



Marketing Science

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors

Dinesh Puranam, Vishal Narayan, Vrinda Kadiyali

To cite this article:

Dinesh Puranam, Vishal Narayan, Vrinda Kadiyali (2017) The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors. Marketing Science 36(5):726-746. <https://doi.org/10.1287/mksc.2017.1048>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2017, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

The Effect of Calorie Posting Regulation on Consumer Opinion: A Flexible Latent Dirichlet Allocation Model with Informative Priors

Dinesh Puranam,^a Vishal Narayan,^b Vrinda Kadiyali^c

^aMarshall School of Business, University of Southern California, Los Angeles, California 90089; ^bNUS Business School, National University of Singapore, Singapore 119245; ^cSamuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York 14853

Contact: puranam@marshall.usc.edu,  <http://orcid.org/0000-0002-2373-9591> (DP); bizvn@nus.edu.sg (VN); kadiyali@cornell.edu (VK)

Received: December 17, 2013

Revised: October 31, 2014; April 3, 2016; January 14, 2017

Accepted: February 4, 2017

Published Online in Articles in Advance: August 21, 2017

<https://doi.org/10.1287/mksc.2017.1048>

Copyright: © 2017 INFORMS

Abstract. In 2008, New York City mandated that all chain restaurants post calorie information on their menus. For managers of chain and standalone restaurants, as well as for policy makers, a pertinent goal might be to monitor the impact of this regulation on consumer conversations. We propose a scalable Bayesian topic model to measure and understand changes in consumer opinion about health (and other topics). We calibrate the model on 761,962 online reviews of restaurants posted over eight years. Our model allows managers to specify prior topics of interest such as “health” for a calorie posting regulation. It also allows the distribution of topic proportions within a review to be affected by its length, valence, and the experience level of its author. Using a difference-in-differences estimation approach, we isolate the potentially causal effect of the regulation on consumer opinion. Following the regulation, there was a statistically small but significant increase in the proportion of discussion of the health topic. This increase can be attributed largely to authors who did not post reviews before the regulation, suggesting that the regulation prompted several consumers to discuss health in online restaurant reviews.

History: Pradeep Chintagunta, Dominique Hanssens, and John Hauser served as the guest editors and Robert McCulloch served as associate editor for this article.

Funding: The second author thanks NUS Business School for generous financial support for this research.

Supplemental Material: Data and the online appendix are available at <https://doi.org/10.1287/mksc.2017.1048>.

Keywords: Bayesian estimation • data mining • word-of-mouth

1. Introduction

In the face of rising obesity, Mayor Michael 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 the 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 (FDA). “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” (Jalonic 2013).

Past research has shown us that regulations pertaining to health claims on food labels affect consumer search and behavior in various ways (Roe et al.

1999, Bollinger et al. 2011, Downs et al. 2013). Unlike these papers, we focus on consumers’ postconsumption opinions of the product. Our data are 761,962 reviews of 9,805 restaurants in New York City, posted on a leading restaurant review site¹ in an eight-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 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 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 short- 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 of 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?

Our analysis provides the following managerial and regulatory insights: First, should there be a postregulation increase 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 et al. 2011, Ghose et al. 2012). That is, we might expect a greater discussion of health across a large number of online reviews to be accompanied by an increased 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 for commissioning costlier investigations into changes in consumer buying behavior. Fourth, how widespread health discussion is (i.e., variance in how deeply reviews discuss health as a topic, conditional on the mean level of health topics across reviews) can provide insights into 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. Note that there is also a 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 et al. 2003), which have been developed by computer scientists (specifically in the machine learning discipline) to analyze words in large sets of original text 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 the position of each word in a review, a topic is first chosen. Conditional on the choice of topic, a word is chosen from the vocabulary to take that position. 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. Estimation challenges arise because (a) we do not observe the topics, (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.

Because 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.³ 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. In sum, we combine current methods from computer science with causal inference techniques.

Within the existing research, our paper is related to the literature on extracting useful information from large masses of text of reviews (Decker and Trusov 2010, Ghose et al. 2012). Methodologically, our work is close to four papers. Lee and Bradlow (2011) automatically extract phrases from consumer reviews, which are then rendered into word vectors that record the frequencies with which words appear in the corresponding phrase. Phrases are clustered according to their similarity, measured as the distance between the word vectors. Netzer et al. (2012) use a similar approach, with the difference being that they define similarity between products based on their co-mention in the data. Tirunillai and Tellis (2014) apply the LDA model on consumer reviews to infer the latent dimensions

of product quality, to understand brands positions on these dimensions, and to estimate how dimensions and brand positions vary over time. Büschken and Allenby (2016) propose an LDA model that uses the sentence structure of reviews, leading to improved prediction of consumer ratings.

Our model generalizes previous marketing approaches 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 might alter the level of discussion of “price” as a topic when consumers review the focal product online. In the papers discussed above, words or phrases are allocated to topics, and topics are interpreted by the researcher. By contrast, our approach allows the analyst to prespecify 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 from 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 (Yaniv and Hogarth 1993, Wierenga 2006). For example, Blattberg and Hoch (1990) show that out-of-sample sales forecasts at the stock-keeping unit (SKU) level improve substantially when data pertaining to product characteristics and price are combined with managers’ intuitions about these forecasts.

Second, current marketing and computer science research assumes that the distribution of topics in a document is independent of the author’s decision on 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 focused solely on health are more likely to lead consumer opinion on health, and to be more important than the general reviewing population for targeting. To the extent that shorter reviews are more likely to discuss one or only a 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 (i.e., 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 postconsumption 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 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 postconsumption opinions.

Because the backdrop for the regulation is concerns over rising obesity, we can safely conjecture that there was a need for regulation since consumers are likely not as health-aware. Therefore, our expectation is that before regulation, the discussion of health topics in reviews was likely to be small; thus, 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 increased 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 postregulation increases in the health topic discussion in restaurant reviews might be cause for celebration and offer 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.

Next, we discuss the model specification. Section 3 presents the data, and discusses specific estimation challenges. Section 4 presents the results from the model, and their implications. Section 5 provides concluding remarks.

2. The Model

2.1. Model Specification

We start with some standard definitions and notation. A corpus is a concatenation of all documents (i.e., reviews) in the data set. Each document d ($d = 1, \dots, D$) is composed of n_d words. The number of total word instances in the corpus is N . The corpus is therefore defined by an N -dimensional vector $\mathbf{w} = \{w_{11}, w_{12}, \dots, w_{d1}, \dots, w_{Dn_D}\}$, where w_{di} is the i th word of document d . A vocabulary is the set of V unique words across all documents; each word is denoted by v . Each word in the vocabulary belongs to each topic, such that topic ϕ_k ($k = 1, \dots, K$) is a probability distribution over all v in $\{1, 2, \dots, V\}$. Element ϕ_{kv} denotes the probability of word v given topic k . Document d is a mixture

of the K topics. A K -dimensional vector (θ_d) represents the proportions of each topic in document d .

We briefly describe our textual data to enable a 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 et al. 2009). After eliminating stop words ("a," "the," etc.) and words that occurred fewer than five 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. We discuss the robustness of our results to this assumption in Section 4.3.

The LDA model assumes a generative process by which the textual data in each document is generated. Topic proportion θ_d is drawn from a Dirichlet distribution with K -dimensional parameter vector α_d . The topic assignment z_{di} for the word i in document d is drawn from a categorical distribution with parameter θ_d . 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 with hyperparameter β . This process is repeated for each word in each document. It ignores the order of words in a document, i.e., LDA is a "bag-of-words" model (Eliashberg et al. 2007, Netzer et al. 2012). The generative process for document d can be summarized as follows:

- (a) $\theta_d | \alpha_d \sim \text{Dirichlet}(\alpha_d)$.
- (b) $z_{di} | \theta_d \sim \text{Categorical}(\theta_d)$.
- (c) $w_{di} | (z_{di} = k), \phi \sim \text{Categorical}(\phi_k)$
where $\phi = [\phi_1, \phi_2, \phi_3, \dots, \phi_k]$.
- (d) $\phi_k | \beta \sim \text{Dirichlet}(\beta)$.

The researcher does not observe the topics, the 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. The process described above defines a joint probability distribution over 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 document characteristics that might affect the distribution of topics. Longer documents might discuss more topics and therefore could 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 levels of author experience might have similar topic distributions. Online reviews tend to be disproportion-

ately 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 five-point ordinal scaled numerical rating), and the past reviewing experience of the author (measured by the number of reviews in our data set posted by the author of document d , before she posted document d). This leads to the following specification:

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

where 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 that 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. If longer documents involve the discussion of more topics entailing a more evenly spread topic distribution, we expect λ_{k1} to be positive.⁴

Next, we address the central question of this research: How did the proportion of health topic discussion in online reviews change due to the regulation? Because the regulation was implemented for chain restaurants only, standalone restaurants fall outside of our domain of substantive interest, except as a control group for inferring the causal effect of the regulation on chain restaurants (which form the treatment group). To draw causal inference, we must 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 in the treatment group, and then compare this to changes in the *exact same topic* in 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:

$$\theta_{kd} = v_{0k} + v_{1k}Chain_d + v_{2k}Post_d + v_{3k}Chain_d \cdot Post_d + v_{4k}ZipCode_d + v_{akd} + v_{fkd} + \epsilon_{kd}, \quad (2)$$

where θ_{kd} is the proportion of topic k in document d (the outcome variable), and $Chain_d$ is the dummy variable that accounts for the effect of belonging to the treatment group. It controls for unobserved factors that might affect topic proportions of chain and standalone restaurants differently. The dummy variable $Post_d$ accounts for the effect of the treatment. It is possible that at the time of implementation of the regulation, there were unobserved events that affected topic proportions of all restaurants (including chains) in New York City. The main effect of $Post_d$ controls for how topic proportions for *all* restaurants changed after implementation of the regulation. The coefficient of the interaction term ($Chain_d \cdot Post_d$) captures the crucial effect of the treatment on the treatment group. Furthermore, we control for spatial variation (across locations in New York City), and temporal variation (over the duration of the data), in the 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. The term $ZipCode_d$ 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 for topic k is captured by⁵ v_{akd} and the random effect of restaurant f is captured by v_{afd} . Both effects are assumed to be normally distributed with zero mean.

We first estimate the model on reviews posted over four years (i.e., 16 quarters) starting July 1, 2006 (we show robustness to the choice of time periods in Section 4.3). This period was chosen so that the duration of time before and after July 1, 2008 (date of regulation implementation), is the same. Sixteen quarters correspond with 16 dummy variables, one for each quarter. The dummy variable corresponding to the quarter in which document d was posted takes the value 1; all others take the value 0. Furthermore, of the 16 dummy variables, 8 correspond to the post regulation period. The eight variables sum to the variable $Post_d$. So, to avoid collinearity issues, we drop one of these eight variables. Similarly, the remaining eight dummy variables correspond to the pre regulation period. These eight variables sum to $(1 - Post_d)$. So, we drop one of these eight dummy variables to avoid collinearity. Error terms are assumed independent and identically distributed (IID) and normally distributed. All parameters are topic-specific.

2.2. Model Estimation

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 IID, the likelihood of the data conditional on the hyperparameters is calculated as follows:

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

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-form 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. The dimensionality problem stems from 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, $P(z|w)$ based on the equation $P(z|w) = P(z, w) / \sum_z P(z, w)$. The numerator of the right-hand side (RHS) of this equation can be factorized and simplified as $P(z, w) = P(w|z)P(z)$. We now turn our attention to $P(w|z)$ and $P(z)$. Given the conjugacy between the distribution of observing word w given topic k and the Dirichlet prior ($\phi_k | \beta \sim \text{Dirichlet}(\beta), \forall k$), the posterior distribution of $P(w|z)$ is as follows:

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

where $\Gamma(\cdot)$ is the standard gamma function. In Equation (4), 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 ($\theta_d | \alpha_d \sim \text{Dirichlet}(\exp(x_d^t \lambda_k))$) yields document specific topic assignments. These topic assignments are conditionally independent across documents and can be multiplied to yield

$$P(z) = \prod_{d=1}^D \left[\frac{\Gamma(\sum_k \exp(x_d^t \lambda_k))}{\prod_k \Gamma(\exp(x_d^t \lambda_k))} \right] \cdot \left[\frac{\prod_{k=1}^K \Gamma(\exp(x_d^t \lambda_k) + n_{kd})}{\Gamma(\sum_k \exp(x_d^t \lambda_k) + n_d)} \right], \quad (5)$$

where 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 a Markov Chain Monte Carlo (MCMC) approach that relies on Gibbs sampling of the latent topic assignment variable z (Griffiths and Steyvers 2004).⁶ 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.⁷ The Gibbs sampling scheme for (2) can be derived from (4) and (5) as follows:

$$P(z_i = j | z_{-i}, w)_d \propto \frac{n_{j,-i}^w + \beta}{V\beta + n_j - 1} \frac{n_{j,-i}^d + \exp(x_d^t \lambda_j)}{\sum_k \exp(x_d^t \lambda_k) + n^d - 1}. \quad (6)$$

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 to 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 the corpus and the document.

We now describe how λ_k is estimated. For each topic k , λ_k is a topic specific $M + 1$ dimensioned vector of parameters, with prior distribution $N(0, \sigma^2 I)$.⁸ The joint probability of z and λ can be written as follows:

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

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 (EM) (Mimno and McCallum 2008). This 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 $\lambda_{k,m}$. Instead of assuming that β is known (Mimno and McCallum 2008), we explicitly estimate the hyperparameter β , in addition to λ , to address this issue. See Online Appendix 1 for details.

Finally, we estimate Equation (2) in 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 MALLLET⁹ 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 1,500 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. Identical results were obtained by computing the moments from 5,000 posterior draws.

2.2.1. Seeding. Our model 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. 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 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.

To initialize our data set, we randomly assign each word in each document to one of K topics. Note that in this stage of the estimation process, topics do not 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 . 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 1,000. Dividing the number of times each word was assigned to topic 0 by n_k yields a probability distribution over the vocabulary.

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. We call these incremental units 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).¹⁰ After some experimentation, we chose five as the value of this pseudocount in our 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 4.3, we discuss the robustness of our results to the choice of seeds.

3. Analysis and Evaluation

We start by reporting the top 10 words across all reviews for chain restaurants in our data. In Figure 1(A), we present their ranking, based on the frequency with which they appear in each of the four groups of interest: chain restaurants preregulation, chain restaurants postregulation, standalone restaurants preregulation, and standalone restaurants postregulation. For example, the word “salad” is the eighth 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 of 44,276 unique words in our corpus. Extending this

analysis to 100 words did not reveal any other substantively useful insights.¹¹

Next, we present a 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, and nutrition. We present the frequencies of occurrence for each month divided by the number of reviews of chain restaurants posted in that month (Figure 1(B)). 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 as to changes of consumer opinion due to the regulation. Because 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 also not helpful. First, 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 corpora.

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 can 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 used in marketing, do not correlate well with human judgments of topic interpretability (Chang et al. 2009). This has led to an increased interest among computer scientists in developing automated metrics that can better predict topic interpretability. A useful insight from this research is that if a topic is highly interpretable (to humans), pairs of words with a high probability of association with this topic 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 coherence metric for each topic.

Figure 1(A). Most Frequently Occurring Words by Restaurant Type and Time Period

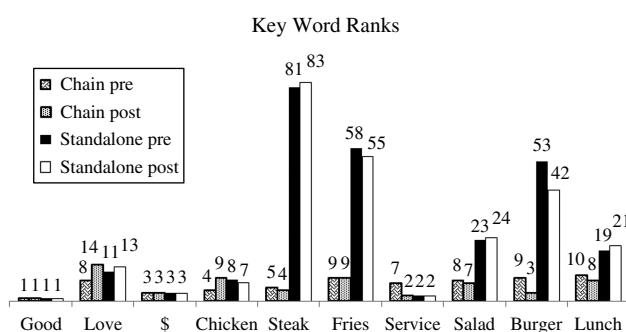
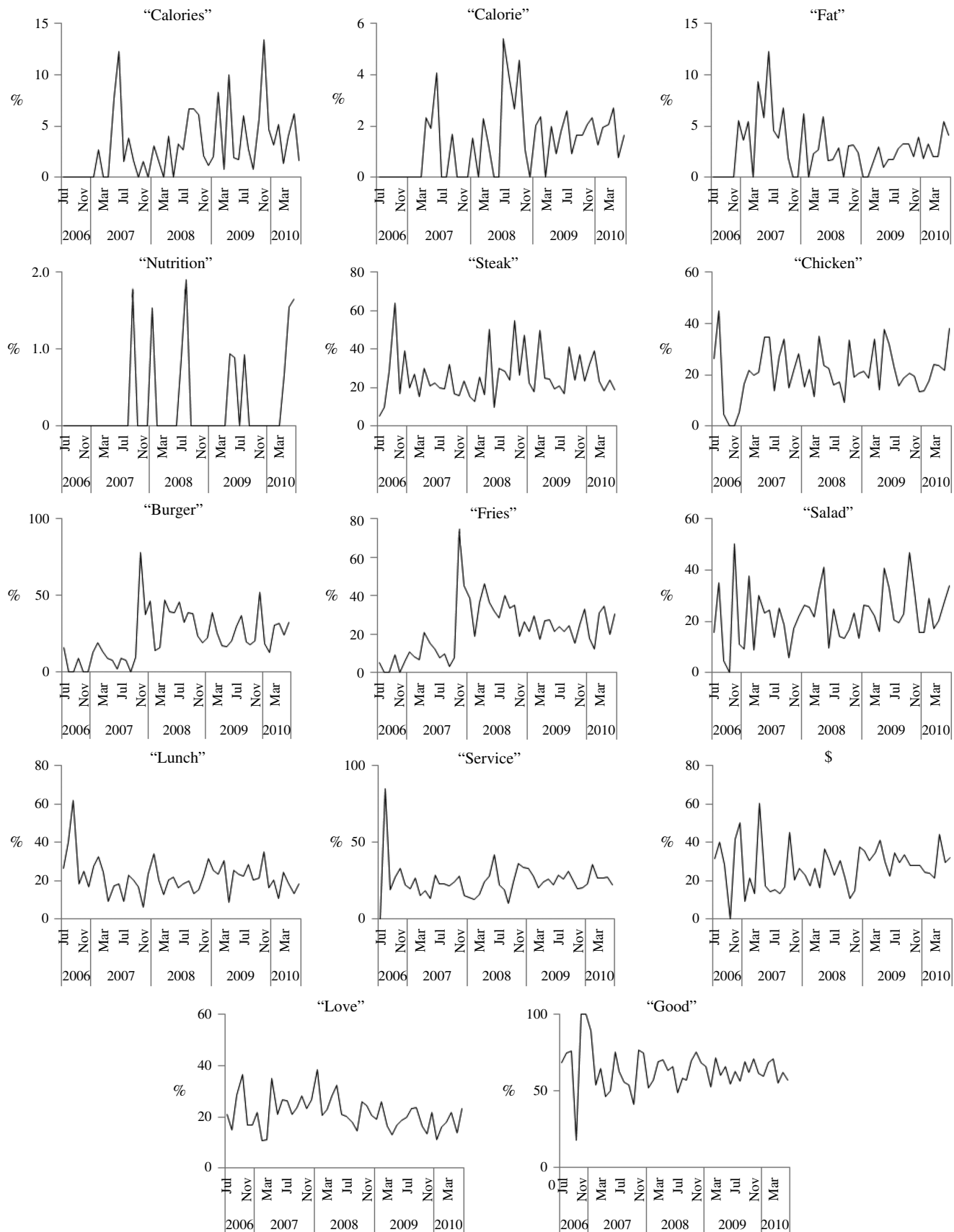


Figure 1(B). Monthly Frequency of Occurrence of Words in Reviews of Chain Restaurants Per Review (July 2006–June 2010)



We now discuss how we choose the number of topics (K) and label each one. 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 is a probability distribution across the vocabulary. We use 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 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. Because not all topics are of substantive interest, 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(A), i.e., the seeded “health” topic discussed earlier, and 19 topics with the greatest values of the topic coherence metric (henceforth referred to as the top 19 topics).¹² Coherence scores of all other topics are available from the authors. Following convention, 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 with the greatest posterior probability of belonging to that topic (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.

4. Results and Implications

4.1. Results

We first discuss the major topics of interest in our data. In Table 1(A), we present the top 10 words in decreasing order of posterior probability of being in each of the 20 topics, as inferred from the analysis of all reviews of chain restaurants. These comprise 19 topics that did better than other topics on established coherence metrics, and the “health” topic.¹³ Each topic is labeled by concatenating the two most probable words in the topic. First, we find that a substantial number of topics focus on specific menu items (e.g., steak:potatoes, bbq:chicken, burger:fries). Second, several topics focus 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, and access). Topic 14 alludes to nonfood related restaurant services (e.g., waitress, seated, check, server).

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 . Because

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 are discussed to a greater extent than the average topic (see Table 1(B)). 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%. We find that just 6.8% of such reviews contain the health topic to an 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. 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%). Figure 2 shows the relative frequency of occurrence of the top 20 words in each of the top 20 topics. See Online Appendix 2 for a list of the rest of the topics and Online Appendix 3 for further details on the extent of the discussion on each topic.

Figure 3 shows the temporal trends over a 48-month window of the posterior 95% of 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). 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 postregulation, is difficult to infer from these charts. See Online Appendix 4 for temporal trends in proportions for the top 19 topics.¹⁴

Next, we present the results of the difference-in-differences analyses for the health topic as specified in Equation (2), with key parameters presented in Table 2. Several insights emerge. First, after controlling for chain characteristics (compared to standalone restaurants), temporal trends, restaurant locations, and time-specific shocks, we find that the proportion of the

Table 1(A). Major Topics, Associated Words, and Topic Proportions (Based on Coherence Scores)

ID	Name	Coherence score for unseeded topics	Top 10 words in decreasing order of posterior probability of belonging to the topic	Mean topic proportion in chain restaurants (%)
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, creme, 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

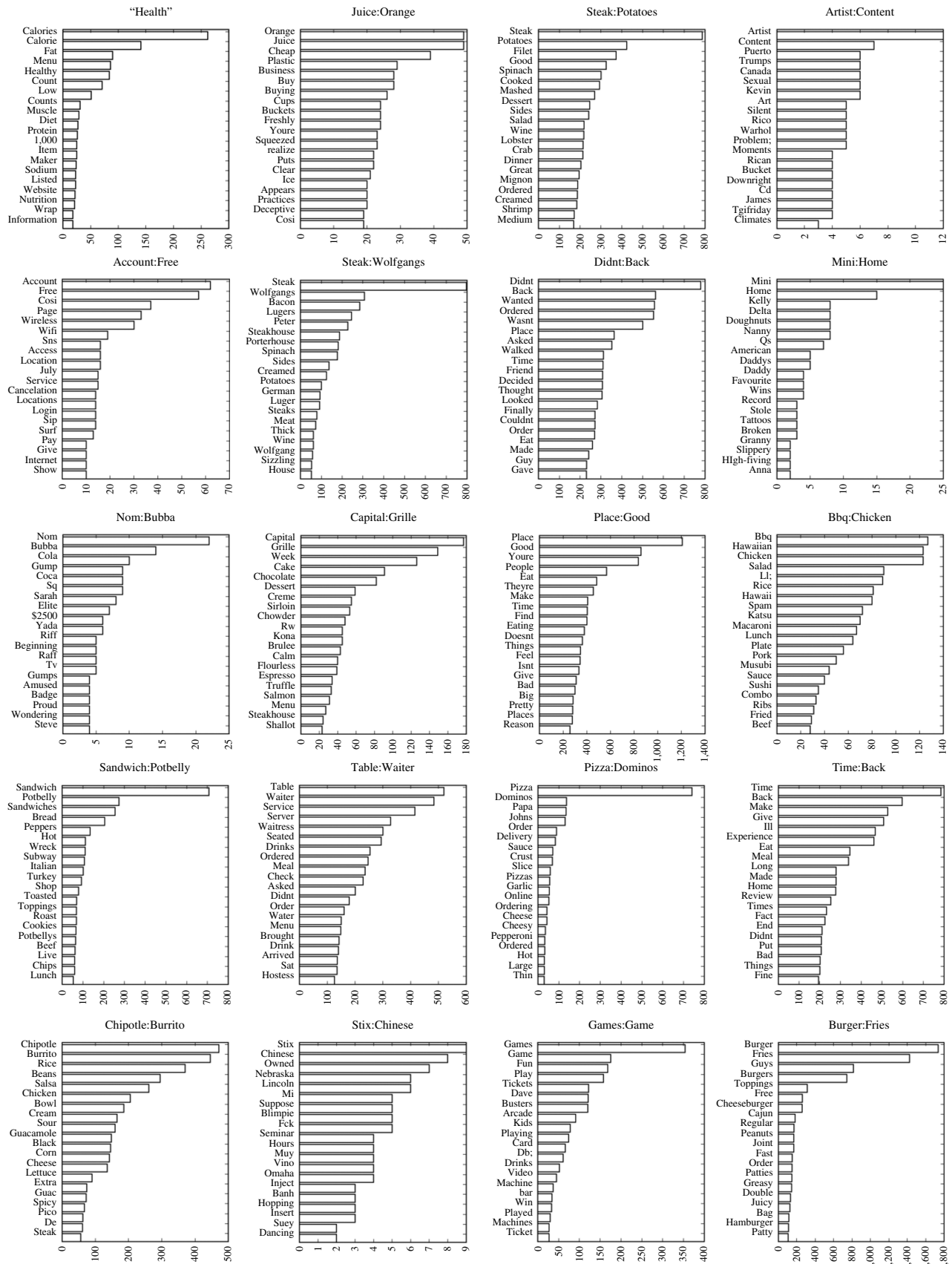
Table 1(B). Major Topics, Associated Words, and Topic Proportions (Based on Topic Proportions)

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

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 supporting the success of the regulation.

The proportions of topics connoting high calorie foods such as “steak:potatoes” and “burger:fries” also increase after implementation of the regulation. From

the coefficients of $Post$, it is evident that topics related to brands such as Potbelly (sandwich:potbelly) and Chipotle (chipotle:burrito) garner a lower proportion of online reviews after 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. However, we cannot draw inferences about

Figure 2. A Visual Representation of Major Topics

Note. Bars represent the frequencies of occurrence of each of the top 20 words in a topic.

Figure 3. Temporal Trend of the Posterior 95% Credible Intervals of the Mean of Topic Proportions (Within a Month) for the Health Topic

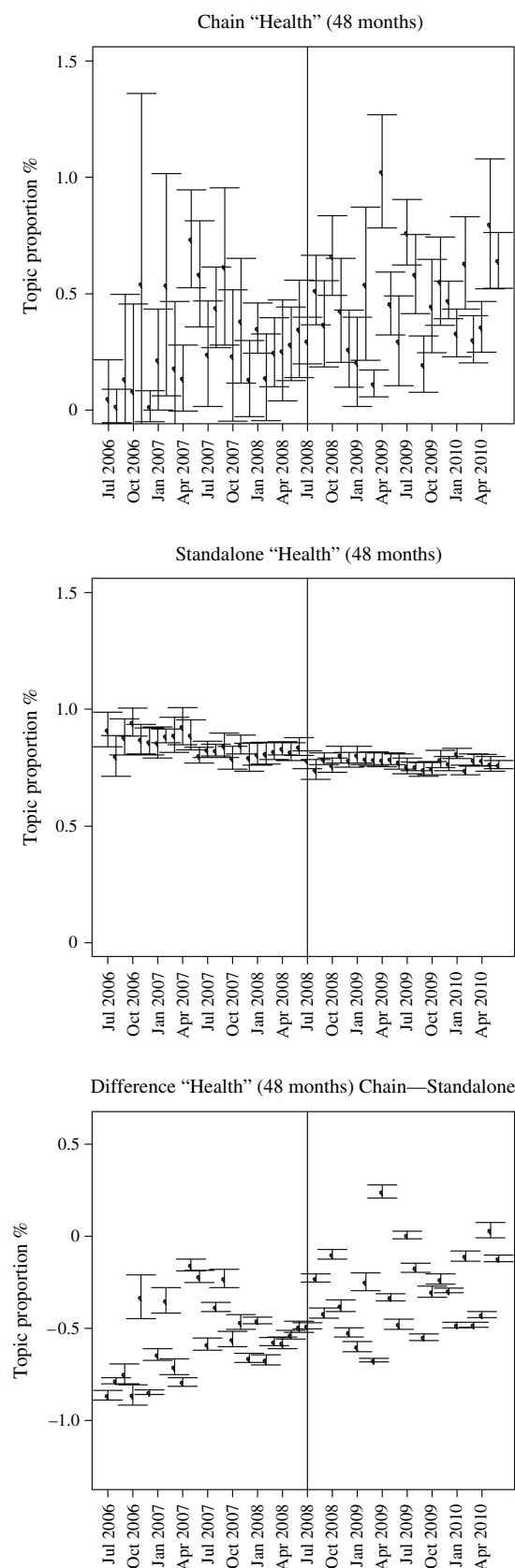


Table 2. The Effect of the Calorie Posting Regulation on Topic Proportions

Topic	Mean topic proportions		
	Coefficient of Chain	Coefficient of Post	Coefficient of Chain × Post
“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

Notes. 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. Posterior Means and Posterior SDs of Coefficients of Equation (2).

health discussion from topics other than the seeded health topic. The health topic, irrespective of whether it is seeded, contains the words calorie, calories, and health with high probabilities, and thus seems pertinent to studying the effect of the 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 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 unsuitable for drawing inferences about health.

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 who are 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 chain restaurant review 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 the reviews exceeds the overall mean health topic proportion (0.41%) in all

Table 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

reviews of chain restaurants. In Table 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. This suggests that the increase in health topic proportion post regulation can be traced to a small group of authors.¹⁵ We then examined the health topic proportions of reviews posted in each of the four cells (high/low health discussants versus pre/postregulation). We find that the health topic proportion substantially increased post regulation in reviews posted by “high” health discussants (who 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 by 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 did not post any review in the preimplementation 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 preimplementation period. This suggests that the overall health discussion in the post implementation period and any increase from the pre period are driven by reviewers who did

not review chain restaurants on the focal website before the regulation. In summary, we find that the 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.

Finally, we discuss how the hyperparameter α_{kd} that affects topic distributions for the health topic is affected by observed characteristics of reviews and the author. Parameter estimates pertaining to Equation (1) are shown in Table 4. As expected, we find that longer reviews have lower values of this parameter, suggesting a more evenly spread distribution of the health topic. The more negative the valence of the review, the lower is the value of α_{kd} . This suggests 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.¹⁶

4.2. Dynamic Topic Model

The proposed model assumes that the topics 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 is possible that the words representing a topic may evolve with time due to changes in word use.¹⁷ To verify this, we estimated a Dynamic Topic Model (Blei and Lafferty 2006) by allowing ϕ_{tkv} , the probability of word v given topic k in time period t to depend on the natural parameter of the same topic in the previous time period as follows: $\phi_{tkv} | \phi_{t-1,kv} \sim N(\phi_{t-1,kv}, \sigma^2 I)$, where I is a V -dimensional identity matrix. We incorporated the review specific characteristics in the model as outlined in Section 2.1. This model may reveal interesting patterns of topic evolution, and may also alleviate any potential biases that could arise from assuming 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 eight years from January 2005 to December 2012. Of these reviews, 10,779 were posted by 7,156 authors from 2006 to 2012. Because reviews were sparsely posted in 2005, we drop 44 from 2005 for estimation of the dynamic topic model. We use reviews for every month post 2005 until December 2012 for estimation. Because of the sparsity of reviews at the weekly level, we use month as our unit of analysis. We modeled topic dynamics from January 1, 2006 until December 31, 2012 at the monthly level. To estimate this model, we

Table 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

Table 5(A). Temporal Evolution of the “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
	Fruit	Fruit	Fruit	Love	Love	Love
Least						

Table 5(B). Temporal Evolution of the “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
	Peanuts	Peanuts	Greasy	Greasy	Greasy	Greasy
Least						

closely followed the VI technique in Blei et al. (2003).¹⁸ We found that an 8-topic model maximizes model fit per the Jensen–Shannon statistic. So this model in effect requires us to estimate eight topics in each of the 84 months (which is effectively $8 \times 84 = 684$ topics). Although these topics are not independent, they are far more numerous than the 200 topics in the static model. Table 5(A) 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 the top 20 words for a topic that is ranked high by coherence and importance metrics in our data set (burger:fries) in Table 5(B). We find little 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 dynamic model estimates, the resulting measure of health discussion in chain restaurants increases from 0.38% in the pre period to 0.45% in the post period; this is 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 a similar fit to the data. Perplexity is a measure of fit for LDA models and is described in more detail in Section 4.5.

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 Code¹⁹ makes the calorie posting regulation effective from May 5, 2008. This code states that the New York City Health Department 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 before 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 implemented. We estimated the model including Equation (2) for different temporal breaks (before and after July 1, 2008), and find that our results hold (Table 6(A)).

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, e.g., 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 could affect topic proportions of chain restaurants only. We estimate the regressions for $X = 24$ months (i.e., on all reviews posted 24 months before and after the regulation), $X = 18$, $X = 12$, and $X = 6$ months. Although the regression coefficients vary in magnitude (Table 6(B)), the coefficients' signs remain the same, showing a positive effect of the regulation from the 12th month onward. The null effect for shorter time 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 by consumers taking time to absorb the information from calorie posting.

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, i.e., the large difference in mean health topic proportion between the

Table 6(A). Difference-in-Differences Analyses with Varying Policy Implementation Dates

Model	Dependent variable = Health topic proportion		
	Coefficient of Chain	Coefficient of Post	Coefficient of Chain \times Post
Policy date			
July 1, 2008			
M	7.70	-1.18	4.66
SD	1.78	0.24	1.90
Policy date			
October 1, 2008			
M	7.80	-0.99	4.70
SD	1.60	0.22	1.74
Policy date			
April 1, 2008			
M	7.80	-1.68	4.44
SD	1.94	0.26	2.05

Notes. 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.

Table 6(B). Difference-in-Differences Analyses with Varying Time Window Around Policy Implementation

Model	Dependent variable = Health topic proportion		
	Coefficient of Chain	Coefficient of Post	Coefficient of Chain \times Post
± 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

Notes. 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.

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 of chain restaurants). For this purpose, we sampled preregulation 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 all preregulation 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 *Post* (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 random), it shows that our results are not affected by differences between chain and standalone restaurants.²⁰

Increase in number of authors and reviews. There are many more unique authors and reviews in the post-period compared to the preperiod for chain and standalone restaurants (Table 3). One concern is that this growth in volume of new authors or reviews might be leading to the increased discussion of “health” in chain restaurants. It is unlikely that this sharp increase in the number of new authors stems from the regulation itself. Indeed, we find in our data that the number of authors posting reviews for the first time increases substantially over time for both types of restaurants.

Our main concern is whether the regulation led to a change in this trend for chain restaurants, which was different from the change for standalone restaurants. To analyze this, we regressed the number of new authors in month t for chain and standalone restaurants ($t = 1, \dots, 48$) 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 important, 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 reviewing chain restaurants.²¹

We repeated the same regression analysis with the total number of authors as the dependent variable, and obtained the same result. Regressing the number of reviews 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.

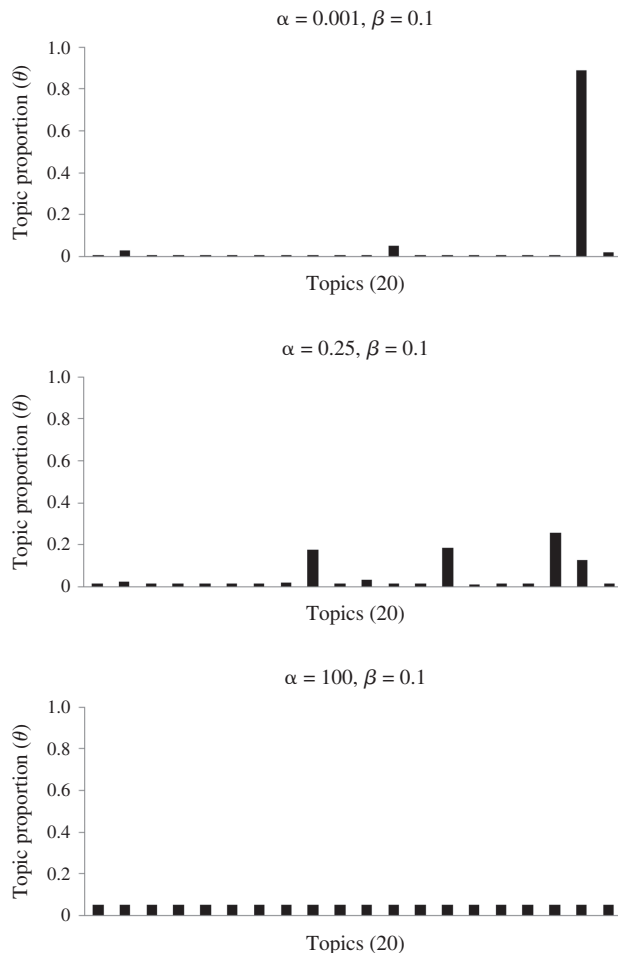
Choice of seeds. To understand if our results are driven by our choice of seed words from the regulation document, we estimated the model using another set of seed words. Specifically, Wikipedia (one of the most widely used sources of information globally) contains a description of the word “nutrition” and lists a table of nutrients that should be consumed for healthy living (<http://en.wikipedia.org/wiki/Nutrition>). We used all of the words describing the nutrients in this table as seeds. These seed words are calories, fat, protein, vitamin, cholesterol, calcium, magnesium, sodium, potassium, iron, and iodine. We reestimated the model using these words as seeds, and found that our results are very similar to those obtained using seeds from this source instead of those from the regulation document. Topic proportions did not vary based on the choice of seeds.²²

Document specific hyperparameters. To demonstrate how allowing for document specific hyperparameters α_{kd} affects the within-topic distribution, we estimated the model on online reviews of all chain restaurants in our data set, 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 $\beta = 0.1$. Topic proportions for the three models (corresponding to three values of α_{kd}) are presented in Figure 4. Larger values of α are clearly associated with more evenly spread topic distributions.

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,²³ 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 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

Figure 4. Distribution of Topic Proportions for a 20-Topic Model for Various Values of α 

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: 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 seven 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 provides a basis for conducting further (and costlier) studies into consumption of healthier prod-

ucts 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 the focal website) after the regulation. Policy makers might be interested in understanding what kinds of reviewers were 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.

Although 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 the average annual weight gain in the entire county population would be averted. Reduction in obesity levels has a monumental social and economic significance in the United States where over 250,000 deaths every year are attributable to obesity (Allison et al. 1999). Obesity related costs in the United States in 2008 were estimated to be a staggering \$147 billion (Finkelstein et al. 2009), and are 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 that consumers use for choice decisions, this serves as a free and externally valid input into product management decisions. For trade-offs between investing in service or menu redesign, it is useful for managers to know that menu items are 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 also important for health regulators to understand whether there are differential effects of the health regulation. Especially, regulators may have a goal that the regulation not leave behind less healthy consumers, i.e., the regulation benefits not just the healthier consumers. In the absence of disaggregated data

on health measures at the zip code level, researchers have often used race as a proxy (Boardman et al. 2005), given the greater prevalence of food-correlated diseases (specifically obesity, hypertension, and diabetes) among African Americans. We separated reviews from African American majority zip codes²⁴ (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% afterward.²⁵ By contrast, 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.

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% for 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 ($M = 0.88$, $SD = 0.51$). This seems to indicate that the regulation had a (10% level of significance) larger impact on potentially less healthy zip codes.

However, replacing the binary race variable with a continuous measure (proportion of African American population) reduces the statistical significance even further. Although the evidence that an increase in health discussion of chain restaurants is greater for African American neighborhoods is not very compelling from a statistical standpoint, we can confidently say that the regulation did not differentially favor healthy zip codes or leave behind less healthy zip codes. This should be reassuring to regulators, even if a larger effect of less healthy zip codes could be more reassuring.

Finally, our analysis reveals useful insights for brand managers of restaurants. Topics in Table 1(A) show words that 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 serve as a useful input for future store choice decisions where one brand might want to avoid proximity to the other. Food items frequently mentioned with a brand indicate the items with which that brand is associated. Based on this, Chipotle is more strongly associated with burritos and chicken, and not as much with tacos or beef. This could serve as input into a formal menu planning exercise, i.e., more items related to burritos and chicken might strengthen these brand-product associations.

4.5. Model Comparison

We assess improvement in model performance due to the incorporation of two features unique to the marketing literature, i.e., (a) allowing the researcher to seed certain topics with specific words that 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, i.e., the likelihood of observing a collection of words given a model of how the words were generated. 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).²⁶ Therefore, seeding enables incorporation of managerial intuition and offers much richer managerial insights, at very little cost in terms of model fit. Note 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. Model B performs worse than the proposed model, providing empirical validity to the notion that longer reviews and reviews with more positive valence have a more even distribution of topics.

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 consumers’ opinions (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., eight 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. 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 knowledge. A recent report by the McKinsey Global Institute (Manyika et al. 2011, p. 23) 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 focused on inferring market structure and product attributes in specific product categories, ascertaining the extent to which these correlate with consumer level data collected from more traditional experimental and survey based techniques, and incorporating measurements of such data in demand models. We extend this work 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 LDA set of models in computer science. Future research can explore more flexible heterogeneity specifications of the hyperparameter α_{kd} (e.g., restaurant and author-specific random effects) and allow for correlated probabilities of words belonging to a topic, across time periods, in the dynamic topic model. We look forward to several strategy- and policy-relevant applications as well as more sophisticated models in this area of topic detection and measurement.

Acknowledgments

The authors thank Claire Cardie, David Mimno, and Anirban Mukherjee for useful comments. The authors also thank seminar participants at INSEAD, Georgetown University, the QME Conference 2014, and the Marketing Science Conference 2014.

Endnotes

¹Per 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.

²Despite large sample sizes, user-generated textual content might 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 free, in larger quantities, and over longer periods of time.

³We 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.

⁴Topic 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.

⁵There are 1.5 reviews per author on average, which restricts our ability to identify author specific effects. To address this issue, we assume the same random effect for authors who posted the same number of reviews in our data.

⁶Variational Inference (VI) methods are also commonly used 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 as they are more common in the marketing literature.

⁷For example, $P(\phi_k | z, w) = \text{Dirichlet}(\beta + n_k^1, \dots, \beta + n_k^V)$ and the estimated mean vector of this distribution from a single MCMC iteration is $((\beta + n_k^1)/(V\beta + n_k), \dots, (\beta + n_k^V)/(V\beta + n_k))$.

⁸Directionally, 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.

⁹<http://people.cs.umass.edu/~mccallum/mallet/>.

¹⁰For 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. It features in another topic with the following top 10 words: health, department, properly, unclear, dirty, roach, grade, equipment, NYC, contact. This topic relates to health safety and hygiene and is unrelated to the calorie content of food. The data overwhelmed the prior and assigned “health” to another topic.

¹¹Because 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, 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.

¹²In 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 Zipf’s law to apply; most of the probability mass in each topic is allocated to a few words. Using this measure did not change the results in the paper.

¹³To improve interpretability, it might be tempting to combine topics that appear similar. This is better 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 a poorer fit. Thus, we follow the standard approach of drawing substantive inferences from the best fitting model.

¹⁴Figure 3 shows the large variation in topic proportions within a month or quarter. To account for this within-period variation, we replaced the quarterly dummies in Equation (2) with the following variable: the posterior standard deviation across reviews posted in that quarter, of the topic proportion θ_{kd} . Our results remain unchanged. Replacing this variable with measures of monthly standard deviation also does not affect our results.

¹⁵Although it is plausible that “low” health discussants contribute to the health discussion by posting more reviews (per author) than “high” health discussants, we find the opposite: “Low” health discussants post an average of 1.43 reviews in the data 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.

¹⁶The 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. Because all documents in our application are authored by a single individual only, we did not adopt this approach.

¹⁷We thank an anonymous reviewer for highlighting this possibility.

¹⁸This 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 $(\pi_{1,k}, \dots, \pi_{i,k}, \dots, \pi_{V,k})$ is $\beta_{i,k} = \log(\pi_{i,k}/\pi_{V,k})$. Mean parameters $(\pi_{k,i})$ for a topic in any period can be recovered using this result.

¹⁹<https://www1.nyc.gov/nycbusiness/description/calorie-posting> (accessed March 2016).

²⁰Our 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 showed that most of these authors also post reviews of standalone restaurants. This suggests overlaps in consumer segments across the two types of restaurants.

²¹These results are robust to inclusion of the following additional covariates in the regression: t^2 , t^3 , *Chain*, and *Post*. Inclusion of more covariates leads to a larger standard error of the coefficient of t_Chain_Post .

²²Because our interest is specific to the discussion of health, seeding only one topic is sufficient in our context. Recent advances in the LDA literature (Jagaramudi et al. 2012) allow for seeding multiple topics, such that each topic is a mixture of two distributions, i.e., a seeded topic and a regular topic. Such methods are more appropriate for contexts with multiple topics of interest.

²³<https://www.health.ny.gov/statistics/chac/general/g74.htm>.

²⁴2010 census conducted by the U.S. Census Bureau. <https://www.census.gov/2010census/>.

²⁵Zip codes capture resident demographics. We assume this is also a proxy for restaurant customer demographics.

²⁶We prefer a seeded model to an unseeded model as 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.

References

- Allison DB, Fontaine KR, Manson JE, Stevens J, VanItallie TD (1999) Annual deaths attributable to obesity in the United States. *J. Amer. Medical Assoc.* 282(16):1530–1538.
- AlSumait L, Barabá D, Gentle J, Domeniconi C (2009) Topic significance ranking of LDA generative models. Buntine W, Grobelnik M, Mladenić D, Shawe-Taylor J, eds. *Proc. Joint Eur. Conf. Machine Learn. Knowledge Discovery Databases: Part I, Lecture Notes Comput. Sci.*, Vol. 5781 (Springer, Berlin Heidelberg), 67–82.
- Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.
- Bird S, Loper E, Klein E (2009) *Natural Language Processing with Python* (O'Reilly Media Inc., Sebastopol, CA).
- Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model + 50% manager. *Management Sci.* 36(8):887–899.
- Blei DM, Lafferty JD (2006) Dynamic topic models. Cohen D, Moore A, eds. *Proc. 23rd Internat. Conf. Machine Learn.* (ACM, New York), 113–120.
- Blei DM, Ng A, Jordan M (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.
- Boardman JD, Saint Onge JM, Rogers RG, Denney JT (2005) Race differentials in obesity: The impact of place. *J. Health Soc. Behav.* 46(3):229–243.
- Bollinger B, Leslie P, Sorensen A (2011) Calorie posting in chain restaurants. *Amer. Econom. J.: Econom. Policy* 3(1):91–128.
- Büschen J, Allenby GM (2016) Sentence-based text analysis for customer reviews. *Marketing Sci.* 35(6):953–975.
- Cao J, Xia T, Li J, Zhang Y, Tang S (2009) A density-based method for adaptive LDA model selection. *Neurocomputing* 72(7–9):1775–1781.
- Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: How humans interpret topic models. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Adv. Neural Inform. Processing Systems*, Vol. 22, 1–9.
- Decker R, Trusov M (2010) Estimating aggregate consumer preferences from online product reviews. *Internat. J. Res. Marketing* 27(4):293–307.
- Deveaud R, SanJuan E, Bellot P (2012) LIA at TREC 2012 Web track: Unsupervised search concepts identification from general sources of information. *Proc. 21th Text Retrieval Conf. (TREC 2012)*, Gaithersburg, MD.
- Dickey JM (1983) Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *J. Amer. Statist. Assoc.* 78(383):628–637.
- Downs JS, Wisdom J, Wansink B, Loewenstein G (2013) Supplementing menu labeling with calorie recommendations to test for facilitation effects. *Amer. J. Public Health* 103(9):1604–1609.
- Eliashberg J, Hui SK, Zhang J (2007) From story line to box office: A new approach for green-lighting movie scripts. *Management Sci.* 53(6):881–893.
- Finkelstein EA, Trogdon JG, Cohen JW, Dietz W (2009) Annual medical spending attributable to obesity: Payer and service specific estimates. *Health Affairs* 28(5):822–831.
- Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.
- Godes D, Silva JC (2013) Sequential and temporal dynamics of online opinion. *Marketing Sci.* 31(3):448–473.
- Griffin A, Hauser JR (1993) The voice of the customer. *Marketing Sci.* 12(1):1–27.
- Griffiths T, Steyvers M (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. USA* 101(Suppl 1):5228–5235.
- Jagaramudi J, Daumé H III, Udupa R (2012) Incorporating lexical priors into topic models. *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 204–213.
- Jalonick MC (2013) FDA head says menu labeling “thorny” issue. *Associated Press* (March 12). <http://www.foxnews.com/health/2013/03/12/fda-head-says-menu-labeling-thorny-issue.html>.
- Kullback S, Leibler R (1951) On information and sufficiency. *Ann. Math. Statist.* 22(1):79–86.
- Kuo T, Jarosz CJ, Simon P, Fielding JE (2009) Menu labeling as a potential strategy for combating the obesity epidemic: A health impact assessment. *Amer. J. Public Health* 99(9):1680–1686.
- Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.
- Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans. Inform. Theory* 37(1):145–151.
- Lu B, Ott M, Cardie C, Tsou B (2011) Multi-aspect sentiment analysis with topic models. Spiliopoulou M, Wang H, Cook D, Pei J, Wang W, Zaiane O, Wu X, eds. *Data Mining Workshops (ICDMW), 2011 IEEE 11th Internat. Conf.* (IEEE Computer Society, Washington, DC), 81–88.
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2011) Big data: The next frontier for innovation, competition, and productivity. Report, McKinsey Global Institute, Chicago.

- Mimno D, McCallum A (2008) Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *Proc. Twenty-Fourth Conf. Annual Conf. Uncertainty Artificial Intelligence (AUAI Press, Arlington, VA)*.
- Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. *Proc. Conf. Empirical Methods Natural Language Processing*, 262–272.
- Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.
- Roe B, Levy AS, Derby BM (1999) The impact of health claims on consumer search and product evaluation outcomes: Results from FDA experimental data. *J. Public Policy Marketing* 18(1):89–105.
- Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. *Proc. 20th Conf. Uncertainty Artificial Intelligence (AUAI Press, Arlington, VA)*, 487–494.
- Steyvers M, Griffiths T (2007) Probabilistic topic models. *Handbook Latent Semantic Anal.* 427(7):424–440.
- Tirunillai S, Tellis G (2014) Mining marketing meaning from chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *J. Marketing Res.* 51(4):463–479.
- Wierenga B (2006) Motion pictures: Consumers, channels, and intuition. *Marketing Sci.* 25(6):674–677.
- Yaniv I, Hogarth RM (1993) Judgmental versus statistical prediction: Information asymmetry and combination rules. *Psych. Sci.* 4(1):58–62.