press.*

Moe, Wendy W., and David A. Schweidel. 2012. Online product opinions: Incidence, evaluation, and evolution. *Marketing Science 31.3*: 372-386.

Moe, W. W. & Yang, S., 2009. Inertial disruption: the impact of a new competitive entrant on online consumer search. Journal of Marketing, 73(1), pp. 109--121.

Netzer, O., R. Feldman, J. Goldenberg, M. Fresko. 2012. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. *Marketing Science*. 31(3), 521--543.
Paul, M., and Dredze, M. 2012. Factorial LDA: Sparse multi-dimensional text models. *Advances in Neural Information Processing Systems* (pp. 2582-2590).

Sood, A. & Tellis, G. J., 2011. Demystifying disruption: a new model for understanding and predicting disruptive technologies. Marketing Science, 30(2), pp. 339--354.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24-54.

Tirunillai, S., G. Tellis. 2014. Mining Marketing Meaning from Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research*.

Wang, H., Lu, Y., & Zhai, C. 2010. Latent aspect rating analysis on review text data: a rating regression approach. *In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 783-792)

Yang, S. et al., 2012. An empirical study of word-of-mouth generation and consumption. Marketing Science, 31(6), pp. 952--963.

CHAPTER 3
THE ENROLLMENT EFFECT: A STUDY OF AMAZON'S VINE PROGRAM

Section 1: Introduction

In 2007 Amazon introduced its Vine program[53]. According to Amazon, "Amazon invites customers to become Vine Voices based on their reviewer rank, which is a reflection of the quality and helpfulness of their reviews as judged by other Amazon customers. Amazon provides Vine members with free products that have been submitted to the program by participating vendors. Vine reviews are the "*independent* opinions of the Vine Voices".[54] There could be potential concerns as to whether this enrollment affects the way reviews are written, introducing, for example, a positive bias.[55]

In this work, we investigate whether enrollment in the Vine program results in changes in the linguistic style used in reviews. We investigate this by looking at reviews by individuals before and after enrollment in the program. Following Feng et al. (2012) and Bergsma et al. (2012), we conduct a stylometric analysis using a number of syntactic and semantic features to detect differences in style. We believe that detecting changes in consumer behavior due to intervention by a firm is a novel natural language processing task. Our approach offers a framework for analyzing text to detect these changes. This work is relevant for social scientists and consumer advocates as research suggests that product reviews are influential (Chevalier and Mayzlin, 2006) and changes in style could potentially influence consumer decisions.

[53] http://blog.librarything.com/main/2007/08/amazon-vine-and-early-reviewers/
[54] http://www.amazon.com/gp/vine/help, words italicized by authors.
[55] http://www.npr.org/blogs/money/2013/10/29/241372607/top-reviewers-on-amazon-get-tons-of-free-stuff.

Section 2: Related Work

Our work lies at the intersection of research in four broad areas — Product Reviews, Product Sampling, Status and Stylometry.

Product Reviews Product reviews have received considerable attention in multiple disciplines including Marketing, Computer Science and Information Science. Research has addressed questions such as the influence of product reviews on product sales and on brands (Gopinath et al. (2014); Chevalier and Mayzlin (2006)), detection of deceptive reviews (Ott et al., 2011) and sentiment summarization (Titov and McDonald, 2008). This list is by no means comprehensive, but it is indicative of the extensive work in this domain.

Product Sampling Here, consumers receive products for free — as a marketing tactic. This is also a well-studied phenomenon. Research in this area has indicated that consumers value free products (Shampanier et al. (2007); Palmeira and Srivastava (2013)); that product sampling affects brand sales (Bawa and Shoemaker, 2004) and that sampling influences consumer behavior (Wadhwa et al., 2008).

Status Research shows that status can influence writing style. Danescu Niculescu Mizil et al. (2012) study discussions among Wikipedia editors and transcripts of oral arguments before the U.S. Supreme Court and show how variations in linguistic style can provide information about power differences within social groups.

Stylometry focuses on the recognition of style elements to identify authors (Rosen-Zvi et al., 2004), detect genders and even determine the venue where an academic paper was presented (Bergsma et al., 2012).

Our work draws from each of these research areas and in turn hopes to make a contribution to each in return. Our primary objective is to establish a framework to detect

behavioral change due to a decision by a firm (in this case enrollment to the Vine program characterized by free products and Vine membership status) by analyzing product reviews. Further, we hope to understand the dimensions on which this behavior may have changed. Consequently, we pursue a novel stylometric task. This type of work is especially important when the traditional numerical measure (rating) suggests there is no difference in the review pre and post-enrollment (see Section 4).

Section 3: Data & Preprocessing Steps

We gathered all reviews by the top 10,000 reviewers ranked by Amazon as of September, 2012. These rankings are partly driven by helpfulness and recency of reviews[5]. The data collected includes the review text, review title, rating assigned, date posted, product URL, product price, whether the reviewed product was received for free via the Vine program (also referred to as "Vine Review"), "helpfulness" votes and badges received by the reviewer.

We collected a total of 2,464,141 reviews of which 282,913 reviews were for products received for free via the Vine program. These reviews covered a total of 9,982 reviewers[56] of which 3,566 were members of the Vine program. Approximately half the reviews belonged to Vine members. We eliminated reviews that did not have a rating. We further excluded reviews where the review text was less than 20 words in length. We were left with 1,189,704 reviews by Vine members.

The date of enrollment to the Vine program for each reviewer is not explicitly available. We infer the date of enrollment in the following manner. We sort in ascending order all the "Vine Reviews" for each reviewer by posted date. We assume the earliest posted date for a "Vine

[56] During the crawling, ranks changed resulting in fewer than 10,000 reviewers in our data set.

review" is the enrollment date. This is an important assumption, as potentially reviewers could have moved in and out of the program at varying points of time. Reviewers can be moved out of the program for reasons such as not posting a "Vine Review" within 30 days of receipt of the product. In our data set we found 47,510 "Vine Reviews" by 163 reviewers who were not actively on the Vine program[57]. We can view these reviewers as having been dropped from the Vine program. Given the small volume of this type of reviews and reviewers, our assumption on date of enrollment appears reasonable.

Table 3.1: Data Summary

Member Type	Free/ Paid	Enrollment Timing	Review Count
Non Vine	Paid	NA	1,169,561
Non Vine	Free	NA	47,510
Vine	Paid	Post	452,729
Vine	Paid	Pre	503,688
Vine	Free	Post	233,287

Section 4: Enrollment Effect

This research seeks to answer the question: does enrollment in the Vine program change the writing styles of reviewers. One naive theory is that perhaps receiving products for free and receiving status badges will result in Vine members posting more positive reviews. Interestingly, the average rating for reviews by Vine members posted before enrollment is 4.22 and after enrollment is 4.21 and this difference is not statistically significant. In contrast, the length of reviews significantly increased from 251 words prior to enrollment to 306 words post-

[57] As these reviewers were not enrolled to Amazon's Vine Program as of September, 2012, they are excluded from our analysis.

enrollment. Natural language techniques are the only option to further investigate possible effects of enrollment. Consequently we focus on the review text posted by Vine members.

4.1 Approach

Following Ashok et al. (2013) and Bergsma et al. (2012) we construct features that represent writing style from each review (discussed in more detail in the next section). We incorporate these features in a classification algorithm that attempts to classify each review as having been written pre or post-enrollment to the Vine program. We report whether the difference in accuracy for this classifier vs. a majority vote classification is statistically significant or not. In order to detect differences in style pre and post-enrollment, we need to address certain confounding factors — Reviewer Specificity, Product Specificity and Time Specificity.

Reviewer Specificity It may be possible that certain users post more reviews post-enrollment than pre-enrollment. Consequently the classifier may simply end up learning the differences in style between reviewers. To avoid this, we construct a balanced sample where we randomly select 25 reviews for each reviewer prior to and post-enrollment (see Table 3.2). This also sets our baseline accuracy at 50%.

Product Specificity As the program started in 2007, the post-enrollment reviews are likely to predominantly contain products released in after 2007. This might result in the classifier simply learning the differences between products (say I Phone vs Palm). Given our focus on style, we do not use word tokens as such thus avoiding the use of product specific features. However, for some products, the product specific details may result in the use of specific syntactic structures. We assume this is not a significant contributor to the prediction performance. A post-hoc analysis of the top features supports this assumption. A second source

of change in writing style could be due to simply whether the product was bought or received for free. We exclude "Vine Reviews" [58] to eliminate this confounding factor.

Time Specificity A similar concern as Product Specificity exists for date references. By focusing on syntactic and semantic style, we avoid the use of time specific features.

Another concern is that perhaps post enrollment, reviewers receive writing guidelines from Amazon. This does not appear to be the case, as the writing guidelines[59] appear to be for all members. We now turn to the extraction of style features.

Table 3.2: Experiment Data

Data Type	Number of Reviews	Number of Reviewers
Training	113,250	2,265
Test	2,500	50

4.2 Feature Extraction

We consider three different features — "Bag of words/ unigrams", "Parse Tree Based Features" and an umbrella category consisting of genre and semantic features (see Section 4.2.3).

4.2.1 Bag of Words

Bag of Words/Unigrams (UNIGRAMS) Unigrams have often been found to be effective predictive features (Joachims, 2001). In our context, this serves as a competitive baseline for the classification task.

4.2.2 Parse Tree Based Features

[58] Reviews where product was received for free via the Vine program.
[59] http://www.amazon.com/gp/community-help/customer-reviews-guidelines

Following Feng et al. (2012) and Ashok et al. (2013) we use Probabilistic Context Free Grammar (PCFG) to construct a parse tree for each sentence. We then generate features from this parse tree and aggregate features to a review level.

All Production Rules (Γ) This set of features include all production rule features for each review, including the leaves of the parse tree for each sentence in the review. This effectively represents a combination of production rules and unigrams as features and represents an additional competitive baseline.

Non Terminal Production Rules (Γ^{N}*)* This excludes the leaves and hence restricts the feature set to nonterminal production rules. This allows us to investigate purely syntactic features from the text.

Phrasal/ Clausal Nodes (PHR/CLSL) We also investigate features that incorporate phrasal or clausal nodes of the parse trees. Please see Table 3.5 and Table 3.6 for examples of these features.

Parse Tree Measures (PTM) We construct a set of measures for each sentence based on the parse tree. These measures are maximum height of parse tree, maximum width of the parse tree and the number of sentences in each review.

Latent Dirichlet Allocation (LDA) We also apply Latent Dirichlet Allocation (Blei et al., 2003) to the production rules extracted from the Probabilistic Context Free Grammar. We use the topics generated as features in our prediction task. Our objective was to determine whether certain co-occurring production rules offered better classification accuracy. Our implementation includes hyper-parameter optimization via maximum likelihood. The number of topics is selected by maximizing the pairwise cosine distance amongst topics. We used the Stanford Parser (Klein and

Manning, 2003) to parse each of the reviews and the Natural Language Toolkit (NLTK) (Bird et al., 2009) to post process the results.

4.2.3 Genre and Semantic Features

Style Metrics (STYLE) This includes three distinct types of metrics. *Character Based* This includes counts of uppercased letters, number of letters, number of spaces and number of vowels. *Word Based* This includes measures such as number of short words (3 characters or less), long words (8 characters or less), average word length and number of different words. *Syntax Based* This includes measures such as number of periods, commas, common conjunctions, interrogatives, prepositions, pronouns and verbs.

Parts of Speech (POS) features have often been surprisingly effective in tasks such as predicting deception (Ott et al., 2011). Consequently we test this feature set as well.

Domain-independent Dictionary We make use of the Linguistic Inquiry and Word Count (LIWC) categorization (Tausczik and Pennebaker, 2010). One key advantage of this categorization is that it is domain independent and emphasizes psycholinguistic cues. We run two variants of this set of features. The first (LIWC ALL) includes all the categories — both subordinate and superordinate categories. The second (LIWC SUB CATEG.) only includes the subordinate categories, thus ensuring the features are mutually exclusive.

Subjectivity Measures (OPINION) We measure number of subjective, objective and other (neither subjective nor objective) sentences in each review. We use the "OpinionFinder System" (Wiebe et al., 2005) to classify each sentence with these measures. We aggregate the count of subjective, objective and other sentences at the review level and use these aggregates as

features.[60] We also report results on experiments where multiple feature types are included simultaneously in the model.

Section 5: Experimental Methodology

All experiments use the Fan et al. (2008) implementation of linear Support Vector Machines (Vapnik, 1998). The linear specification allows us to infer feature importance. We learn the penalty parameter via grid search using 5 fold cross-validation and report performance on a held-out balanced sample of reviews from 50 randomly selected users (all of whom were excluded from the training set) from the group of reviewers with at least 25 reviews in pre and post enrollment periods. While reporting the results, for some features we report the threshold (Thr) value set to exclude the least frequent features. These thresholds were also learned via the 5 fold cross validation process. Finally, text features can be binarized, mean centered and/or normalized. Each of these options were also selected via 5 fold cross validation.

Section 6: Results & Analysis

[60] One drawback is that the classifiers are trained on sentences from the MPQA corpus. Domain specificity is likely to yield poorer classification performance on our data.

Table 3.3: Experimental Results

Baselines		
Style Features	**Feature Count**	**Accuracy**
UNIGRAMS	796,826	60.9 %
Γ (Thr =50)	29,362	62.0 %
By Feature Type		
Style Features	**Feature Count**	**Accuracy**
Γ^N (Thr=200)	2,730	59.2 %
PHR/CLSL	23	57.4 %
PTM	3	55.8 %
LDA	200	54.0 %
STYLE	26	57.6 %
POS	45	57.5 %
LIWC ALL	76	59.8 %
LIWC SUB CATEG.	67	60.3 %
OPINION	3	56.3 %
Feature Combinations		
Style Features	**Feature Count**	**Accuracy**
Γ^N (THR=200) + STYLE	2,756	57.9 %
Γ^N (THR=200) + OPINION	2,733	56.2 %
PHR/CLSL + OPINION	26	58.0 %
PHR/CLSL + STYLE	49	57.5 %
LIWC + STYLE	93	60.2 %
LIWC + PHR/CLSL	90	60.2 %
LIWC + Γ^N (Thr=200)	2,797	59.1 %
LIWC + OPINION	70	60.3 %
PTM + OPINION	6	57.2 %
STYLE + OPINION	29	58.7 %
STYLE + PTM	29	57.4 %
LIWC +STYLE+PHR/CLSL	116	60.1 %

All of the feature sets perform statistically better[61] than a majority vote (50%). Baselines Unsurprisingly, the feature set containing all production rules (Γ) yields the best ac curacy (62.0 %). Unfortunately, as expected, the top features all included terminal production rules that signal time or product specificity. For example in the pre-enrollment reviews the top 10 features for include NNP → 'Update', CD → '2006', NNP → 'XP' and NNP → 'Palm'. In the post-enrollment reviews the top 10 features include CD → '2012',CD → '2011', NN → 'iPad' and NN → 'iPhone'. We observe the same issue with the UNIGRAMS feature set. This supports our

[61] as indicated by a paired t-test at p=0.05 on the held out sample.

contention that the analysis should restrict itself to style and domain-independent features. The best performing style feature set is LIWC SUB CATEG. followed by Non Terminal Production Rules (Γ^{N}). OPINION is the most parsimonious feature set that performs significantly better than a majority vote.

Non Terminal Production Rules (Γ^{N}): Table 3.7 presents the top Non Terminal Production Rules. We observe the following: First, pre-enrollment reviews have noun phrases(NP) that contain fewer leaf nodes than in the post-enrollment reviews. This appears to be due to the inclusion of determiners (DT), adjectives (JJ), comparative adjectives (JJR), personal pronouns (PRP $) or simply more nouns (NN). This might indicate that topics are discussed with more specifics in post-enrollment reviews. Second, clauses(S) begin with action oriented verb phrases (VP) in the pre-enrollment reviews. In contrast in the post-enrollment reviews clauses connect two clauses using coordinating conjunctions(CC) or prepositions(IN). One possibility is that reviewers are offering more detail/concepts per sentence (where each clause is a detail/concept) in the post-enrollment reviews. Finally, we observe that pre-enrollment reviews include adjectival phrases (ADJP) connect to superlative adverbs (RBS)which convey certainty. We will revisit this finding when we review the results from the LIWC model below.

Phrasal/Clausal (PHR./CLSL.): Tables 3.5 and 3.6 suggest that post-enrollment reviews emphasize information using descriptive phrases — adjectival phrases (ADJP) and adverbial phrases (ADVP) — and quantifier phrases (QP). Pre-enrollment reviews appear to have more complex clause structures (SBAR, SINV, SQ, SBARQ see table 3.5 for definitions).

Parse Tree Metrics (PTM): The three features used are number of sentences, maximum height of parse tree and the maximum width of the parse tree, listed here in descending order of importance for the post-enrollment reviews. As mentioned earlier in section 4 the average review

length is higher in the post-enrollment reviews so the finding that the number of sentences predict post-enrollment reviews is consistent. Maximum tree width predicts the pre-enrollment reviews. This flat structure indicates a more complex communication structure.

Latent Dirichlet Allocation (LDA): This model did not perform very well, being statistically marginally better than majority vote. As mentioned before, we selected the number of topics by maximizing the average cosine distance amongst topics. Even with 200 topics, this measure was 0.39, suggesting that the topics were themselves not well separated. In the limit, each topic would be a nonterminal production rule. This is the same as Non Terminal Production Rules (Γ^N) feature set discussed earlier in this section.

Table 3.4: Style Metrics: Top Features

Predicts PRE Enrollment
'number of different words', 'uppercase', 'alphas', 'vowels' , 'short words', 'words per sentence', 'to be words' , 'punctuation symbols', 'long words', 'common prepositions'
Predicts POST Enrollment
'average word length', 'spaces', 'verbs are', 'chars per sentence' , 'verbs be', 'common conjunctions', 'verbs were', 'personal pronouns' , 'verbs was', 'verbs am'

Style (STYLE): Table 3.4 presents the top features for this feature set. The features suggest that reviewers used a more varied vocabulary (number of different words), more words per sentence (words per sentence) and more long words (long words) in pre-enrollment than in post-enrollment reviews. This might indicate that sentences in the pre-enrollment reviews were longer and more complex. Interestingly, the average word length did go up in the post-enrollment reviews as did the characters per sentence. In addition, more personal pronouns and conjunctions are used — a finding replicated in the model using LIWC features (see below).

Parts of Speech (POS): The top features for post-enrollment are commas, periods, comparative adjectives, verb phrases and coordinating conjunctions. The top features for pre-enrollment are nouns, noun phrases, determiners, prepositions and superlative adverbs. These results are more difficult to interpret though the use of comparative adjectives suggests more comparisons between different objects in the post enrollment reviews.

LIWC SUB CATEG: The top 10 LIWC features are shown in Table 3.8. LIWC features are categories that are contained in broader categories. For example POSEMO (see Table 3.8, first feature for "Predicts POST enrollment") refers to the class of positive emotion words. POSEMO itself is contained in a category called "Affective Features" which in turn is classified as a Psychological Process (abbreviated to Pscyh.). The analysis of the categories of features is in itself interesting. Psych./ Cognitive Features occur higher up in features predictive of pre-enrollment reviews than in the features predictive of post-enrollment reviews. "Psych./ Affective Features" occurs as a top feature for the post-enrollment reviews. The actual feature from the "Psych./ Affective Features" category is POSEMO suggesting that the positive emotion is more strongly conveyed in the post-enrollment reviews than in the pre-enrollment reviews. Interestingly the corresponding negative feature NEGEMO is in the top 10 features predicting the pre-enrollment reviews. This is especially intriguing since the average rating for reviews in the pre and post-enrollment reviews is the same (see 4). We were concerned that possibly our sampling had induced a bias in the ratings. But the average ratings in our sample are 4.18 and 4.19 pre and post-enrollment respectively (difference is not statistically significant).

FUNCTION WORDS occur extensively in the post-enrollment reviews. We also observe that inclusive (INCL) and exclusive (EXCL) terms are used more in the post-enrollment reviews. It's possible that reviewers are seeking to be more balanced. Products are described in personal

(I), perceptual (FEEL) and relativistic (SPACE) terms. Pre-enrollment reviews discuss personal concerns (LEISURE, RELIG) , indicate a level of certainty (CERTAIN) and opinions are presented in terms of thought process (INSIGHT). Interestingly, the pre-enrollment reviews address the reader (YOU).

Opinions (OPINION): Features predicting post-enrollment are number of objective sentences, number of subjective sentences and finally number of other (neither subjective nor objective) sentences. This suggests that reviewers try to write somewhat more objectively in the post-enrollment reviews.

Feature Combinations With the exception of the combinations STYLE + OPINION , PHR/CLSL +OPINION and PTM + OPINION which improve on either feature set used alone, none of the other combinations improved performance over all component feature sets modeled individually. Overall, none of the combinations improved over LIWC SUB CATEG. Hence we do not delve further into features from these models.

Summary Overall pre-enrollment reviews are more complex (complex clauses, wide parse trees, varied vocabulary, more words per sentence), have fewer concepts per sentence, contain negative emotions, addresses the reader directly and are more certain. Post-enrollment reviews are longer, more descriptive, contain comparisons, contain quantifiers, have more positive emotion and describe the product experience in physical and personal terms. These reviews are specific, balanced and contain more objective sentences as well.

Discussion on Readability One possibility is that the "Enrollment" effect leads to reviewers writing more readable reviews. To test this hypothesis we performed a paired t-test between readability scores for pre and post-enrollment reviews. Table 3.9 suggests that indeed this is the case. Flesch Reading Ease is the only measure where a higher score indicates simpler text. For

the rest of the measures a higher score implies more complex text. All of the measures are within the average readability range and the magnitude of the differences is small. Nevertheless, these differences are statistically significant [62]with one exception lending support to the idea that "Enrollment" effect might lead to reviewers writing more readable reviews.

[62] The cell size for each class is 57,875, making the modest difference in magnitude statistically significant

Table 3.5: Phr/Clsl: Top Features PRE

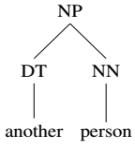
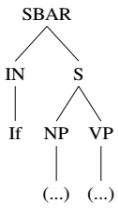
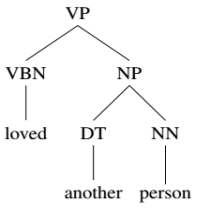
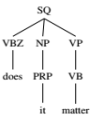
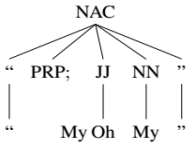
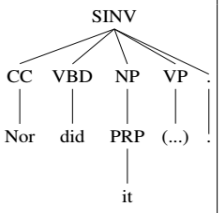
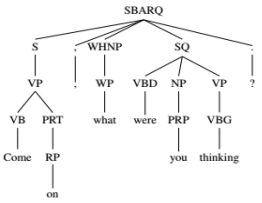
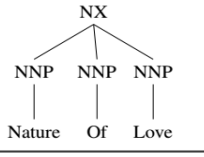
Predicts PRE Enrollment	
1 NP (Noun Phrase)	6 LST (List marker. Includes surrounding punctuation)
EXAMPLE 	EXAMPLE (3)
2 SBAR (Clause introduced by a (possibly empty) subordinating conjunction)	7 VP (Verb Phrase)
EXAMPLE 	EXAMPLE 
3 SQ (Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ)	8 PRN (Parenthetical)
EXAMPLE 	EXAMPLE (p. 73)
4 NAC (Not a Constituent; used to show the scope of certain prenominal modifiers within an NP)	9 SINV (Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal)
EXAMPLE 	EXAMPLE 
5 SBARQ (Direct question introduced by a wh-word or a wh-phrase)	10 NX (Used within certain complex NPs to mark the head of the NP)
EXAMPLE 	EXAMPLE 

Table 3.6: Phr/Clsl: Top Features POST

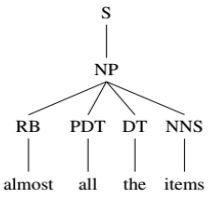
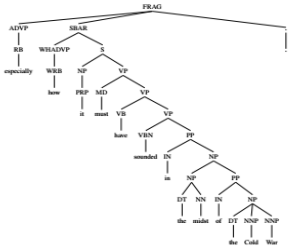
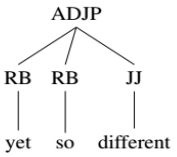
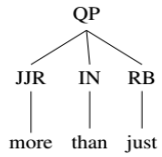
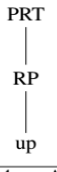
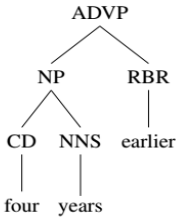
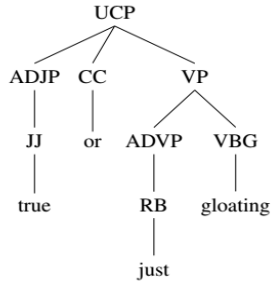
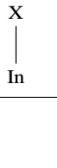
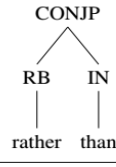
Predicts POST Enrollment	
1 S (Simple declarative clause)	6 FRAG (Fragment)
EXAMPLE 	EXAMPLE 
2 ADJP (Adjective Phrase)	7 QP (Quantifier Phrase)
EXAMPLE 	EXAMPLE 
3 PRT (Particle. Category for words that should be tagged RP)	8 WHNP (Wh-noun Phrase)
EXAMPLE 	EXAMPLE 
4 ADVP (Adverb Phrase)	9 UCP (Unlike Coordinated Phrase)
EXAMPLE 	EXAMPLE 
5 X (Unknown, uncertain, or unbracketable)	10 CONJP (Conjunction Phrase)
EXAMPLE 	EXAMPLE 

Table 3.7: Γ^{N}: Top Features (PCFG Non Terminal)

Predicts PRE Enrollment	
Feature	**Examples**
ROOT → S	(1) And nearly every single item seemed cute and usable to me. (2) Look closely, (...) overwhelming personal and cultural upheaval.
NP → NNP NNP	(1) Tim Bess (2) Jennifer Fitch
PP → IN NP	(1) for its psychological and emotional richness (2) of loyalty
NP → DT NN	(1) the price (2) a book
NP → NNP POS	(1) Frost 's (2) Clough 's
ADJP → RBS JJ	(1) most assuredly (2) most entertaining
WHNP → WP	(1) who (2) what
NP → NNP	(1) Blessed (2) India
PP → TO NP	(1) to the crime (2) to me
S → VP	(1) linking Pye to the crime scene (2) Gripping due to (...)
Predicts POST Enrollment	
Feature	**Examples**
S → S , IN S .	(1) It is functionally the same as Apple's 10 watt charger which outputs 2.1 A , so it is also suitable for charging the iPad. (2) It has 3 levels of trays that spread as you open the box, so you can easily access contents in all trays.
S → IN NP VP .	(1) So I don't think the investment in graphics (...) enjoyability in the game. (2) So we decided to try it again this year.
ROOT → NP	(1) Some kind of (...) disorder ? (2) Proper Alignment and Posture; This segment (...) .
S → S CC S .	(1) Mage and Takumo (...) but lacking in depth.(2) The light feature is great and it powers off (...).
NP → PRP$ NNP NN	(1) your Alpine yodeling (2) my MacBook Pro
S → VP .	(1) Enough negativity. (2) Suffice it to say that (...) .
NP → DT JJR NN	(1) a better future (2) a slower flow
NP → DT JJ , JJ NN	(1) an immediate , visceral reaction (2)a roots-based, singer-songwriter effort
NP → DT NNP NNP NNP NNP	(1) the Post-Total Body Weight Training (2) The Gunfighter DVD Gregory Peck
WHADVP → WRB RB	(1) How far (2) how well

Table 3.8: LIWC Sub Category: Top Features

Predicts PRE Enrollment		
Feature	**Category**	**Examples**
leisure	Personal Concerns	Cook, chat, movie
verb	Function words	Walk, want, see
certain	Psych./Cognitive Processes	always, never
insight	Psych./Cognitive Processes	think, know, consider
negemo	Psych./Affective Processes	Hurt, ugly, nasty
exclam	Exclamation	!
period	Period	.
you	Function words	2^{nd} person , you, your
preps	Function words	to, with, above
relig	Personal Concerns	2^{nd} synagogue, sacred
Predicts POST Enrollment		
Feature	**Category**	**Examples**
posemo	Psych./Affective Processes	Love, nice, sweet
article	Function words	a, an, the
i	Function words	1^{st} person singular.
space	Psych./Relativity	Down, in, thin
ingest	Psych./Biological Processes	Dish, eat, pizza
ipron	Function words	Impersonal Pronouns, it its , those
incl	Psych./Cognitive Processes	Inclusive, and, with , include
conj	Function words	and, but, whereas
excl	Psych./Cognitive Processes	Exclusive but, without, exclude
feel	Psych./Perceptual Processes	feels , touch

Table 3.9: Readability Measures

Reading Measure /Cite	Pre Mean	Post Mean	t Value
ARI /(Senter and Smith, 1967)	9.16	9.15	(0.45)
Coleman Liau /(Coleman and Liau, 1975)	8.76	8.68	(6.39)*
Flesch Kincaid /(Kincaid et al., 1975)	8.75	8.71	(2.19)*
Flesch Reading Ease /(Kincaid et al., 1975)	65.63	66.18	6.61*
Gunning Fog /(Gunning, 1952)	11.75	11.70	(2.18)*
LIX /(Anderson, 1983)	38.24	38.07	(2.89)*
RIX /(Anderson, 1983)	3.74	3.71	(3.05)*
SMOG /(McLaughlin, 1969)	10.59	10.56	(2.56)*
* Significant at 5% level			

Section 7: Discussion

So far we have ignored the possibility that writing styles of reviewers may simply continuously evolve with experience and we are simply detecting a difference due to this underlying trend.[63] To address this question we investigated the sub-periods within the pre and post enrollment periods.

We split the post enrollment period (i.e. from date of enrollment to the date the most recent review was posted) further into two equal time periods for each reviewer. As before, we learn a classifier to discriminate between the sub periods. Interestingly the classifier performed the same as chance at p=0.05 (Test Accuracy= 51.0%).[64] [65]However a similar analysis in the pre-enrollment period results in a test set accuracy of 63.3% (significant at p=0.05). So there is a

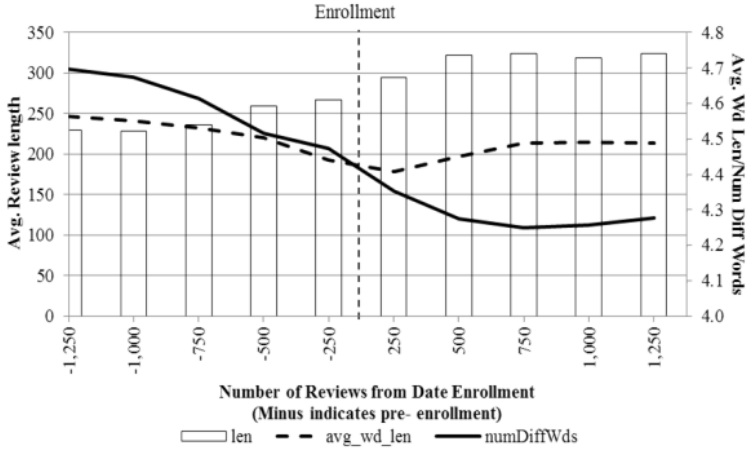
[63] Ideally, if a) the enrollment date had been the same for all reviewers and b) the enrollment was random, we would have a clean experimental framework to detect whether a similar trend exists for non-vine reviewers. Unfortunately, this is not the case.

[64] We report the results only on POS for conciseness. The other feature sets performed similarly.

[65] As before the test sample includes 50 users. However we sampled only 10 reviews in each sub period. Corresponding down sampled performance for Pre vs Post enrollment accuracy is 57.5% (significant at p=0.05)using POS features.

change in writing style within the pre-enrollment period, but there is no continued change post-enrollment. This is not consistent with the continuous style evolution hypothesis. One account would be that Amazon enrolls reviewers whose styles have stabilized. This remains a possibility as Amazon actively selects the members (and we are not aware of the specific rules used by Amazon). The trends (see Figure 3.1) suggest that there are changes right upto the enrollment date and some levelling out in the post enrollment period, providing some evidence against the hypothesis.

Figure 3.1



Feature Trends

Table 3.10: Sub Period Results

		Train Size	Test Size	Accuracy
Within	Pre Enrollment	44,800	1000	63.3%
Within	Post Enrollment	59,250	1000	51.0%
Pre vs Post (Down Sampled)	Enroll.	53,840	1000	57.5%

Section 8: Conclusion

We view this work as a first step toward investigating this phenomenon further. In particular, we plan to test the robustness of our results w.r.t. product specificity, to investigate stylistic differences (a) between reviews for purchased products versus for products received for free amongst Vine members and (b) between reviews by Vine reviewers and non-Vine reviewers. Another line of inquiry involves decomposing the "Enrollment" effect into a reputation/status effect (the influence of the status badge - Vine membership) and a product sampling effect (the influence of receiving goods for free). Finally, investigating the temporal dynamics of style for these reviewers might prove interesting as would determining whether these subtle differences in style affect the readers and influence purchase decisions.

References

Jonathan Anderson. Lix and rix: Variations on a little-known readability index. Journal of Reading, pages 490–496, 1983.

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. Poetry, 580(9): 70, 2013.

Kapil Bawa and Robert Shoemaker. The effects of free sample promotions on incremental brand sales. Marketing Science, 23(3):345– 363, 2004.

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 327–337. Association for Computational Linguistics, 2012.

Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media, 2009.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.

Judith A Chevalier and Dina Mayzlin. The effect of word of mouth on sales: Online book reviews. Journal of marketing research, 43(3): 345–354, 2006.

Meri Coleman and TL Liau. A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2):283, 1975.

Cristian Danescu Niculescu Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. Echoes of power: Language effects and power differences in social interaction. In Proceedings of the 21st international conference on World Wide Web, pages 699–708. ACM, 2012.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Lib-linear: A library for large linear classification. The Journal of Machine Learning Research, 9: 1871–1874, 2008.

Song Feng, Ritwik Banerjee, and Yejin Choi. Characterizing stylistic elements in syntactic structure. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1522–1533. Association for Computational Linguistics, 2012.

Shyam Gopinath, Jacquelyn S Thomas, and Lakshman Krishnamurthi. Investigating the relationship between the content of online word of mouth, advertising, and brand performance. Marketing Science, 2014.

Robert Gunning. Technique of clear writing. 1952.

Thorsten Joachims. A statistical learning learning model of text classification for support vector machines. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pages 128–136. ACM, 2001.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics Volume 1, pages 423– 430. Association for Computational Linguistics, 2003.

G Harry McLaughlin. Smog grading: A new readability formula. Journal of reading, 12(8):639–646, 1969.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies Volume 1, pages 309– 319. Association for Computational Linguistics, 2011.

Mauricio M Palmeira and Joydeep Srivastava. Free offer6= cheap product: A selective accessibility account on the valuation of free offers. Journal of Consumer Research, 40(4):644–656, 2013.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author topic model for authors and documents. In Proceedings of the 20th conference on Uncertainty in artificial intelligence, pages 487–494. AUAI Press, 2004.

RJ Senter and EA Smith. Automated readability index. Technical report, DTIC Document, 1967.

Kristina Shampanier, Nina Mazar, and Dan Ariely. Zero as a special price: The true value of free products. Marketing Science, 26(6):742– 757, 2007.

Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of Language and Social Psychology, 29(1):24–54, 2010.

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL08: HLT, pages 308–316, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Vladimir N. Vapnik. Statistical Learning Theory. WileyInterscience, 1998.

Monica Wadhwa, Baba Shiv, and Stephen M Nowlis. A bite to whet the reward appetite: The influence of sampling on reward-seeking behaviors. Journal of Marketing Research, 45 (4):403–413, 2008.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. Language resources and evaluation, 39(23):165–210, 2005.

CHAPTER 4
IN SEARCH OF U-CURVES:
LIKELIHOOD AND THE OPTIMAL NUMBER OF TOPICS

Section 1: Introduction

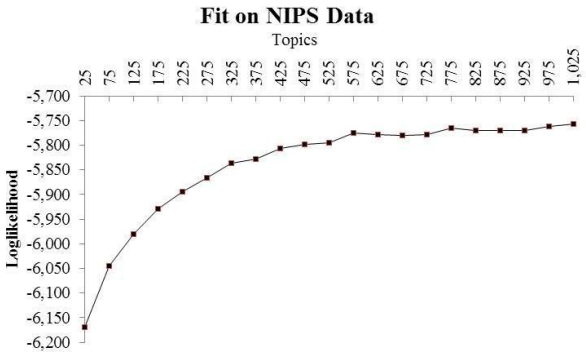
There is a widespread assumption that held-out probability can be used to find the optimal number of topics for an LDA model on a specific data set. Users may be skeptical of results based on seemingly arbitrary parameters. As topic modeling becomes common outside of machine learning and NLP contexts, it has become increasingly important to provide guidance to non-expert users.

Previous work in topic modeling (Blei et al., 2003; Steyvers and Griffiths, 2007) addresses the selection of the number of topics K in two ways. In some cases there are domain-specific justifications for particular numbers. More often, the choice of number of topics cannot be argued from theory so K must be empirically determined.

The most common method is to plot the log-likelihood of a held-out test set for varying K Griffiths and Steyvers (2004). The assumption is that as we add more topics, the model becomes more expressive, increasing our ability to predict combinations of words in unseen documents, up to a specific ideal point. If we go beyond that point, we assume that the model becomes too expressive, causing topics to overfit and simply memorize the input corpus, therefore reducing the likelihood of previously unseen documents. This increase, inflection, and decrease result in an inverted U-Shaped plot. We select the number of topics where this plot reaches its maximum. There are several problems with this argument. First, we often find that increasing K continues to

improve fit to arbitrarily large values. Figure 4.1 demonstrates this phenomenon on the NIPS data set[66].

Figure 4.1



Likelihood increases and then flattens with more topics.
We typically do not observe an inflection point.

Second, even if we find a turning point it is not obvious that that turning point will recover the correct number of topics. Third, the method does not explicitly penalize adding additional parameters to a model, but rather relies on an assumption that "overfitting" will punish overly expressive models. Finally, the basic concept of a "correct" number of topics is questionable. Although LDA is surprisingly effective at recovering semantically related words, no one would argue that it is a realistic model of human language.

In this work we investigate multiple criteria for selecting K. In section 3 we evaluate the effectiveness of likelihood based metrics currently used in this task. We control for the question of whether a correct K is a meaningful concept by using synthetic data generated from the LDA process with a known number of topics (see section 2). We find that even under these ideal conditions, likelihood-based estimation of K is usually unable to determine a unique optimal number of topics, and when it does, that number is consistently too large. Although held-out

[66]http://www.cs.nyu.edu/ roweis/data.html

likelihood is unlikely to find a single "correct" number of topics K, we do find that it consistently identifies values of K that are too large. We explore several potential causes of this phenomenon such as choice of estimation method (Gibbs Sampling vs variational inference), choice of fit measure, and empirical relationships between size of the vocabulary and the size of the corpus.

We then turn examine parameter convergence properties in a Bayesian setting in section 4. In section 5 we examine empirical measures as well as present conditions under which it is possible to recover the correct number of topics. Finally, we offer modeling guidelines for the engaged practitioner. To our knowledge this is the first comprehensive study of this problem.

Section 2: Synthetic Data Generation

We construct synthetic data sets that are consistent with the LDA generative model and have specific, known numbers of topics K. In the standard LDA model Blei et al. (2003) the topics are categorical distributions with Dirichlet priors. To model the topics we use a symmetric Dirichlet prior = 0:01. This prior is used to draw independent draws of a given vocabulary size. Each draw is a topic. For simplicity we treat document length as a fixed, known constant. The topic proportions in each document are a categorical distribution with a Dirichlet prior. We use a symmetric Dirichlet prior = 0:01 to make draws for each document in the corpus. This distribution will concentrate most probability in a given document on a small number of topics, theoretically improving our ability to separate topics. For each document we sample a sequence of topic indicator variables, and then for each topic variable in the sequence we sample a word from that topic's distribution.

We construct synthetic data sets where the number of topics varies from 25 to 200 (inclusive) in increments of 25. The vocabulary size is set at 5,000 and the number of documents is 10,000 for all of these data sets. The length of each document is fixed at 100 words. In a

separate set of experiments we increase the corpus size to 20,000 documents and all other factors remain the same.

Section 3: An Examination of Likelihood

3.1 Measures of Fit

Computing the probability of a document requires summing over all possible token-topic assignments, so we rely on approximate estimators. We consider four methods. The simplest method Document Completion or Half-Heldout evaluation. We split held-out documents so that half of each test document is included in the training set.

We then evaluate the testing half using the estimated topic distribution of the training half. The Harmonic Mean estimator Griffiths and Steyvers (2004) runs a Gibbs sampler over the held-out document and computes the harmonic mean of likelihoods from saved samples. Wallach et al. (2009b) show that the Harmonic Mean method consistently overestimates the log-likelihood of documents, but we find that the relative values between different settings of K are consistent with other fit metrics. Wallach et al. (2009b) propose two more accurate methods, Importance Sampling and the left-to-right Marginal Probability estimator.

3.2 LDA Model specifications

We use two methods for training topic models, MCMC and variational inference. We report results on test sets comprising 20% of the corpus size. In running inference on the test set, the assumption is that the word-topic distributions remain unaffected — we are only inferring document-topic proportions. The experiments are run on synthetic corpora, for which we know the real number of topics. We estimate topic models with varying numbers of topics for each data set, and report multiple measures of fit.

We estimate models using Gibbs Sampling and variational inference. We run Gibbs sampling using MALLET (McCallum, 2002) for 2,000 burn-in iterations and then save samples every 10 iterations for an additional 3000 iterations (Nguyen et al.). We observe convergence in loglikelihood for the training data for all models. We then run inference on the test set for 5,000 iterations. We estimate models with hyperparameter optimization (Wallach et al., 2009a) because researchers do not expect there to be specific known hyperparameter values, and it is advisable to estimate the values from data. In addition, we also present results with fixed hyperparameters, including the original parameters used to generate the corpus. For variational inference we estimate models with high and low tolerance levels for model estimation using lda-c. All hyperparameters are optimized in the variational inference specification.

3.3 Results

Figures 4.2 to 4.9 present results from the different synthetic data sets using various estimation methods. Columns indicate the actual number of topics. The fit measures are listed on the left and the cell values indicate the number of topics associated with the best fit for the corresponding metric. Similar to the real-data results in Fig. 4.1 we find that performance increases and then levels off, but unlike the real-data results we frequently find an optimal value that is less than the largest tested number of topics. Looking for changes in first differences to identify "flattening-out" points did not produce better results than taking the max. Boot-strapped standard errors are far too small to make any difference in the estimates of fit.

We fit every data set with the following number of topics: 25, 50, 75, 100, 125, 150, 175, 200, 300, 500. We use color to indicate whether the number of topics selected is greater than (red), less than (blue), or equal to (grey) the actual number of topics in the synthetic data set. We use shading to indicate when the estimated number of topics is off by more than 50 topics. If the

number of topics selected is 500, this indicates that the likelihood continued to improve, and not necessarily that we obtained a maximum. We report results on the whole test set (labelled "Test") and on the half held-out test set (labelled "Half Heldout").

When we use hyperparameter optimization evaluation, all methods typically overestimate the number of topics K. Importance Sampling (Imp. Samp.) appears to overestimate K to a point but then begins to fail. Harmonic Mean prefers more topics and will continue to improve fit as the number of topics increase. The Marginal Probability Estimator (MPE) consistently overestimates the number of topics by 25–50 topics. The method of estimation (Gibbs vs. variational) does not appear to make a difference, so we report only results trained with Gibbs sampling for the remainder of this section. The actual hyperparameters estimated are typically wrong by at least an order of magnitude.

Figure 4.2

Optimized Alpha, Beta		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	25	200	75	200	300	500	50	50
Imp. Samp.	Half Heldout	25	50	150	200	300	50	50	50
Harm. Mean	Test	500	500	500	500	500	500	500	500
Harm. Mean	Half Heldout	25	75	100	150	175	200	300	300
MPE	Test	25	75	100	150	175	200	200	300
MPE	Half Heldout	25	75	100	150	175	200	200	300

Gibbs sampling with optimized hyperparameters α,β

Figure 4.3

Variational Inference		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Harm. Mean	Test	50	75	100	175	175	200	200	175
MPE	Test	50	75	100	175	175	200	200	200
MPE	Half Heldout	25	75	100	150	150	200	200	300

Variational Inference

We now consider four settings of fixed hyperparameters. These settings affect the estimate of K for all fit metrics.

Small α,β : Both and are known and are set equal to 0.01 (Figure 4.4). Although this is the actual setting used to generate the corpus, all metrics overestimate K.

Large α,β : and are fixed, but set to values that indicate greater uncertainty over distributions. We use =0.2 and =0.2 (Figure 4.5). In this case fit metrics consistently underestimate K.

Large, decaying α : β is known, and varies as $\frac{50}{K}$ .This and the next specification assume that we know the sum of the parameters but not the number of topics (Figure 4.6). This heuristic performs surprisingly well given its distance from the true = 0:01.

Smaller, decaying α: β is known, and varies as $\frac{25}{K}$. (Figure 4.7). This specification has the same functional form as the previous, but performs similarly. The magnitude of the Dirichlet parameter matters. However, additional testing indicates that if the choice of beta is wrong (e.g. 0.2), then the results become quite unreliable (not shown here).

Using more data. Adding more documents does not improve results. We ran multiple experiments, with both fixed and optimized hyperparameters. The number of topics K was either over- or understated in most cases.

Figure 4.4

Alpha is 0.01, beta=0.01		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	50	75	100	175	300	300	500	500
Imp. Samp.	Half Heldout	50	75	125	150	200	300	500	500
Harm. Mean	Test	50	100	125	125	175	300	300	300
Harm. Mean	Half Heldout	50	75	100	125	175	200	200	300
MPE	Test	50	75	100	125	175	175	200	300
MPE	Half Heldout	50	75	100	125	175	175	200	300

Small α,β: overestimates K.

Figure 4.5

Alpha is 0.2, beta=0.2		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	25	50	50	75	50	75	75	100
Imp. Samp.	Half Heldout	25	50	50	75	75	75	75	75
Harm. Mean	Test	50	75	100	100	100	125	150	150
Harm. Mean	Half Heldout	25	50	100	100	100	125	150	150
MPE	Test	25	50	75	100	100	125	150	150
MPE	Half Heldout	25	50	100	100	100	125	150	150

Large α,β:underestimates K.

Figure 4.6

Alpha is 50/K, beta=0.01		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	25	50	75	100	125	150	175	200
Imp. Samp.	Half Heldout	25	50	75	100	125	150	175	200
Harm. Mean	Test	100	125	150	175	200	225	250	275
Harm. Mean	Half Heldout	25	50	75	125	125	150	200	225
MPE	Test	25	50	150	100	125	150	200	225
MPE	Half Heldout	25	50	150	100	125	150	175	225

Large, decaying α performs well.

Figure 4.7

Alpha is 25/K, beta=0.01		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	25	50	75	100	125	150	175	200
Imp. Samp.	Half Heldout	25	50	75	100	125	150	175	200
Harm. Mean	Test	25	125	200	300	300	300	300	300
Harm. Mean	Half Heldout	25	50	75	100	125	175	200	200
MPE	Test	25	50	75	100	125	175	200	200
MPE	Half Heldout	25	50	75	100	125	175	200	200

Smaller, decaying α performs well.

Complexity penalization. In some settings, adding additional free parameters will lead to overfitting that reduces predictive performance. As shown in Fig 4.1, we often do not observe over-fitting: likelihood continues to improve as we add more topics. In this case we may want to directly penalize additional free parameters using methods such as AIC. AIC is defined as -2 log likelihood + 2 (number of parameters). Smaller values are better. We also considered the BIC metric, but it produces the same results. A model with K topics, D documents, a V -word vocabulary has K(V-1)+D(K-1) free parameters in the basic LDA model. Optimizing symmetric Dirichlet hyperparameters and adds two additional parameters. For the purposes of held-out

inference we are only fitting the D(K-1) document-topic parameters (if we include the K(V-1) topic-word parameters, AIC will always recommend a 1-topic model). Figure 4.8 shows results for AIC. The number of parameters dominates the log-likelihood, causing the metric to prefer small numbers of topics in all cases. We show results for optimized hyperparameters, results for fixed hyperparameters are similar. One might argue that we are over-penalizing additional topics, which may in practice use significantly fewer than V-1 free parameters. We test an alternative specification where we treat multinomial parameters that are smaller than the corresponding prior as effectively zero, and exclude these from the total number of parameters estimated on the test set. The resulting "effective" number of parameters is used in the AIC computation. The effective number of parameters is smaller, but still grows linearly with the number of topics. AIC now picks larger but still inaccurate values of K.

Figure 4.8

AIC		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Imp. Samp.	Test	25	25	25	25	25	25	25	25
Imp. Samp.	Half Heldout	25	25	25	25	25	25	25	25
Harm. Mean	Test	25	50	50	50	50	25	25	25
Harm. Mean	Half Heldout	25	25	25	25	25	25	25	25
MPE	Test	25	50	50	25	25	25	25	25
MPE	Half Heldout	25	25	25	25	25	25	25	25

AIC underestimates K

Figure 4.9

AIC Effective Parameters		Actual Topics							
Metric	**Data Set***	**25**	**50**	**75**	**100**	**125**	**150**	**175**	**200**
Harm. Mean	Test	300	75	100	150	175	200	300	500
Harm. Mean	Half Heldout	25	75	100	150	150	175	200	300
MPE	Test	25	75	75	150	175	200	200	500
MPE	Half Heldout	25	75	100	150	175	200	200	500

AIC with only "effective parameters" overestimates K

3.4 Discussion

We find that held-out likelihood can be effective at estimating a number of topics under ideal circumstances, in which the model is correct, topics are well-separated, and hyperparameters are set appropriately. It is important to stress that none of these conditions hold in real data.

We find that fit metrics generally agree on an optimal K, but that this value is consistently too large, especially under hyperparameter optimization. The estimated number of topics grows roughly linearly with the correct number of topics, but is also consistently too large. In our experiments, using a constant sum of the alpha hyperparameters appears to work well, except when using the Harmonic Mean computed on the entire test set. The three options available — estimating hyperparameters from the data, using some arbitrary constants and choosing a fixed beta and holding the sum of alpha hyperparameters constant — all over or underestimate the actual number of topics. Metrics using effective number of parameters perform modestly better by consistently over-estimating but by a small amount.

Section 4: Model Convergence Properties

A Bayesian approach in estimation allows us to use some of the standard measures to evaluate model convergence. By focusing on aggregate likelihood (as we did in section 3) perhaps our analysis misses out on information contained in parameter level evaluation of the model. A standard Bayesian model evaluation approach would consider statistics[67] such as the Geweke statistic to check if a parameter is converged, the effective sample size to check whether adequate iterations have been estimated and Highest Posterior Density(HPD) intervals to check if the parameter is different from zero. We construct a variant of the last metric to check whether

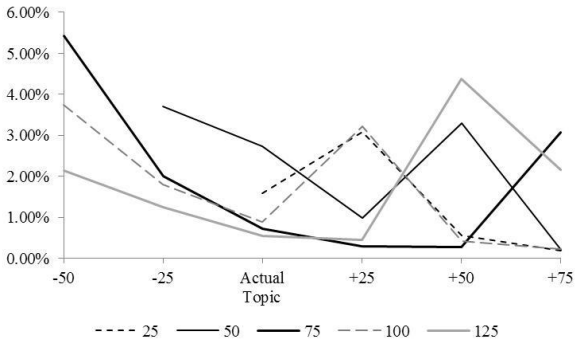
[67] see (Gelman et al., 2014)

the estimated parameter is different from random. In our application the latter is more relevant as parameters no different from 0 reflect useful information - whereas parameter estimates close to initial random values indicate the randomness of the parameter estimate. We have many thousands of parameters and we evaluate model convergence measures for all parameters for each model and report aggregate measures across topic distribution parameters (ϕ) for convenience. [68] Clearly, these measures rely entirely on the training data set.

4.1 Results

We present results for ϕ - but true for θ as well. Figure 4.10 describes the percentage of parameters that would be identified as not converged using the Geweke criterion. This figure (and others following this) maybe read as follows. The dashed black line represents the data generated from a model with 25 topics (25-topic). The x-axis reflects the different number of topics estimated on synthetic data in relation to the actual number of topics. +25 for the 25-topic data indicates 50 topics. The y-axis indicates the percentage of parameters converged.

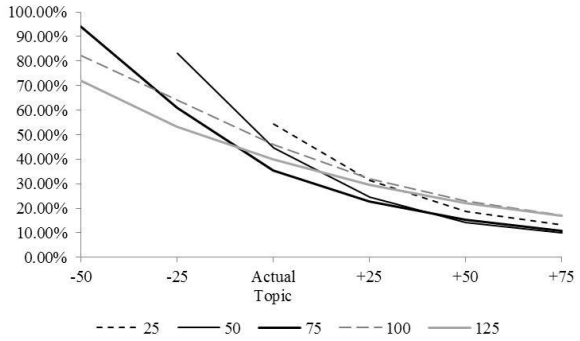
Figure 4.10



Percentage of Parameters Not Converged

[68] Similar results hold for document level topic proportions parameters.

Figure 4.11



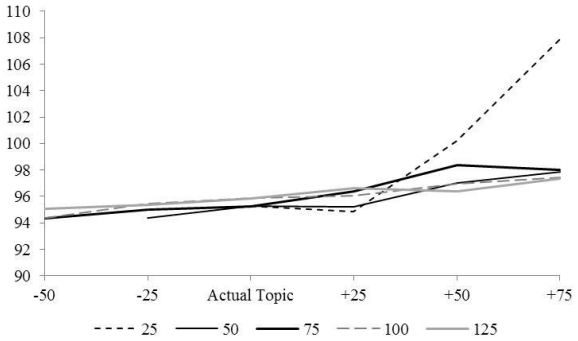
Percentage Effectively Random

The graph of the dashed black line starts at ”Actual Topic” for the 25-topic data. We observe that the percentage of not converged parameters goes up and then declines as we increase the number of topics estimated beyond 25. Similarly the grey line represents the data generated from a model with 125 topics (125-topic). Again, the percentage of parameters not converged declines and continues to decline beyond the actual number for topics and obtains a minimum at +25 or 150 topics.

Figure 4.11 uses the HPD criterion to identify percentage of parameters that are effectively random. This metric declines as the number of topics increases for each synthetic data set and overestimates the true number of topics. In figures 4.12 and 4.13, we report the average effective sample size and the percentage of the parameters whose sample size is less than 30. Clearly, the results suggest that neither of these metrics identify the correct number of topics for any of the synthetic data sets tested.

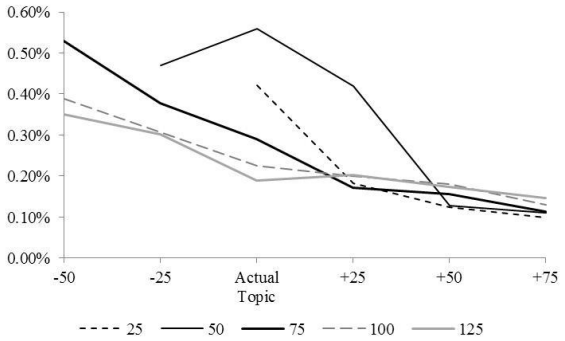
One concern is that perhaps low frequency terms in documents affect the robustness of the estimation. Our results do not change when we exclude low frequency terms.

Figure 4.12



Average Effective Sample Size

Figure 4.13



Effective Sample Size < 30

Section 5 Measures of Support

One purely empirical approach is motivated by the following intuition. If more topics are estimated than the correct number underlying topics, the additional topics should effectively be (uniformly) randomly distributed. We consider the Jensen Shannon distance to measure the distance from a uniform distribution for each topic.

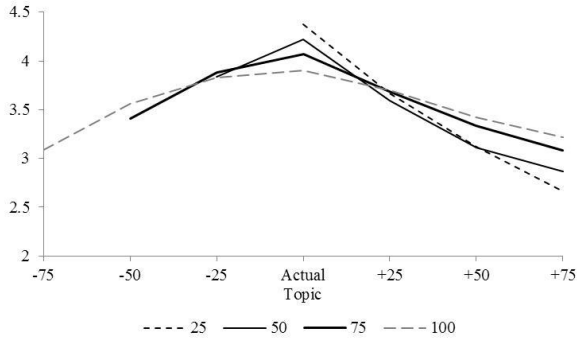
A similarly motivated idea is that most pairs of words in the top topic words should not appear in many documents for topics with little or no information. Consequently, additional number of topics should contain little or no information. We measure this in two ways - co

document frequency (the number of documents in which top words in a topic appear, pairwise) and coherence score(Mimno et al., 2011), that relies on the co-document frequency. As with model convergence metrics, these measures are reported on the training data set.

5.1 Results

Both figures 4.14 and 4.15 demonstrate inverted U-shapes - correctly identifying the number of topics. Co-document frequency and coherence scores yield similar results. Clearly, examining the underlying support does allow us to identify the correct number of topics. However, it must be noted that these results hold when we consider synthetic data generated using symmetric priors[69]. Data generated using asymmetric priors do not present similar inverted U- curves on these measures.

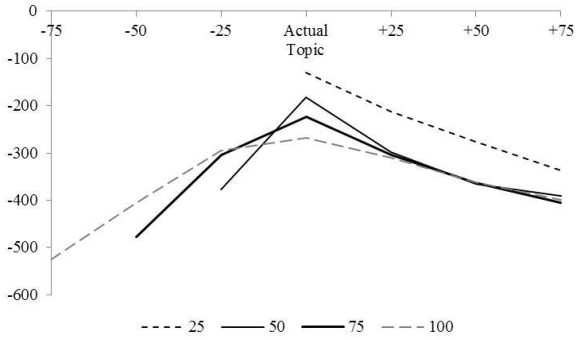
Figure 4.14



Average Distance From Uniform

[69] By symmetric priors, we mean that the topic distributions over the vocabulary have symmetric priors.

Figure 4.15



Average Coherence Score

Our investigation addresses three potential sources of error that might make the task of identifying the optimal number of topics difficult. The first is the choice of estimation framework - variational inference versus Bayesian inference. Our analyses suggest that, at least for the standard model, this choice does not affect the identification of optimal number of topics. A second source of potential error is the choice of priors. We considered both (arbitrarily) fixed priors and priors estimated from the data. Small fixed parameters do seem to work better. Examining parameter level convergence metrics reveals the possibility that some parameters have not converged. A researcher using these parameter estimates in the LDA model as measures is advised to check whether the parameters of interest have indeed converged or not. We Finally, we examine whether the support for the topic parameters offers insight on identifying the correct number of topics. Our examination of support for parameters identifies the particular conditions under which we may be able to correctly recover the number of topics and its (limited) applicability in general settings. It does, however, offer a measure of support that can prove helpful in determining the reliability of each topic.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. Bayesian data analysis, 2014.
Tom Griffiths and Mark Steyvers. Finding scientific topics. PNAS, 101(suppl. 1):5228–5235, 2004.

Andrew McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic mod-els. In Proceedings of the Conference on Empir-ical Methods in Natural Language Processing, pages 262–272. Association for Computational Linguistics, 2011.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Sometimes average is best: The importance of averaging for prediction using MCMC inference in topic modeling.

Mark Steyvers and Tom Griffiths. Probabilistic Topic Models. Handbook of Latent Semantic Analysis, 427(7):424–440, 2007.

Hanna M Wallach, David Mimno, and Andrew McCallum. Rethinking lda : Why priors matter. In NIPS, 2009a.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In ICML, pages 1105–1112, 2009b.