NBER WORKING PAPER SERIES

OPTIMAL AGGREGATION OF CONSUMER RATINGS: AN APPLICATION TO YELP.COM

Weijia Dai
Ginger Z. Jin
Jungmin Lee
Michael Luca

Working Paper 18567
http://www.nber.org/papers/w18567

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2012

We are grateful to John Rust, Matthew Genzkow, Connan Snider, Phillip Leslie, Yossi Spiegel, and participants at the 2012 UCLA Alumni Conference and the Fifth Workshop on the Economics of Advertising and Marketing for constructive comments. Financial support from the University of Maryland and the Sogang University Research Grant of 2011 (#201110038.01) is graciously acknowledged. All errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Optimal Aggregation of Consumer Ratings: An Application to Yelp.com
Weijia Dai, Ginger Z. Jin, Jungmin Lee, and Michael Luca
NBER Working Paper No. 18567
November 2012
JEL No. D8,L15,L86

ABSTRACT

Consumer review websites such as Yelp.com leverage the wisdom of the crowd, with each product being reviewed many times (some with more than 1000 reviews). Because of this, the way in which information is aggregated is a central decision faced by consumer review websites. Given a set of reviews, what is the optimal way to construct an average rating? We offer a structural approach to answering this question, allowing for (1) reviewers to vary in stringency (some reviewers tend to leave worse reviews on average) and accuracy (some reviewers are more erratic than others), (2) reviewers to be influenced by existing reviews, and (3) product quality to change over time. We apply this approach to reviews from Yelp.com to derive optimal ratings for each restaurant (in contrast with the arithmetic average displayed by Yelp). Because we have the history of reviews for each restaurant and many reviews left by each reviewer, we are able to identify these factors using variation in ratings within and across reviewers and restaurants. Using our estimated parameters, we construct optimal ratings for all restaurants on Yelp, and compare them to the arithmetic averages displayed by Yelp. As of the end of our sample, a conservative finding is that roughly 25-27% of restaurants are more than 0.15 stars away from the optimal rating, and 8-10% of restaurants are more than 0.25 stars away from the optimal rating. This suggests that large gains could be made by implementing optimal ratings. Much of the gains come from our method responding more quickly to changes in a restaurant's quality. Our algorithm can be flexibly applied to many different review settings.

Weijia Dai
University of Maryland
Department of Economics
3114 Tydings Hall
College Park, MD 20742
dai@umd.edu

Ginger Z. Jin
University of Maryland
Department of Economics
3115F Tydings Hall
College Park, MD 20742-7211
and NBER
jin@econ.umd.edu

Jungmin Lee
Sogang University
Seoul, Korea
junglee@sogang.ac.kr

Michael Luca
Soldiers Field Road
Boston, MA 02163
mluca@hbs.edu

1 Introduction

The digital age has transformed the way that consumers learn about product quality. Websites ranging from Yelp and TripAdvisor to eBay and Amazon use crowdsourcing to generate product ratings and reviews. This has dramatically increased the amount of information consumers have when making a decision. The value of this information increases in the number of reviews being left. However, the more reviews that are left, the more time-consuming and difficult it becomes for a consumer to process the underlying information. This calls for the platform to generate an easy-to-understand metric that summarizes existing reviews on a specific subject. In this paper, we develop a method to analyze and aggregate reviews, and apply this method to restaurant reviews from Yelp.com.

How should consumer review websites present information to readers? In principle, one could simply present all of the underlying reviews and allow consumers to decide for themselves how to aggregate information. In fact, there are some websites that do this, and hence avoid the need to aggregate information. Yet a growing literature has demonstrated that the impact of information depends not only on the informational content but also on the salience and simplicity of the information (Brown et al 2010, Luca and Smith *forthcoming*, Pope 2009). In the case of Yelp, for instance, consumers respond directly to the average rating even though this is coarser than the underlying information (Luca 2011). Because of consumer inattention, the method chosen to aggregate information is of first-order importance.

Currently, many review websites (including Yelp) present an arithmetic mean of all reviews written for a given product. Implicitly, this method of aggregation treats each review as an equally informative noisy signal of quality. In other words, arithmetic average is only optimal under very restrictive conditions - such as when reviews are unbiased, independent, and identically distributed signals of true quality.

The goal of this paper is to move toward optimal aggregation of consumer reviews. We consider an aggregate rating to be optimal if two conditions are met. First, observable preferences and biases of different types of consumers must be separated from a consumer's vertical signal of quality. Second, reviews must be weighted to account for informational content, with more weight endogenously assigned to reviews containing more information. This includes both the fact that some reviewers may be more accurate than others and the fact that product quality may change over time. The product of this paper is a single aggregated measure of vertical quality.

To derive an optimal aggregation algorithm, we develop a structural framework that allows reviewers to vary in accuracy (some reviewers are more erratic than others), stringency (some reviewers leave systematically lower ratings), and reputational concerns (some reviewers care more about reputation on Yelp than others, and hence prefer not to deviate from prior reviews). Our framework also accounts for the fact that a restaurant's quality can change over time, which implies that concurrent restaurant quality is better reflected in recent reviews than in early reviews.

Because we have the entire history of reviews for each restaurant and many reviews left by each reviewer, we are able to identify these factors using variation in ratings within and across reviewers and restaurants. For example, stringency of a reviewer can be identified using restaurant and reviewer-type fixed effects. Similarly, accuracy of a reviewer can be identified by variation in how far different reviewers are (in expectation) from the long-run average rating of the restaurants they review. To identify changes in restaurant quality, we impose the assumption that the evolution of restaurant quality follows a martingale process by calendar time, and estimate the underlying parameters. Our model also allows restaurant ratings to follow a common time trend since the first Yelp review of a restaurant, which could capture a linear trend of reviewer stringency relative to the first review of the same restaurant, or a linear trend of true quality in addition to the martingale evolution of quality.

Using our estimated parameters, we then construct optimal average ratings for each restaurant on Yelp, and compare them to the simple arithmetic mean by Yelp. The results depend on how we interpret a significant downward trend of ratings within a restaurant. If this "chilling" effect is interpreted as reviewer bias only (relative to the first review), we find that, by the end of the sample, more than half of restaurants have their Yelp-style simple average ratings differ from the optimal by more than 0.15 stars, and more than one-quarter of restaurants have Yelp-style average ratings differ from the optimal by more than 0.25 stars. If the above chilling effect is interpreted as changes in true quality, the absolute difference between simple and optimal average ratings is still more than 0.15 stars for 25-27% of restaurants, and more than 0.25 stars for 8-10% of restaurants by the end of the data sample.

Most of the optimal-vs-simple-average difference is driven by evolution of restaurant quality. This is because the simple average weights a restaurant's first review the same as it weights the thousandth review. In contrast, our algorithm reduces the weight assigned to early reviews and hence more quickly adapts to changes in quality. Reviewer reputation, on the other hand, has little impact on the optimal average in the Yelp setting, even though reputational concerns may be an important part of the decision to become a Yelp reviewer to begin with. For example, "elite" status is a designation given by Yelp to prolific reviewers, who leave what Yelp deems to be higher quality reviews.[1] Our model shows that elite and non-elite reviewers have different signal precision and reputational concerns. Elite reviewers provide ratings with higher precision; these ratings are also closer to a restaurant's long-run average rating. Moreover, estimates suggest that elite reviewers are more likely to incorporate previous reviews of the same restaurant, which can be explained by elite reviewers having a greater reputation concern on Yelp. These reputation concerns (i.e. popularity concerns), as well as the high signal precision of elite reviewers, suggest that the aggregate rating should give more weight to elite reviews. However, at least in our data, reviewer heterogeneity in signal precision and reputational concern explain much less of the optimal-vs-simple-average

[1] Elite status initiates from a nomination from the Yelp reviewers (can be the reviewer herself), and the final designation decision is made by Yelp based on the reviewer's Yelp activeness.

difference than the martingale evolution of restaurant quality and the overall time trend in consumer reviews.

Although our algorithm is derived from Yelp reviews, it could be applied to virtually any website that relies on consumer ratings to convey information of product or service quality. This contributes to the small, but growing literature on information aggregation as well as the literature on consumer reviews. Li and Hitt (2008) find that book reviews on Amazon tend to trend downward overtime, which they attribute to selection, with early purchasers tending to be those who have the strongest preferences for the book, providing further motivation for the need for optimal aggregation. Glazer et al. (2008) have theoretically considered optimal ratings in the context of health plan report cards. Another approach to aggregate the information is via demand estimation. Based on hotel reservation data from Travelocity.com, which include consumer-generated reviews from Travelocity.com and TripAdvisor.com, Ghose, Ipeirotis and Li (forthcoming) estimate consumer demand for various product attributes and then rank products according to estimated "expected utility gain." In comparison, we attempt to aggregate consumer reviews without complementary data on how consumers use such reviews when they choose a product. This situation is faced by many opinion generation websites that offer consumer ratings but do not sell the rated products. Readers interested in consumer usage of Yelp reviews can refer to Luca (2011) who combines the same Yelp data as in this paper with restaurant revenue data from Seattle[2]. Finally, our model of reputation concerns is also related to the vast literature on information cascade and the growing literature on observation learning (e.g. Banerjee 1993).

The rest of the paper is organized as follows. Section 2 presents the model and describes how we estimate and identify key parameters in the model. Section 3 describes the data and presents reduced-form results. Section 4 presents structural estimates. Section 5 presents counterfactual simulations, and compares optimal average ratings to arithmetic average ratings. Section 6 concludes.

2 Model and Estimation

Consider a consumer review website that has already gathered many consumer reviews on many products over a period of time. Our goal is to optimally summarize existing reviews into a single metric of concurrent quality for each product. Simple average assumes that every consumer review follows an i.i.d. distribution around a stable level of true product quality. This assumption can be violated if true quality evolves over time, if reviews are sequentially

[2] More generally, there is strong evidence that consumer reviews are an important source of information in a variety of settings. Chevalier and Mayzlin (2006) find predictive power of consumer rating on book sales. Both Godes and Mayzlin (2004) and Duan, Gu, and Whinston (2008) find the spread of word-of-mouth affect sales by bringing the consumer awareness of consumers, the former measure the spread by the "the dispersion of conversations across communities" and the latter by the volume of reviews. Duan et al. (2008) argues that after the endogenous correlation among ratings, online user reviews has no significant impact on movies' box office revenues.

correlated, and if reviewers differ in stringency and accuracy. This section presents a structural model that captures all these elements in a coherent framework. Our goal in the model is to incorporate economically important parameters while maintaining econometric tractability.

2.1 Basic Setup

Consider reviewer i who writes a review for restaurant r at calendar time t_n.[3] As the n^{th} reviewer of r, she observes her own signal s_{rt_n} as well as all the $n-1$ reviews of r before her $\{x_{r1}, x_{r2}, ..., x_{rn-1}\}$. s_{rt_n} is assumed to be an unbiased but noisy signal of the true quality μ_{rt_n} such that $s_{rt_n} = \mu_{rt_n} + \epsilon_{rn}$ where $\epsilon_{rn} \sim N(0, \sigma_i^2)$. We assume the noise has the same variance when reviewer i visits different restaurants. This way, we can denote the precision of reviewer i's information as $v_i = \frac{1}{\sigma_i^2}$. Because r and n jointly identify a unique reviewer, we use i interchangeably with the combination of r and n.

We consider two incentives for reviewer i to report. The first incentive is to speak out her own emotion and obtain personal satisfaction from it. If satisfaction comes from expressing the true feeling, this incentive motivates her to report her own signal. If i obtains psychological gains from reporting the signal with certain deviation, which we denote as stringency $\theta_{rn} \neq 0$, then she will be motivated to report her signal plus her stringency measure.[4] The second incentive of submitting a Yelp review is to make a best guess of restaurant quality so that she can earn popularity or reputation in the future. For simplicity, we assume popularity is defined by the extent to which Yelp readers agree with her review. If there are many more readers than reviewers and each reader expresses her true feeling when she collates her experience with the review, popularity of i would decrease with the distance between her review x_{rt_n} and the restaurant's actual quality μ_{rt_n}.[5] Combining the above two incentives, reviewer i, as the n^{th} reviewer of restaurant r, chooses her review x_{rt_n}in order to minimize the following objective function:

$$F_{rn} = (1-\rho_i)(x_{rt_n} - s_{rt_n} - \theta_{rn})^2 + \rho_i[x_{rt_n} - E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{n-1}}, s_{rt_n})]^2$$

where $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{\{n-1\}}}, s_{rt_n})$ is the posterior belief of true quality μ_{rt_n} and ρ_i is the weight that i puts on her popularity on Yelp. The optimal review to minimize F_{rn} is:

$$\begin{aligned} x_{rt_n} &= (1-\rho_i)(\theta_{rn} + s_{rt_n}) + \rho_i E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ..x_{rt_{n-1}}, s_{rt_n}) \\ &= \lambda_{rn} + (1-\rho_i)s_{rt_n} + \rho_i E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ..x_{rt_{n-1}}, s_{rt_n}) \end{aligned}$$

where $\lambda_{rn} = (1-\rho_i)\theta_{rn}$ represents the stringency or bias of reviewer i for restaurant r.

[3] We assume that a reviewer submits one review for a restaurant. Therefore, the order of the review indicates the reviewer's identity. On Yelp.com, reviewers are only allowed to display one review per restaurant.

[4] Some reviewers are by nature generous and obtain psychological gains from submitting reviews that are more favorable than what they actually feel. In this case, $\theta_{rn} > 0$ represents leniency.

[5] In this sense the reviewer is a truth seeker. Another interpretation of the popularity incentive is the desire to contribute to a public good by submitting the best estimate of restaurant quality.

Note that popularity weight ρ_i can be negative if reviewer i enjoys being far away from the best estimate of restaurant quality. This incentive to deviate is different from stringency θ_{rn} because θ_{rn} is independent of any reviewer before i but the incentive to deviate due to negative ρ_i depends on the reviews available before t_n.

Popularity weight ρ_i allows sequential correlations among reviews of the same restaurant, even after taking out restaurant fixed effects. This is because a reviewer that cares about popularity on Yelp will extract useful information from past reviews and incorporate them into her own estimate of restaurant quality. This assumption is consistent with the findings of Chen et al. (2010), who use a field experiment on MovieLens to show that some users dramatically change their movie ratings when they are presented with some social information (e.g. the median user's total number of movie ratings). Of course, this is not the only explanation for correlations across reviews; one can argue that the first review attracts certain types of patrons who have similar (or opposite) tastes to the first reviewer and this selection process generates correlations across reviews. Without external information to pin down such selection, the resulting sequence of reviews can be observationally equivalent to what arises from popularity concerns. In this sense, we would like readers to take popularity weight ρ_i as an indicator of how a review correlates with past reviews. As long as later reviews capture information from past reviews, optimal aggregation needs to weigh early and late reviews differently.

2.2 Restaurant Quality Change

If restaurant quality is constant over time and every reviewer is unbiased, then aggregation of consumer reviews is straightforward: even a simple average of reviews will generate an unbiased indicator of true quality, and optimal aggregation can only improve efficiency by giving more weight to more precise reviewers or reviewers with greater popularity concerns.

However, the assumption of constant restaurant quality is unrealistic. The restaurant industry is known for high labor turnover as well as high entry and exit rates. A new chef or a new manager could change a restaurant significantly; even a sloppy waiter could generate massive consumer complaints in a short time. In reality, consumer reviews and restaurant quality may move together because reviews reflect restaurant quality, or restaurant owners may adjust a restaurant's menu, management style, or labor force in response to consumer reviews. Without any direct data on restaurant quality, it is difficult to separate the two. In light of the difficulty, we impose an independent structure on restaurant quality change and shy away from an endogenous generation of restaurant quality in response to consumer reviews. This way, we focus on measures of restaurant quality rather than reasons underlying quality change.

In particular, we assume quality evolution follows a martingale process:

$$\mu_{rt} = \mu_{r(t-1)} + \xi_t$$

where t denotes the units of calendar time since restaurant r has first been reviewed and the t-specific evolution ξ_t conforms to $\xi_t \sim i.i,d\ N(0, \sigma_\xi^2)$. This martingale process introduces a positive correlation of restaurant quality over time,

$$Cov(\mu_{rt}, \mu_{rt'}) = E(\mu_{r0} + \sum_{\tau=1}^{t} \xi_\tau - E(\mu_{rt}))(\mu_{r0} + \sum_{\tau=1}^{t'} \xi_\tau - E(\mu_{rt'}))$$
$$= E(\sum_{\tau=1}^{t} \xi_\tau \sum_{\tau=1}^{t'} \xi_\tau) = \sum_{\tau=1}^{t} E(\xi_\tau^2)\ if\ t < t',$$

which increases with the timing of the earlier date (t) but is independent of the time between t and t'.

Recall that x_{rt_n} is the n^{th} review written at time t_n since r was first reviewed. We can express the n^{th} reviewer's signal as:

$$s_{rt_n} = \mu_{rt_n} + \epsilon_{rn}$$
$$where \quad \mu_{rt_n} = \mu_{rt_{n-1}} + \xi_{t_{n-1}+1} + \xi_{t_{n-1}+2} + ... + \xi_{t_n}.$$

Signal noise ϵ_{rn} is assumed to be $i.i.d.$ with $Var(s_{rt_n}|\mu_{rt_n}) = \sigma_i^2$ where i is the identity of the n^{th} reviewer. The variance of restaurant quality at t_n conditional on quality at t_{n-1} is,

$$Var(\mu_{rt_n}|\mu_{rt_{n-1}}) = Var(\xi_{t_{n-1}+1} + \xi_{t_{n-1}+2} + ... + \xi_{t_n}) = (t_n - t_{n-1})\sigma_\xi^2 = \Delta t_n \sigma_\xi^2.$$

Note that the martingale assumption entails two features in the stochastic process: first, conditional on $\mu_{rt_{n-1}}$, μ_{rt_n} is independent of the past signals $\{s_{rt_1}, ..., s_{rt_{n-1}}\}$; second, conditional on μ_{rt_n}, s_{rt_n} is independent of the past signals $\{s_{rt_1}, ..., s_{rt_{n-1}}\}$. As shown later, these two features greatly facilitate reviewer n's Bayesian estimate of restaurant quality. This is also why we choose martingale over other statistical processes (such as AR(1)).

2.3 Reviewer Heterogeneity and Reviewer-Restaurant Match

In addition to random changes of restaurant quality and random noise in reviewer signal, reviewers may differ in stringency, popularity concern, and signal precision. Optimal information aggregation - in our definition - would correct for these differences. For the purposes of this paper, we are estimating match based on characteristics that we observer - which is the review history for each reviewer and their Yelp-granted elite status.

Yelp assigns elite status to a subset of reviewers who have been nominated - either by themselves or by other Yelp users - due to a perceived high quality of reviews. We take Yelp "elite" as a signal of a reviewer's type, and hence take elite status as given. We then allow elite reviewers to have $\{\rho_e, \sigma_e^2\}$ while all non-elite reviewers have $\{\rho_{ne}, \sigma_{ne}^2\}$. If elite reviewers are able to obtain more precise signals of restaurant quality and care more about their reputation on Yelp, we expect $\rho_e > \rho_{ne}$ and $\sigma_e^2 < \sigma_{ne}^2$. Elite and non-elite reviewers could also differ in stringency λ_e and λ_{ne}. However, we do not know true restaurant quality and can at best only identify the stringency difference between elite and non-elite reviewers.

In theory, reviewer stringency, popularity concern and signal precision can all vary over time. From observed reviewer history, we define several reviewer attributes at the time of a particular review. One is the number of reviews that reviewer i has submitted for Seattle restaurants before writing a new review for restaurant r at time t. This reflects reviewer experience with Seattle restaurants. We denote it as $NumRev_{it}$. The second is review frequency of i at t, defined as the number of reviews i has submitted up to t divided by the number of calendar days from her first review to t. Review frequency allows us to capture the possibility that a reviewer who has submitted two reviews 10 months apart is fundamentally different from a reviewer who has submitted two reviews within two days, even though both reviewers have the same number of reviews on Yelp. We denote review frequency of i at t as $FreqRev_{it}$.

The third and fourth reviewer attributes attempt to capture reviewer-restaurant match. In reality, reviewers may have their own preference for cuisine type and sort themselves into different restaurants at different times. Although we do not have enough information to model the sorting explicitly, we can describe reviewer-restaurant match by characteristics of the restaurants a reviewer has written reviews for in the past. In particular, we collect 15 cuisine type indicators describing whether a restaurant is traditional American, new American, European, Mediterranean, Latin American, Asian, Japanese, seafood, fast food, lounge, bar, bakery/coffee, vegetarian, or others. These categories are defined by Yelp and not mutually exclusive. We also use Yelp's definition of price categories (1,2,3,4) and code a missing price category as category 0. With these restaurant characteristics in hand, we use factor analysis to decompose them into eight orthogonal factors $F_r = [f_{r,1}, ..., f_{r,8}]$. By construction, the sample mean of each factor is normalized to 0 and sample variance normalized to 1. We then collapse a reviewer history into two metrics: the first metric, C_{it}, measures the average restaurant that this reviewer has written reviews for before she writes her m^{th} review at time t, the second metric, $TasteVar_{it}$, measures the variety of restaurants that she has written reviews for before her m^{th} review at time t. In particular, they are defined as:

$$C_{it} = \frac{1}{m-1} \sum_{l=1}^{m-1} F_{il},$$

$$TasteVar_{it} = \sqrt{\sum_{q=1}^{8} \frac{1}{m-2} \sum_{l=1}^{m-1} (f_{il,q} - \overline{f}_{il,q})^2}$$

where $m-1$ is the number of Seattle restaurants reviewer i has written reviews for before t, F_{il} denotes the vector of factors of the l^{th} restaurant that i visited, and $\overline{f}_{il,q} = \frac{1}{m-1} \sum_{l=1}^{m-1} f_{il,q}$ is the mean in factor q among the $m-1$ restaurants that i visited. If reviewer i has not reviewed any restaurant yet, we set her taste equal to the mean characteristics of restaurants ($C_{it} = 0$). When reviewer i writes a review for restaurant r, we have a pair of $\{C_{it}, F_r\}$ to describe the reviewer taste and restaurant characteristics. Assuming that reviewer i reviews restaurant r at time t, we define the reviewer-restaurant matching distance $MatchD_{rit}$ as

$$MatchD_{rit} = (C_{it} - F_r)'(C_{it} - F_r).$$

The smaller the matching distance ($MatchD_{rit}$) , the better the match is between the restaurant and the reviewer's review history.

To summarize, we have five reviewer attributes: elite status ($Elite_i$), number of reviews ($NumRev_{it}$), frequency of reviews ($FreqRev_{it}$), matching distance between reviewer and restaurant ($MatchD_{rit}$), and taste for variety ($TasteVar_{it}$). By construction, all but $Elite_i$ vary within a reviewer over time, and only $MatchD_{rit}$ depends on the restaurant that the reviewer is about to review at time t.

Readers should take $MatchD_{rit}$ and $TasteVar_{it}$ as controls for observable sorting between restaurants and reviewers. In reality, who reviews which restaurant at what time can be driven by past reviews of every restaurant and thus endogenous. Some unobservable tastes of reviewers will lead to specific values of $MatchD_{rit}$ and $TasteVar_{it}$; hence controlling for $MatchD_{rit}$ and $TasteVar_{it}$ indirectly controls for these unobservable tastes. Other unobservable attributes of reviewers may not have any influence on $MatchD_{rit}$ and $TasteVar_{it}$, but they affect how reviewers read past reviews and then visit the restaurant and write their own reviews. This will generate correlations along the order of reviews, and such correlations are captured in popularity weight ρ_i.

2.4 Time Trend

In addition to all the above, we also record the number of calendar days since restaurant r received its first review on Yelp until a reviewer is about to enter the review for r at time t. This variable, denoted as Age_{rt}, attempts to capture any linear trend in consumer reviews that is missed by the above-mentioned reviewer or restaurant variables. By definition, this trend – which turns out to be negative over time – is subject to multiple interpretations. It is possible that true restaurant quality declines over time for every restaurant. Note that this decline is in addition to the martingale evolution of restaurant quality because the martingale deviation is assumed to have mean zero. It is also possible that later reviewers are always harsher than early reviewers. Either interpretation can be a result of a "chilling" effect as described in Li and Hitt (2008), who also lay out these competing hypotheses. We are unable to distinguish these underlying stories, although the interpretation does affect how we calculate optimal estimate of true quality. We will come back to this point when we present the optimal average ratings in Section 5.

To summarize, we assume:

$$\begin{aligned}
\rho_i &= Elite_i \cdot \rho_e + (1 - Elite_i) \cdot \rho_{ne} \\
\sigma_i^2 &= Elite_i \cdot \sigma_e^2 + (1 - Elite_i) \cdot \sigma_{ne}^2 \\
\lambda_{ri} &= Age_{rt} \cdot \alpha_{age} + NumRev_{it} \cdot \alpha_{numrev} + FreqRev_{it} \cdot \alpha_{freqrev} \\
&\quad + MatchD_{rit} \cdot \alpha_{matchd} + TasteVar_{it} \cdot \alpha_{tastevar} \\
&\quad + Elite_i \cdot [\lambda_{(e-ne)0} + Age_{rt} \cdot \beta_{day} + NumRev_{it} \cdot \beta_{numrev} + FreqRev_{it} \cdot \beta_{freqrev} \\
&\quad + MatchD_{rit} \cdot \beta_{matchd} + TasteVar_{it} \cdot \beta_{tastevar}]
\end{aligned}$$

where $\{\rho_e, \rho_{ne}\}$ capture the popularity concern of elite and non-elite reviewers, $\{\sigma_e^2, \sigma_{ne}^2\}$ capture the signal precision of elite and non-elite reviewers, $\{\alpha_{age}\}$ captures the catch-all trend in quality or stringency change, $\{\alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}\}$ capture how restaurant and reviewer attributes change the stringency of non-elite reviewers, and $\{\lambda_{(e-ne)0}, \beta_{day}, \beta_{numrev}, \beta_{freqrev}, \beta_{matchd}, \beta_{tastevar}\}$capture how restaurant and reviewer attribute change the stringency difference between elite and non-elite reviewers. We have tried to allow ρ_i and σ_i^2 to vary by restaurant and reviewer attributes other than elite status, but none of them turns out to be significant from zero, so we ignore them here for the simplicity of illustration.

2.5 Data Generation Process

The above model includes random change in restaurant quality, random noise in reviewer signal, reviewer heterogeneity in stringency, popularity concern, and signal precision, and a linear time trend, as well as the quality of the match between the reviewer and the restaurant. Overall, one can consider the data generation process as the following three steps:

1. Restaurant r starts with an initial quality μ_{r0} when it is first reviewed on Yelp. Denote this time as time 0. Since time 0, restaurant quality μ_r evolves in a martingale process by calendar time, where an i.i.d. quality noise $\xi_t \sim N(0, \sigma_\xi^2)$ is added on to restaurant quality at t so that $\mu_{rt} = \mu_{r(t-1)} + \xi_t$.

2. A reviewer arrives at restaurant r at time t_n as r's n^{th} reviewer. She observes the attributes and ratings of all the previous $n-1$ reviewers of r. She also obtains a signal $s_{rt_n} = \mu_{rt_n} + \epsilon_{rn}$ of the concurrent restaurant quality where the signal noise $\epsilon_{rn} \sim N(0, \sigma_\epsilon^2)$.

3. The reviewer chooses an optimal review that minimizes her loss of deviating from her own experience and her best estimate of concurrent restaurant quality. The optimal review takes the form

$$x_{rt_n} = \lambda_{rn} + \rho_n E(\mu_{rt_n} | x_{rt_1}, x_{rt_2}, .., x_{rt_2}, ..., s_{rt_n}) + (1 - \rho_n) s_{rt_n}$$

where $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_2}, ..., s_{rt_n})$ is the best guess of the restaurant quality at t_n by Bayesian updating.

4. The reviewer is assumed to know the attributes of all past reviewers so that she can debias the stringency of past reviewers. The reviewer also knows that there is a linear trend in reviewer stringency which changes λ by $d\lambda$ per unit of calender time. If there is a linear trend in restaurant quality, it is just a linear term added to $E(\mu_{rt_n}|x_{rt_1}, .., x_{rt_n}, s_{rt_n})$ and not distinguishable from the linear trend in λ in the above expression for x_{rt_n}.

In the Bayesian estimate of $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_2}, ..., s_{rt_n})$, we assume the n^{th} reviewer of r is fully rational and has perfect information about the other reviewers' observable attributes, which according to our model determines the other reviewers' stringency (λ), popularity preference (ρ), and signal noise (σ_ϵ). With this knowledge, the n^{th} reviewer of r can back out each reviewer's signal before her thus the Bayesian estimate of $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_2}, ..., s_{rt_n})$ can be rewritten as $E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n})$. Typical Bayesian inference implies that reviewer's posterior about restaurant quality is a weighted average of previous signals and her own signal, while weight increases with signal precision. What is complicated is that restaurant quality evolves by martingale and therefore current restaurant quality is better reflected in recent reviews. Accordingly, the Bayesian estimate of $E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n})$ should give more weight to more recent reviews even if all reviewers have the same stringency, popularity preference and signal precision. The analytical derivation of $E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n})$ is presented in Appendix A.

2.6 Maximum Likelihood Estimation

According to the derivation of $E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n})$ in Appendix A, we can write out the probability distribution of all the N_r reviews of restaurant r, namely $L(x_{rt_1}, x_{rt_2}, ..x_{rt_{N_r}})$, and then estimate parameters by maximizing the combined log likelihood of all reviews of all R restaurants $logL = \sum_{r=1}^{R} logL(x_{rt_1}, x_{rt_2}, ..x_{rt_{N_r}})$. The parameters to be estimated are restaurant quality at time 0 ($\{\mu_{r0}\}_{r=1}^{R}$), the standard deviation of the martingale noise of restaurant quality change (σ_ξ), the standard deviation of reviewer signal noise (σ_e, σ_{ne}), reviewer popularity concerns (ρ_e, ρ_{ne}), parameters affecting reviewer stringency ($\alpha_{numrev}, \alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}$, $\lambda_{(e-ne)0}$, $\beta_{day}, \beta_{numrev}, \beta_{freqrev}, \beta_{matchd}, \beta_{tastevar}$), and the parameter for the catch-all time trend (α_{age}).

Note that consistent estimation of all other parameters depends on the consistency of $\{\mu_{r0}\}_{r=1}^{R}$, which requires the number of reviews of each restaurant goes to infinity. But in our data, the number of reviews per restaurant has a mean of 33 and a median of 14. When we use simulated data to test the MLE estimation of observed reviews, we find that poor convergence of $\{\mu_{r0}\}_{r=1}^{R}$ affects the estimation of other key parameters of interest.

To circumvent the problem, we estimate the joint likelihood of $\{x_{r2}-x_{r1}, x_{r3}-x_{r2}, ..., x_{rN_r}-x_{rN_r-1}\}_{r=1}^{R}$ instead. In this way we subtract the initial restaurant qualities $\{\mu_{r0}\}_{r=1}^{R}$ and only need to estimate the other parameters. Because the covariance structure of $\{x_{rt_2}-x_{rt_1}, x_{rt_3}-$

$x_{rt_2}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}}\}$ is complicated, we use the change of variable technique to express the likelihood $f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}})$ by $f(s_{rt_2} - s_{rt_1}, ..., s_{rt_{N_r}} - s_{rt_{N_r-1}})$,

$$f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}}) = |J_{\Delta s \to \Delta x}|^{-1} f(s_{rt_2} - s_{rt_1}, ..., s_{rt_{N_r}} - s_{rt_{N_r-1}}).$$

More specifically, $f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}})$ is calculated in three steps:

- Step 1: To derive $f(s_{rt_2} - s_{rt_1}, ..., s_{rt_{N_r}} - s_{rt_{N_r-1}})$, we note that $s_{rt_n} = \mu_{rt_n} + \epsilon_n$ and thus, for any $m > n,\ n \geq 2$, the variance and covariance structure can be written as:

$$\begin{aligned} & Cov(s_{rt_n} - s_{rt_{n-1}}, s_{rt_m} - s_{rt_{m-1}}) \\ =& Cov(\epsilon_{rn} - \epsilon_{rn-1} + \xi_{t_{n-1}+1} + ... + \xi_{t_n}, \epsilon_{rm} - \epsilon_{rm-1} + \xi_{t_{m-1}+1} + ... + \xi_{t_m}) \\ =& \begin{cases} -\sigma_{rn}^2 & if\ m = n+1 \\ 0 & if\ m > n+1 \end{cases} \\ & Var(s_{rt_n} - s_{rt_{n-1}}) \\ =& \sigma_{rn}^2 + \sigma_{rn-1}^2 + (t_n - t_{n-1})\sigma_\xi^2. \end{aligned}$$

  Denoting the total number of reviewers on restaurant r as N_r , the vector of the first differences of signals as $\Delta s_r = \{s_{rt_n} - s_{rt_{n-1}}\}_{n=2}^{N_r}$, and its covariance variance structure as $\Sigma_{\Delta s_r}$, we have

$$f(\Delta s_r) = (2\pi)^{-\frac{N_r-1}{2}} |\Sigma_{\Delta s_r}|^{-(N_r-1)/2} exp(-\frac{1}{2}\Delta s_r' \Sigma_{\Delta s_r}^{-1} \Delta s_r).$$

- Step 2: We derive the value of $\{s_{rt}, ...s_{rt_{N_r}}\}_{r=1}^R$ from observed ratings $\{x_{rt_1}, ...x_{rt_{N_r}}\}_{r=1}^R$. Given

$$x_{rt_n} = \lambda_{rn} + \rho_n E(\mu_{rt_n} | s_{rt_1}, ...s_{rt_n}) + (1 - \rho_n) s_{rt_n}$$

  and $E(\mu_{rt_n} | s_{rt}, ...s_{rt_n})$ as a function of $\{s_{rt_1}, ...s_{rt_n}\}$ (formula in Appendix A), we can solve $\{s_{rt_1}, ...s_{rt_n}\}$ from $\{x_{rt_1}, ...x_{rt_n}\}$ according to the recursive formula in Appendix B.

- Step 3: We derive $|J_{\Delta s \to \Delta x}|^{-1}$ or $|J_{\Delta x \to \Delta s}|$, where $J_{\Delta x \to \Delta s}$ is such that

$$\begin{bmatrix} s_{rt_2} - s_{rt_1} \\ ... \\ s_{rt_n} - s_{rt_{n-1}} \end{bmatrix} = J_{\Delta x \to \Delta s} \begin{bmatrix} x_{rt_2} - x_{rt_1} \\ ... \\ x_{rt_n} - x_{rt_{n-1}} \end{bmatrix}$$

  the analytical form of $J_{\Delta x \to \Delta s}$ is available given the recursive expression for x_{rt_n} and s_{rt_n}.

2.7 Identification

Since our model includes restaurant fixed effects (denoted as time-0 quality μ_{r0}), all our parameters are identified from within-restaurant variations.

In particular, reviewer popularity weight ρ and signal variance σ^2 are identified by the variance-covariance structure of reviews within a restaurant. To see this point, consider a simple case where restaurant quality is stable (i.e. $\sigma_\xi^2 = 0$). If every one has the same signal variance σ_ϵ^2, for the n^{th} review, we have

$$Var(x_{rn}) = \rho_n(2-\rho_n)\frac{\sigma_\epsilon^2}{n} + (1-\rho_n)^2\sigma_\epsilon^2.$$

As we expect, it degenerates to σ_ϵ^2 if the n^{th} reviewer puts zero weight on popularity ($\rho_n = 0$). When $\rho_n > 0$, $Var(x_{rn})$ declines with n. If the n^{th}reviewer cares about popularity only ($\rho_n = 1$), we have the familiar form of $Var(x_{rn}) = \frac{\sigma_\epsilon^2}{n}$. In other words, ρ determines the degree to which the variance of reviews shrinks over time, while σ_ϵ^2 determines the variance of the first review.

There are overidentifications for ρ and σ_ϵ^2, because they affect not only the variance of reviews but also the covariance between reviews. In the above simple case, the covariance of x_{rm} and x_{rn} for $m < n$ is:

$$Cov(x_{rm}, x_{rn}) = \frac{\rho_n}{\sum_{j=1}^{n} v_j}$$

which declines with n, increases with ρ_n, and does not depend on the distance between m and n. This is because the covariance of reviews is generated from reviewer n's belief of restaurant quality, and reviewer n values the information content of each review equally according to the Bayesian principle.

Nevertheless, popularity concern is not the only force that generates correlation between reviews within a restaurant. The other force is restaurant quality evolution. How do we separate the two? The above description has considered the case with popularity concern but no restaurant quality change ($\sigma_\xi^2 = 0$ and $\rho > 0$). Now let us consider a model with $\sigma_\xi^2 > 0$ and $\rho = 0$, which implies that restaurant quality evolves over time but reviewers do not incorporate information from previous reviews. In this case, the correlation between the n^{th} and the $(n-k)^{th}$ reviews only depends on the common quality evolution *before* the $(n-k)^{th}$ reviewer, not the order distance (k) or time distance ($t_n - t_{n-k}$) between the two reviews. In the third case of $\sigma_\xi^2 > 0$ and $\rho > 0$, the n^{th} reviewer is aware of quality evolution and therefore put more weight on recent reviews and less weight on distant reviews. In particular, one can show that the correlation between the n^{th} and the $(n-k)^{th}$ reviews depends on not only the order of review but also the time distance between the two reviews. In short, the separate identification of the noise in quality evolution (σ_ξ^2) from reviewer popularity concern and signal precision$\{\rho, \sigma_\epsilon^2\}$ comes from the calendar time distance between reviews.

As stated before, we allow both ρ and σ_ϵ^2 to differ between elite and non-elite reviewers. Because we observe who is elite and who is not, $\{\rho_e, \sigma_e^2, \rho_{ne}, \sigma_{ne}^2\}$ are identified by the variance-covariance structure of reviews as well as the arrival order of elite and non-elite reviewers.

The constant bias difference between elite and non-elite reviewers ($\lambda_{(e-ne)0}$) is identified by the mean difference of elite and non-elite reviews on the same restaurant. The other parameters that affect reviewer stringency, namely $\{\alpha_{age}, \alpha_{numrev}, \alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}\}$, $\{\beta_{day}, \beta_{numrev}, \beta_{freqrev}, \beta_{matchd}, \beta_{tastevar}\}$, are identified by how the observed reviews vary by restaurant age, reviewer attributes at time t, reviewer-restaurant match, and their interaction with elite status.

2.8 Optimal Estimate of Restaurant Quality

Following the above model, if we interpret the linear trend of ratings ($Age_{rt} \cdot \alpha_{age}$ in λ_{rn}) as reviewer bias (relative to the first reviewer), the optimal estimate of restaurant quality at time t_n is defined as $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_n})$, which is equivalent to $E(\mu_{rt_n}|s_{rt_1}, s_{rt_2}, .., s_{rt_n})$ and we know how to calculate it according to Appendix A. If we interpret the linear trend of ratings ($Age_{rt} \cdot \alpha_{age}$ in λ_{rn}) as changes in true quality, the optimal estimate of quality at t_n is $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_n}) + Age_{rt_n} \cdot \alpha_{age}$. We will report both in Section 6.

3 Data and Reduced Form Results

Our empirical setting is the consumer review website Yelp.com. Yelp began in 2004, and contains reviews for a variety of services ranging from restaurants to barbers to dentists, among many others, although most Yelp reviews are for restaurants. For a more complete description of Yelp, see Luca (2011). In this paper, we use the complete set of restaurant reviews that Yelp displayed for Seattle, WA at our data download time in February 2010. In total, we observe 134,730 reviews for 4,101 Seattle restaurants in a 64-month period from October 15, 2004 to February 7, 2010[6]. These reviews come from 18,778 unique reviewers, of which 1,788 are elite reviewers and 16,990 are non-elite as of the end of our data period. Elite reviewers are determined via a nomination process, where a reviewer can self-nominate or be nominated by someone else. We do not observe the nomination process, and instead only observed whether someone ultimately becomes elite. For our purposes, we take elite status as fixed. Since Yelp reviewers can leave reviews for restaurants throughout the US but our data cover Seattle only, we do not have the complete Yelp history of each reviewer. Another data limitation is that our data contain star ratings given in each review (one to five), but do not include the text. Each reviewer is only allowed to display one review per restaurant, but Yelp allows reviewers to update their existing reviews. If a review has been updated, the review

[6] Yelp identifies reviews that either violate terms of service or seem to be fake, as determined by an algorithm, and removes these reviews from the main Yelp webpage and ratings. We do not observe these reviews, and do not consider them in our analysis.

date records the time of update and there is no information indicating the date or content of the replaced review. Due to this data limit, we treat updated reviews the same as other reviews. In our data set, 64.53% of reviewers have written at least two reviews and 23.7% have written at least five reviews. which provide us with within-reviewer variation.

Table 1 summarizes the main variables in our analysis. In the first panel of restaurant characteristics, we note that on average each restaurant receives 33 reviews but the distribution is highly skewed to the right, ranging from 1 to 698 with a standard deviation of 50 and median of 14. At the end of our data, an average restaurant receives 0.156 reviews per day. This masks enormous heterogeneity of review frequency; if we calculate review frequency per restaurant at any time of a new review, it varies from 0.001 to as large as 28 reviews per day. The arrival of reviews also varies over the lifetime of a restaurant: on average, the second review of a restaurant comes 155 days later than the first review, while the average lag is 34 days between the 11st and 12nd reviews and 21 days between the 21st and 22nd reviews. This is partly driven by the fact that most restaurants receive only a handful number of reviews far apart, while a small fraction of restaurants receive more reviews that arrive much more frequently.

The second panel of Table 1 summarizes the data by reviewers. Although less than 10% of reviewers are elite, an average elite reviewer writes five times more reviews than a non-elite reviewer (24 versus 5). As a result, elite reviewers account for 32.5% of all reviews. Comparing elite and non-elite reviewers, they are similar in average rating per review (both around 3.7 stars), but elite reviewers have a higher review frequency, a closer match with the restaurants they review, and slightly higher variety of taste. The latter two are partly driven by elite reviewers writing more reviews in our data.

3.1 What Explains the Variation in Yelp ratings?

Although the goal of Yelp is to provide information about a restaurant's quality, there are many other factors that determine a restaurant's Yelp rating, for all of the reasons discussed throughout this paper. To get a feel for how significant these other factors are, Table 2 presents the variance explained by different factors.

A linear regression using reviewer fixed effects shows that reviewer fixed effects alone accounting for 23.3% of the total variation in Yelp ratings. This suggests that individual stringency can have a large effect on the final rating. One way to think about restaurant quality is to use restaurant fixed effects, and its variation alone explains 20.86% of total variation in Yelp ratings.

Incorporating both reviewer and restaurant fixed effects, we can explain almost 36% of total variations. This is less than adding the variations accountable by reviewer or restaurant fixed effects separately, suggesting that there is systematic match between reviewers and restaurants. In fact, we are able to control for some of this matching through our proxies for match quality, which further explains the variation in Yelp ratings.

3.2 Elite Reviewers

The data shows quite clearly that different reviewers behave differently. It is possible to segment these individuals into groups of reviewers. In particular, crowdsourced settings such as Yelp, TripAdvisor, and Wikipedia identify reviewers they expect to be especially influential and give them a special certification. On Yelp, this is the "elite" system. In this section, we investigate the ways in which elite reviewers differ from other reviewers. We are interested in this for two reasons. First, this is an increasingly common way to structure review websites, and therefore is of direct interest. Second, by segmenting reviewers, we can allow the weights assigned to a given review to endogenously adjust for different groups of reviewers.

Table 3 shows that elite and non-elite reviews are, in fact, systematically different. The significantly negative coefficient of the elite dummy suggests that elite reviews deviate less from the long-run average rating of the restaurant, suggesting that elite reviewers have more precise signals ($\sigma_e^2 < \sigma_{ne}^2$) or care more about their popularity on Yelp ($\rho_e > \rho_{ne}$).

The elite versus non-elite difference of rating is also presented in Figure 1. The left (right) graph of Figure 1 shows the kernel density of a rating minus the restaurant's average rating beforehand (afterward), for elite and non-elite reviewers separately. An elite reviewer tends to give a rating closer to the restaurant's average ratings before or after her, one phenomenon to be expected if elite reviewers have either more precise signal or greater popularity concerns.

3.3 Dynamics of the Review Process

This makes five empirical observations related to the dynamics of the reviews, which we use to inform our model. In this section, we describe these effects, and highlight the relationship between the reduced form and structural results.

3.3.1 Ratings are less variable over time

As detailed in Section 2, identification of our model relies on the extent to which the variance of reviews shrinks over time within a restaurant. If reviewers tend to incorporate a restaurant's previous reviews ($\rho > 0$), we should observe reviews to vary less and less over time around the restaurant's fixed effect. To check this intuition, we first obtain residual $\widehat{\epsilon_{ri,yr}}$ after regressing observed rating on reviewer, restaurant and year fixed effects (i.e. $x_{ri} = \mu_r + \alpha_i + \gamma_{year} + \epsilon_{ri,yr}$), and then associate residual square ($\widehat{\epsilon_{ri,yr}}^2$) with the order of a review (N_{ri}) and whether the review is written by an elite reviewer (i.e. $\widehat{\epsilon_{ri,yr}}^2 = \beta_0 + \beta_1 D_{ri,elite} + \beta_2 N_{ri} + \beta_3 N_{ri} \times D_{ri,elite} + \zeta_{ri}$). Table 3 shows the results of the latter regression. The significantly negative coefficient on the order of review in suggests that reviews deviate less and less over time from the average rating of the restaurant. Such variance reduction is consistent with reviewers incorporating previous reviews in their own review ($\rho > 0$).

3.3.2 Ratings trend downward over time

Our data shows that ratings within a restaurant tend to decline over time. Figure 2 plots $\widehat{\epsilon_{ri,yr}}$ by the order of reviews within a restaurant, in the fitted fractional polynomial smooth and the corresponding 95% confidence interval. This is consistent with Li and Hitt (2008), who document a downward trend in a product's Amazon reviews over time. There are multiple factors that could contribute to this downward trend.

First, it could be a selection effect, where a restaurant with a good rating tends to attract new customers who do not like the restaurant as much as the old clientele. This is the mechanism outlined by Li and Hitt (2008). If this were the primary driver of the result in our setting, then we would expect later reviewers to be a worse fit for the restaurant. We do find this, as discussed in section 3.3.4. However, this only explains a small component of the downward trend. Because any estimated matching index clearly does not capture all aspects of the match between a customer and a restaurant, we believe that there are contributing factors as well.

An alternative possibility is that restaurants decline in quality over time.

3.3.3 Ratings are serially correlated

Positive popularity concerns also imply positive serial correlation of ratings within a restaurant. In other words, we should see a stronger correlation between the 3rd and 4th review than between the 2nd and 4th. To check this, we regress the above-obtained rating residual $\widehat{\epsilon_{ri,yr}}$ on its lags within the same restaurant. As shown in Table 4, the residuals show strong, positive correlation over time, while the correlation dampens gradually by the order distance between reviews. This is clear evidence that reviews cannot be treated i.i.d. as the simple-average aggregation assumes.

3.3.4 Restaurants find reviewers with less diverse taste over time

Table 5 presents reduced-form analysis regarding variations of reviewer-restaurant matching distance ($MatchD_{rit}$) and reviewer's taste for variety ($TasteVar_{rit}$). The first two columns of Table 5 show that, within a restaurant, later reviewers tend to have less of a diverse taste. This is consistent with the positive sorting hypothesis. And we did not find significant evidence for "chilling" effect where the matching between between restaurant and reviewers gets worse over time.

3.3.5 Reviewers find better matches over time

The last two columns of Table 5 examine variations within a reviewer, which turn out to be quite different from what we have seen within a restaurant. Within a reviewer, the later visited (and reviewed) restaurants are better matched with the reviewer's taste and the reviewer has more taste for variety when she visits and reviews the later restaurants. This suggests that an average reviewer finds better matches over time, but is also more willing to seek variety. In other words, $MatchD_{rit}$ and $TasteVar_{rit}$ capture at least part of the dynamic sorting between restaurants and reviewers, although we do not model the sorting explicitly.

4 Results from Structural Estimation

The goal of our model is to estimate parameters that can then be used for optimal information aggregation. As described in the model section, the parameters of interest pertain to (1) a reviewer's stringency and accuracy, (2) the extent to which a reviewer takes into account prior reviews, (3) the likelihood that a restaurant has changed quality, and (4) the quality of the match between the reviewer and the restaurant. We allow these parameters to vary between groups of reviewers. As an example of how this would work, we compare the parameters for elite and non-elite reviewers. We choose these subsets of reviewers because elite reviewers are such a central part of the review system, as documented in section 3.

Table 6 presents structural estimation results in four columns. In Column (1), we estimate the model under the assumptions that restaurant quality is fixed and reviewers have the same signal precision, popularity weight, and stringency.The estimated signal precision and popularity weight will ultimately be used to optimally aggregate reviews. Note that popularity weight is statistically different from zero, suggesting that reviewers are taking into account the content of previous reviews. As we will see in the simulation section, this will cause later reviews to receive more weight than early reviews.

In the rest of this section, we relax the assumptions to allow for elite reviewers to have different reviewing behavior, and to allow restaurants to change quality over time.

4.1 Elite Reviewers

In Column (2), we allow signal precision, popularity weight, and stringency to differ by reviewer type. Specifically, we allow for elite reviewers to behave differently than other reviewers. The estimates, as well as a likelihood ratio test between Columns (1) and (2), clearly suggests that elite and non-elite reviewers differ in both signal precision and popularity weight. Elite reviewers put higher weight on past reviews and have better signal precision. That being said, all reviewers put more than 75% weight on their own signals, and the noise in their signal is quite large considering the fact that the standard deviation of ratings in the whole sample is of similar magnitude as the estimated $\sigma_e$and σ_{ne}. In terms of stringency, Column (2) suggests insignificant difference between elite and non-elite reviewers.

4.2 Restaurants with Changing Quality

Column (3) allows restaurant quality to change in a martingale process every quarter. As we expect, adding quality change absorbs part of the correlation across reviews, and has significantly reduced the estimate of ρ, but the magnitude of $\rho_e - \rho_{ne}$ is stable at roughly

11-12%. With quality change, ρ_{ne} is estimated to be significantly negative, suggesting that a non-elite reviewer tends to deviate from the mean perspective of the crowd before him, after we allow positive autocorrelation across reviews due to restaurant quality change. Compared to non-elite, elite reviewers are more positively influenced by the past crowd because their popularity/reputation concerns motivate them to be closer to the other reviewers. Although the quarterly noise in restaurant quality ($\sigma_\xi = 0.1452$) is estimated much smaller than the noise in reviewer signal ($\sigma_e = 0.9293$ and $\sigma_{ne} = 0.9850$), this amounts to substantial noise over the whole data period because a random draw of ξ adds up to restaurant quality *every* quarter. A likelihood ratio test between Column 3 and Column 2 favors the inclusion of restaurant quality change.

In addition to restaurant quality change, Column (4) allows reviewer stringency to vary by:

- the restaurant's tenure (on Yelp) Age_{rt}, the quality of the match between the reviewer and the restaurant $MatchD_{rit}$,
- the reviewer's taste for variety $TasteVar_{ri}$,
- the number of reviews a reviewer has written $NumRev_{it}$,
- the frequency with which a reviewer writes reviews $RevFreq_{it}$, and
- the reviewer's elite status.

The set of coefficients that starts with $\mu + \lambda_{ne}$ describes the stringency of non-elite reviewers (which are not identifiable from the time-0 restaurant quality), while the set of coefficients that starts with $\lambda_e - \lambda_{ne}$ describes the stringency difference between elite and non-elite reviewers. According to these coefficients, reviewers are more stringent over time, indicating that there is a "chilling effect." This "chilling" effect is less for elite reviewers. Moreover, reviewers that have written more reviews on Yelp match better with a restaurant, and have more diverse tastes tend to be more stringent. In comparison, an elite reviewer behaves similar in terms of matching distance and taste for variety, but her stringency does not vary significantly by the number of reviews on Yelp. Again, likelihood ratio tests favor the full model of Column 4 over Columns 1-3, suggesting that it is important to incorporate restaurant quality change, reviewer heterogeneity, and signal noise all at once.

A remaining question is, at what frequency does restaurant quality evolve? Given the lack of hard evidence on this, Table 7 shows the full model estimation results with restaurant quality evolving by month, quarter, and half-year. The main changes occur in the estimates for noise of reviewer signal (σ_e, σ_{ne}), noise of quality evolution (σ_ξ), and reviewers' popularity weight (ρ_e, ρ_{ne}). This is not surprising because they are all identified by the variance-covariance structure of reviews within a restaurant. Nevertheless, we are able to identify quality evolution from reviewer signal and popularity preference because there are enormous variations in how close sequential reviews arrive. Clearly, the more frequent we allow restaurant quality to vary, the smaller σ_ξ is (because it captures quality change in a smaller calendar window), the lower the popularity weight, and and the lower the reviewers' signal noise (because more frequent quality change absorbs more autocorrelations of nearby reviews). However, the difference between elite and non-elite reviewers remains similar across the three columns of Table 7. The likelihood reported at the end of Table 7 suggests that the raw data are better explained by more frequent changes of restaurant quality.

4.3 Comparing to a Model of Limited Attention

One assumption underlying our structural model is reviewer rationality. One may argue that the assumption of full rationality is unrealistic, given consumer preference for simple and easy-to-understand metrics. Anecdotally, we know reviewers tend to pay more attention to detailed information than those who only read reviews. To address the concern more rigorously, we estimate an alternative model in which we assume that reviewers are naive and use the simple average of a restaurant's past rating as the best guess of quality. Recall in the full model that the n^{th}reviewer's optimal review should be

$$x_{rt_n} = (1-\rho_n)(\theta_{rn} + s_{rt_n}) + \rho_n E(\mu_{rt_n} | x_{rt_1}, x_{rt_2}, ..x_{rt_{n-1}}, s_{rt_n})$$

where $E(\mu_{rt_n} | x_{rt_1}, x_{rt_2}, ...x_{rt_{n-1}}, s_{rt_n})$ is the Bayesian posterior belief of true quality μ_{rt_n}. If the reviewer is naive, the optimal review will change to:

$$x_{rt_n} = (1-\rho_n) \times (\theta_{rn} + s_{rt_n}) + \rho_n \times (\frac{1}{n-1}\sum_{i=1}^{n-1} x_{rti})$$

where a simple average of past reviews $\frac{1}{n-1}\sum_{i=1}^{n-1} x_{rti}$ replaces the Bayesian posterior estimate of quality $E(\mu_{rt_n} | x_{rt_1}, x_{rt_2}, ..x_{rt_{\{n-1\}}}, s_{rt_n})$.

In an unreported table, we compare the MLE result and log the likelihood of the Bayesian and naive models, while allowing restaurant quality to update by quarter or half year.[7] According to the Akaike information criterion (Akaike 1974), if we assume quality updates by half year, the Bayesian model is 49,020.8 times as probable as the naive model to minimize the information loss.[8] Similarly, if we assume quality updates by quarter, we find the Bayesian model is 2.41×10^8 times as probable as the naive model. This suggests that the Bayesian model is more suitable for our data.

[7] In all specifications, we assume that the reviewer stringency term (λ_{rt}) only depends on $MatchD_{rit}$, $TasteVar_{it}$, and Age_{rt}. Later on, we will redo this by adding $NumRev$ and $FreqRev$ in λ_{rt} but they are unlikely to change the results.

[8] Specifically, we have $exp(AIC_{Bayesian} - AIC_{Naive}) = exp(logL_{Bayesian} - logL_{Naive}) = 49,020.8$.

5 Counterfactual Simulations

This section presents three sets of counterfactual simulations. The first set highlights the role of each modeling element in the optimal aggregation of simulated ratings. The second set compares optimally aggregated ratings - as determined from our algorithm - to the arithmetic average ratings currently presented on Yelp (and many other websites). The third set shows the impact of shocks on a restaurant's simple and optimal average ratings over time.

5.1 Counterfactuals Across Model Variations Based on Simulated Ratings

The structural results presented in Section 4 stress the importance of incorporating many modeling elements in one single model. But how important is each element in its contribution to an optimal aggregation of ratings? When optimally aggregating, we are making adjustments to remove biases and choosing weights to maximize efficiency. Essentially, we are trying to find weights that will lead to information that is unbiased and as precise as possible. We analyze this question through a series of counterfactual simulations.

The condition in which the simple average is an unbiased and efficient summary of restaurant quality is the following: reviewer signals are i.i.d., restaurant quality is stable, and there is no reviewer popularity weight or bias. To start, we take this condition as the benchmark. In order to highlight how much optimal average is superior than the simple average in each model variation, we add each variation separately to the benchmark and compare by simulation.

For figures 3 and 4, we consider a hypothetical restaurant with a fixed true quality/rating. We then simulate the 95% confidence interval of average ratings that would occur under different aggregation procedures. For figure 5, we consider a hypothetical restaurant with quality change following a martingale process. Because the quality is random variable in this model, we compare the mean absolute error and the mean squared error of the two aggregation procedures in estimating the true restaurant quality when each review is written.

The first model variation we consider allows reviewers to put positive weight on popularity. When popularity concern is the only deviation from the assumption that reviews are i.i.d., then the arithmetic average is unbiased but inefficient. Because later reviews have already incorporated past reviews, an arithmetic average across all reviews assigns too much weight to early reviews. As a result, the optimal average of ratings should give more weight to later reviews. Figure 3 presents two cases, one with $\rho = 1$ and the other with $\rho = 0.6$, while restaurant quality is fixed at 3 and reviewer's signal noise is fixed at $\sigma_\epsilon = 1$. We create these figures by simulating a large number of ratings according to the underlying model, and then computing optimal versus simple average of ratings at each time of review. As shown in Figure 3, optimal average is more efficient than simple average, and the efficiency improvement is greater if reviewers are more concerned about their Yelp popularity. However, the right graph suggests that efficiency gain over simple average is small even if the popularity weight is as large as $\rho = 0.6$. Recall that our structural estimation of ρ never exceeds 0.25 , which suggests that the efficiency gain from accounting for ρ in the optimal average is likely small in the real data.

The second model variation is to allow elite reviewers to have signals that are of different precision than non-elite reviewers. Again, since we are not allowing for reviewers to differ in stringency or for restaurants to change quality, an arithmetic average is going to be unbiased but inefficient. Optimal aggregation endogenously assigns more weight to elite reviews, since elite reviewers have reviews that are more precise. As shown in Figure 4, the more precise the elite reviewers' signals are relative to other reviewers, the larger the efficiency gain is for optimal average versus simple average.

The third model variation adds restaurant quality evolution to the benchmark. Unlike the first two deviations from an i.i.d. distribution of reviews, failing to account for quality change does lead to bias in the arithmetic average ratings. We present three graphs in Figure 5: the first two allow the variance to change every quarterly restaurant quality change with different standard deviation in the noise of quality update, while the third one allows restaurant quality to update monthly with the same σ_ξ as in the second graph. Review frequency is simulated using its empirical distribution as in the raw data. Comparison across the three graphs suggests that the optimal average rating, which accounts for restaurant quality evolution, leads to significant reduction in mean square errors especially when quality update is noisy or frequent.

All three graphs of Figure 5 show the advantage of optimal average over simple average as an estimator of the true restaurant quality when quality follows the assumed random process with initial quality 3. To illustrate the magnitude of bias of optimal and simple average in one realized path of quality, Figure 6 focuses on a hypothetical change of quality from 3 at the beginning, to 2.5 at the 20th review, and to 3.25 at the 40th review. Reviewers believe that true quality is updated by quarter. To focus on the effect of restaurant quality evolution, Figure 6 fixes review frequency at 4.5 days per review. As shown in Figure 6, optimal average tracks the actual quality change better than simple average.

Figure 7 highlights the importance of reviewer stringency (and its heterogeneity). Compared to the benchmark situation, we allow reviewer stringency (λ) to vary by restaurant and reviewer characteristics (including the time trend by restaurant age) according to the coefficients presented in the last column of Table 6. Reviewer and restaurant characteristics are simulated using their empirical distribution as observed in the raw data. The first graph of Figure 7 assumes that the reviewer bias changes with restaurant age, but the restaurant quality does not. And the second graph of Figure 7 assumes that the reviewer bias does not change with restaurant age, and only the restaurant quality does. Both graphs show that optimal average has corrected the bias in reviewer stringency and therefore reflects the true quality, but simple average is biased due to the failure to correct reviewer bias. Figure 8 allows everything else (popularity weight, reviewer signal noise, and restaurant quality change by quarter) and reruns the simulation. Again, because the restaurant quality is random, we use the mean absolute and squared errors to compare the optimal and simple average. The graphs shows that the optimal rating represents the restaurant quality better and the gain is larger when a restaurant receives larger number of reviews.[9]

5.2 Optimal Versus Simple Average for Real Yelp Data

We now compare optimal and simple average based on real Yelp ratings as observed in our data. According to our structural estimates in Table 7, the noise of quality update (σ_ξ) has a standard deviation of 0.081 per month, which amounts to an average deviation of 0.28 stars per year. This is a substantial variation over time as compared to the standard deviation of 1.14 stars in the whole data set over six years. Noise in reviewer signal is even larger, with a standard deviation estimated to be between 0.9 and 1.

These two types of noise have different implications for the relative advantage of optimal average ratings: quality update implies that optimal average needs to give more weight to recent reviews, which is not taken into account by simple average rating. In comparison, simple average reduces the amount of signal noise by law of large number and will do so efficiently unless different reviewers differ in signal precision. Our estimates show a relatively small difference between σ_e and σ_{ne} ($\leq$0.05), implying that optimal weighting due to reviewer heterogeneity in signal noise is unlikely to lead to large efficiency improvement. Another difference between elite and non-elite reviewers is their weight on popularity, but the absolute magnitudes of ρ_e and ρ_{ne} never exceed 0.2, suggesting that the efficiency gain of optimal average due to popularity concern is likely to be small as well. Reviewer's stringency bias is important in magnitude. We know from Table 1 that on average the second review is 155 days apart from the first review; according to the coefficients on Age_{rt}, the second reviewer (if non-elite) will give a rating 0.05 stars lower. Over the six year period in our data, the stringency difference could be as substantial as 0.66 stars.

Including all these elements, we compute simple and optimal average ratings at the time of every observed rating. This calculation is done for monthly, quarterly, and half-year quality update separately, according to the structural estimates in Table 7. We then calculate the difference between simple and optimal average, $\mu_{rn}^{simple} - \mu_{rn}^{optimal}$, for every observation, and plot the mean and confidence interval of this difference by the order of review.

As shown in the first row of Figure 9 (assuming quality updates by quarter), simple average rating is on average close to optimal average, but the confidence interval of their difference ranges from -0.1 to 0.2 stars in early reviews and enlarges gradually as more reviews accumulate. Within each restaurant, we calculate the percent of observations in which simple average rating is more than 0.15 stars away from the optimal average rating. The bar chart on the right hand side plots the histogram of restaurants by this percent. For example, the

[9] In the simulation with full model specifications, the assumption for restaurant age affecting restaurant quality or reviewer bias is nonessential for comparing the mean absolute and squared errors of the two aggregating methods. Optimal rating always corrects any bias in reviewer bias, and simple rating always reflects the sum of the changes in quality and reviewer bias.

second bar shows that roughly 124 restaurants (out of the total 4,101) have 5-10% of times with simple average ratings more than 0.15 stars away from the optimal. Overall, 1,630 restaurants have at least 30% of times with simple average ratings more than 0.15 stars away from the optimal. This suggests that optimal average rating is likely to generate substantial improvement over simple average. The middle and bottom row-blocks of Figure 9 lead to similar conclusion when we model quality update as monthly or half-year instead of quarterly.

Table 8 attempts to decompose the forces driving the difference between optimal and simple average ratings. The above discussion treats time trend ($Age_{rt} \cdot \alpha_{age}$) as reviewer stringency (relative to the first reviewer). By this setting, the optimal average, calculated by $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_n})$, corrects for this trend while the simple average does not. The first panel of Table 8 shows that, by this calculation, 53-55% of restaurants have simple average ratings more than 0.15 stars away from the optimal by the end of the sample, and 26-29% of restaurants have simple average more than 0.25 stars away from the optimal. If instead we interpret the time trend ($Age_{rt} \cdot \alpha_{age}$) as changes in restaurant quality, the optimal average should be $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_n}) + Age_{rt} \cdot \alpha_{age}$. The second panel of Table 8 shows that, after counting the linear trend as quality change (which means it does not contribute to the difference between optimal and simple average), the absolute difference between simple and optimal average rating is still more than 0.15 stars for 25-27% of restaurants, and more than 0.25 stars for 8-10% of restaurants by the end of the data sample.

Overall, in the Yelp setting, the difference between optimal and simple average is mostly driven by restaurant quality updates (σ_ξ) and the time trend ($Age_{rt} \cdot \alpha_{age}$), and less by popularity concerns (ρ), reviewer's signal noise (σ_ϵ), or other terms in reviewer stringency (λ_{rt}).

5.3 Impact of Shocks on Simple and Optimal Averages

Our last set of counterfactual simulations highlight the impact of shocks on rating summary. Following the previous example where a restaurant's true quality starts at 3 stars, jumps down to 2.5 at the time of the 20th review, and jumps back to 3.25 at the 40th review, we consider two types of shocks: one is a shock on true restaurant quality, and the other is a shock in the form of an outlier rating from a particular reviewer.

For the former, Figure 10 presents three cases where the true quality shock (at 1.5 stars) occurs exactly once at either the first, third, or fifth review. For each case, we simulate a large number of reviewer ratings and compute the mean of simple average and optimal average at each review time. The left graph of Figure 10 shows that, when the true quality shock occurs at the time of the first review, both simple average and optimal average captures this shock immediately. But over time when the shock is gone, optimal average recovers faster than simple average and better tracks the concurrent restaurant quality. This is because, in an environment of quarterly quality update, optimal average is designed to give more weight to recent reviews. The other two graphs of Figure 10 confirm this understanding: when the true quality shock happens in the middle of the review history, optimal average is not only closer to the true quality right after the shock, but also catches up faster with the recovered quality later on.

What if the shock occurs in a submitted review instead of true restaurant quality? We simulate this situation in Figure 11. In particular, the three graphs of Figure 11 assume an outlier review of 1.5 stars was submitted as the first, third, or fifth review separately, while the true restaurant quality remains at 3 stars until the time of the 20th review. All the other reviews are simulated in a large number according to the underlying model. We then plot true quality, simple average, and optimal average by order of review. The left graph of Figure 11 shows that, if the outlier review is the first review, over time optimal average has a better ability to shed the influence of this outlier review, because it gives more weight to recent reviews. The right and bottom graphs of Figure 11 suggest that optimal average is not always the best; because it gives more weight to recent reviews, it gives more weight to the outlier review right after it has been submitted, which makes optimal average ratings further away from the actual quality in the short window after the outlier review. However, for the same reason, optimal average also forgets about the outlier review faster than simple average, and better reflects true quality afterward.

6 Conclusion

As consumer reviews are beginning to offer unprecedented amounts of information, this paper argues that the way in which information is aggregated becomes a central design question. To address this question, we offer a method to aggregate consumer ratings into an adjusted weighted average for a given product, where the weights and adjustments are based on the informational content of each review. The informational content, in turn, is empirically determined based on variation in the reviewer characteristics (and review histories), as well as the inferred likelihood that product quality has changed, with parameters set by a model of reviewer behavior. Using the model, it is clear that optimally aggregated information deviates significantly from arithmetic averages. Our method is applicable in a variety of settings. While we have focused on consumer reviews, we could also use this to aggregate expert opinion. In this section, we discuss limits of our model and directions for future research.

6.1 Selection of Reviewers

One limitation of our paper is that we do not observe the selection of consumers who decide to leave a review. In practice, reviewers have selected to purchase a product and also selected to leave a review. In principle, selection into a product would tend to skew ratings upward (you are more likely to eat at a restaurant that you think is a good match). The decision to review has an ambiguous effect, depending on whether people are more likely to review something after a good or bad experience. One could structurally measure this selection function by imposing further assumptions on restaurant preferences. In fact, the information systems literature has documented bimodal distributions of reviews in Amazon (Hu et al 2009), and attributed this to tendency to review when opinions are extreme. While we do not model this selection, we estimate the quality of a reviewer's match to a restaurant using the history of reviews (e.g. some reviewers tend to leave more favorable reviews for Thai food, while others leave better reviews for pizza). Moreover, Yelp reviews do not have the bimodal distribution found by Hu et al as evidence of significant selection problems. This may in part be due to Yelp's focus on encouraging social image and community interaction. We also account for time trend and serial correlation of Yelp reviews within a restaurant, both of which could be generated by reviewer selection.

Two concurrent papers are currently investigating the selection process. Chan et al. (2010) uses a Bayesian learning model on data from a Chinese restaurant review website similar to Yelp.com in order to estimate the way consumers use reviews. They focus on studying how reviewer sorting is affected by social network connection and review content, so they do not consider reviewers' strategic reporting behavior as well as quality change. Our objective is to uncover the optimal representation of restaurant quality, which is quite different from theirs. Wang et al. (2012) also study restaurant reviews, and examine how reviews can influence reviewer behavior in exploring new restaurant choices. Although consumers' variety seeking behavior is not the main theme of our study, we treat it as a heterogeneous reviewer characteristic that may influence reviewer rating. We find that reviewers with a wider variety of reviewing experience are relatively more stringent in ratings.

6.2 Incentives to Write Reviews

Our paper has focused on taking an existing set of reviews and optimally aggregating them to best reflect the quality of a product. An alternative mechanism to achieve this goal is to use incentives to encourage people to leave more representative reviews. These incentives often seem to rely on social image. There is a large theoretical literature studying social image (Akerlof 1980, Bénabou and Tirole 2006). Theoretically modelling a crowdsourced setting, Miller, Resnick and Zeckhauser (2005) present a model arguing that an effective way to encourage high-quality review is rewarding reviewers if their ratings predict peer ratings. Consistent with this theory, Yelp allows members to evaluate each other's reviews, chat online, follow particular reviewers, and meet at offline social events. It also awards elite status to some qualified reviewers who have written a large number of reviews on Yelp. As shown in our estimation, elite reviewers are indeed more consistent with peer ratings, have more precise signals, and place more weight on past reviews of the same restaurant. Wang (2010) compares Yelp reviewers with reviewers on completely anonymous websites such as CitySearch and Yahoo Local. He finds that Yelp reviewers are more likely to write more reviews, productive reviewers are less likely to give extreme ratings, and the same restaurants are less likely to receive extreme ratings on Yelp. Wang (2010) also finds that more prolific Yelp reviewers have more friends on Yelp, receive more anonymous review votes per review, and display more compliment letters per review. These findings motivate us to explicitly model reviewers' reputational concern on Yelp and allow elite and non-elite reviewers to place different value on Yelp popularity.[10]

6.3 Transparency and Aggregation Decisions

Part of the motivation for this paper is that in almost every consumer review website, reviews are aggregated, prompting questions about how one should aggregate reviews. In practice, the most common way to aggregate reviews is using an arithmetic average, which is done by Amazon, Yelp, TripAdvisor, and many others. As we have highlighted in this paper, arithmetic average does not account for reviewer biases, reviewer heterogeneity, or changing quality.

Another important caveat of our method is that there are reasons outside of our model that may prompt a review website to use an arithmetic average. For example, arithmetic averages are transparent and uncontroversial. If, for example, Yelp were to use optimal information aggregation, they may be accused of trying to help certain restaurants due to a conflict of interest (since Yelp also sells advertisements to restaurants). Hence, a consumer review website's strategy might balance the informational benefits of optimal information aggregation against other incentives that may move them away from this standard, such as conflict of interest (or even the desire to avoid perceived conflict of interest).

6.4 Customized Recommendations

Our paper has attempted to aggregate information into a single comparable signal of quality. Once this is done, it could be extended to then customize recommendations based on the readers horizontal prefences. For example, Netflix tailors recommendations based on other reviewers with similar tastes. This type of recommendation relies both on an understanding of underlying quality (as in this paper), as well as a sense of horizontal preferences of readers.

6.5 Text Analysis

In this paper, we have focused on using only numerical data. However, a productive literature has begun to use text analysis to extract informational content from reviews. For examples, see Ghose and Ipeirotis (2011), Archak, Ghose and Ipeirotis (2011) and Ghose, Ipeirotis, and Li (2012). Ghose, Ipeirotis, and Li (2012) estimate the consumer demand model and argues that the ranking systems should be designed to reflect consumer demand besides price and star ratings. Ghose and Ipeirotis (2011) and Archak, Ghose, and Ipeirotis (2011) examine the

[10] There is a large literature on social image and social influence, with most evidence demonstrated in lab or field experiments. For example, Ariely et al. (2009) show that social image is important for charity giving and private monetary incentives partially crowd out the image motivation.

impact of different product attributes and reviewer opinions on product sales, and propose a model to identify segments of text review that describe products' multifaceted attributes. Although this is beyond the scope of the current paper, one could incorporate text analysis methods into optimal information aggregation.

6.6 Generalizing Our Approach

In principle, the method offered in our paper could be applied to a variety of review systems. Implementing this could also be done in conjunction with the other considerations discussed above. Moreover, when generalizing our method, the relative importance of various factors in our model could vary by context. For example, quality change is not an issue for a fixed product such as book, movie, etc, reviewer heterogeneity could be much more important when expert and non-expert reviews are common but treated differently by the market (say for books and movies). The flexibility of our model allows it to be robust to this type of variation, while also allowing for new insights by applying the model to different settings.

References

Archak, Nikolay, Anindya Ghose, and Panagiotis G. Ipeirotis. "Deriving the pricing power of product features by mining consumer reviews." *Management Science* 57, no. 8 (2011): 1485-1509.

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* 19 (6): 716–723.

Akerlof, George A. (1980) "A Theory of Social Custom, of Which Unemployment May Be One Consequence." *Quarterly Journal of Economics*, 94(4): 749-75.

Ariely, Dan, Anat Bracha, and Stephan Meirer. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99(1): 544-555.

Banerjee, Abhijit V. "A simple model of herd behavior." *The Quarterly Journal of Economics* 107.3 (1992): 797-817.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96(5): 1652-78.

Brown, Jennifer, Tanjim Hossain and John Morgan (2010) "Shrouded Attributes and Information Suppression: Evidence from the Field." *Quarterly Journal of Economics.* 125(2): 859-876.

Chan, Tat, Hai Che, Chunhua Wu, and Xianghua Lu, working paper "Social Network Learning: How User Generated Content on Review Website Influence Consumer Decisions"

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

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li (2010) "Social Comparison and Contributions to Online Communities: A Field Experiment on MovieLens." *American*

Economic Review, 100(4): 1358-98.

Dellarocas, Chrysanthos (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, 52(10):1577–1593.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston. "Do online reviews matter?—An empirical investigation of panel data." *Decision Support Systems* 45, no. 4 (2008): 1007-1016.

Ghose, Anindya and Panagiotis G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, October 2011.

Ghose, A., P. Ipeirotis, B. Li (forthcoming) "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content" *Marketing Science*.

Glazer, Jacob, Thomas G. McGuire, Zhun Cao, and Alan Zaslavsky. 2008. "Using Global Ratings of Health Plans to Improve the Quality of Health Care." *Journal of Health Economics*, 27(5): 1182–95.

Godes, David and Dina Mayzlin. Firm-created word-of-mout communication: Evidence from a field test. *Marketing Science*, 28(4):721–739, 2009.

Goldenberg, Jacob, Barak Libai, and Eitan Muller. "The chilling effects of network externalities." *International Journal of Research in Marketing* 27, no. 1 (2010): 4-15.

Alevy, Jonathan E., Michael S. Haigh, and John A. List. "Information cascades: Evidence from a field experiment with financial market professionals." *The Journal of Finance* 62, no. 1 (2007): 151-180.

Hitt, Lorin and Xinxin Li. Self-selection and information role of online product reviews. *Information Systems Research*, 19:456–474, 2008.

Hu, N.; Liu, L., and Zhang, J. (2008) "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects" *Information Technology Managment* 9(3): 201-214.

Hu, Mingqing and Bing Liu (2004). "Mining and summarizing customer reviews" *Proceedings of the tenth ACM SIGKDD.*

Hu, Nan, Jie Zhang and Paul Pavlou (2009) "Overcoming the J-shaped distribution of product reviews", *Communication ACM.*

Li, Xinxin and Lorin Hitt (2008) "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* 19(4): 456-474.

Luca, Michael and Jonathan Smith (forthcoming) "Salience in Quality Disclosure: Evidence from The US News College Rankings," *Journal of Economics & Management Strategy*.

Luca, Michael (2011) "Reviews, Reputation, and Revenue: The Case of Yelp.com." *Harvard Business School working paper*.

Mayzlin, Dina. "Promotional chat on the Internet." *Marketing Science* 25, no. 2 (2006): 155-163.

Miller, Nolan, Paul Resnick, and Richard J. Zeckhauser (2005). "Eliciting Informative

Feedback: The Peer- Prediction Method." *Management Science*, 51(9): 1359–73.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012). "Mine Your Own Business: Market Structure Surveillance Through Text Mining." *Marketing Science* 31 (3).

Pope, Devin (2009). "Reacting to Rankings: Evidence from 'America's Best Hospitals'," *Journal of Health Economics*, Vol. 28, No. 6, 1154-1165.

Wang, Qingliang, Khim Yong Goh, Xianghua Lu, 2012. "How does user generated content influence consumers' new product exploration and choice diversity? An empirical analysis of product reviews and consumer variety seeking behaviors," working paper.

Wang, Zhongmin "Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews" *The B.E. Journal of Economic Analysis & Policy*, 2010.

Appendix A: Derive $E(\mu_{rt}|s_{rt_1},...s_{rt_n})$

For restaurant r, denote the prior belief of μ_{rt_n} right before the realization of the n^{th} signal as

$$\pi_{n|n-1}(\mu_{rt_n}) = f(\mu_{rt_n}|s_{rt_1},...s_{rt_{n-1}})$$

and we assume that the first reviewer uses an uninformative prior

$$\mu_{1|0} = 0, \sigma^2_{1|0} = W,\ W\ arbitrarily\ large$$

Denote the posterior belief of μ_{rt_n} after observing s_{rt_n} as

$$h_{n|n}(\mu_{rt_n}) = f(\mu_{rt_n}|s_{rt_1},...s_{rt_n})$$

Hence

$$\begin{aligned} h_{n|n}(\mu_{rt_n}) = f(\mu_{rt_n}|s_{rt_1},...s_{rt_n}) =& \frac{f(\mu_{rt_n},s_{rt_1},...s_{rt_n})}{f(s_{rt_1},...s_{rt_n})} \\ \propto& f(\mu_{rt_n},s_{rt_1},...s_{rt_n}) \\ =& f(s_{rt_n}|\mu_{rt_n},s_{rt_1},...s_{rt_{n-1}})f(\mu_{rt_n},s_{rt_1},...s_{rt_{n-1}}) \\ =& f(s_{rt_n}|\mu_{rt_n},s_{rt_1},...s_{rt_{n-1}})f(\mu_{rt_n}|s_{rt_1},...s_{rt_{n-1}})f(s_{rt_1},...s_{rt_{n-1}}) \\ \propto& f(s_{rt_n}|\mu_{rt_n})f(\mu_{rt_n}|s_{rt_1},...s_{rt_{n-1}}) \\ =& f(s_{rt_n}|\mu_{rt_n})\pi_{n|n-1}(\mu_{rt_n}) \end{aligned}$$

where $f(s_{rt_n}|\mu_{rt_n},s_{rt_1},...s_{rt_{n-1}}) = f(s_{rt_n}|\mu_{rt_n})$ comes from the assumption that s_{rt_n} is independent of past signals conditional on μ_{rt_n}.

In the above formula, the prior belief of μ_{rt_n} given the realization of $\{s_{rt_1},...,s_{rt_{n-1}}\}$, or

$\pi_{n|n-1}(\mu_{rt_n})$, depends on the posterior belief of $\mu_{rt_{n-1}}$, $h_{n-1|n-1}(\mu_{rt_{n-1}})$ and the evolution process from $\mu_{rt_{n-1}}$ to μ_{rt_n}, denoted as $g(\mu_n|\mu_{n-1})$. Hence,

$$\pi_{n|n-1}(\mu_{rt_n}) = g(\mu_n|\mu_{n-1})f(\mu_{rt_{n-1}}|s_{rt_1}, ...s_{rt_{n-1}}) = g(\mu_n|\mu_{n-1})h_{n-1|n-1}(\mu_{rt_{n-1}})$$

Given the normality of $\pi_{n|n-1}$, $f(s_{rt_n}|\mu_{rt_n})$ and $g(\mu_n|\mu_{n-1})$, $h_{n|n}(\mu_{rt_n})$ is distributed normal. In addition, denote $\mu_{n|n}$ and $\sigma^2_{n|n}$ as the mean and variance for random variable with normal probability density function $p_{n|n-1}(\mu_{rt_n})$, $\mu_{n|n-1}$ and $\sigma^2_{n|n-1}$ are the mean and variance of random variable with normal pdf $h_{n|n}(\mu_{rt_n})$. After combining terms in the derivation of $p_{n|n-1}(\mu_{rt_n})$ and $h_{n|n}(\mu_{rt_n})$, the mean and variance evolves according to the following rule:

$$\begin{aligned}
\mu_{n|n} &= \mu_{n|n-1} + \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_n}(s_n - \mu_{n|n-1}) \\
&= \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_n} s_n + \frac{\sigma^2_n}{\sigma^2_{n|n-1} + \sigma^2_n} \mu_{n|n-1} \\
\sigma^2_{n|n} &= \frac{\sigma^2_n \sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_n} \\
\mu_{n+1|n} &= \mu_{n|n} \\
\sigma^2_{n+1|n} &= \sigma^2_{n|n} + (t_{n+1} - t_n)\sigma^2_\xi
\end{aligned}$$

Hence, we can deduct the beliefs from the initial prior,

$$\begin{aligned}
\mu_{1|0} &= 0 \\
\sigma^2_{1|0} &= W > 0 \textit{ and arbitrarily large} \\
\mu_{1|1} &= s_1 \\
\sigma^2_{1|1} &= \sigma^2_1 \\
\mu_{2|1} &= s_1 \\
\sigma^2_{2|1} &= \sigma^2_1 + (t_2 - t_1)\sigma^2_\xi \\
\mu_{2|2} &= \frac{\sigma^2_1 + (t_2 - t_1)\sigma^2_\xi}{\sigma^2_1 + \sigma^2_2 + (t_2 - t_1)\sigma^2_\xi} s_2 + \frac{\sigma^2_2}{\sigma^2_1 + \sigma^2_2 + (t_2 - t_1)\sigma^2_\xi} s_1 \\
\sigma^2_{2|2} &= \frac{\sigma^2_2(\sigma^2_1 + (t_2 - t_1)\sigma^2_\xi)}{\sigma^2_1 + \sigma^2_2 + (t_2 - t_1)\sigma^2_\xi} \\
\mu_{3|2} &= \mu_{2|2} \\
\sigma^2_{3|2} &= \frac{\sigma^2_2(\sigma^2_1 + (t_2 - t_1)\sigma^2_\xi)}{\sigma^2_1 + \sigma^2_2 + (t_2 - t_1)\sigma^2_\xi} + (t_3 - t_2)\sigma^2_\xi \\
&...
\end{aligned}$$

$E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n}) = \mu_{n|n}$ is derived recursively following the above formulation.

Appendix B: Solve $\{s_{rt_1}, ...s_{rt_n}\}$ from $\{x_{rt_1}, ...x_{rt_n}\}$ according to the following recursive formula:

$$
\begin{aligned}
x_1 &= s_1 + \lambda_1 \\
s_1 &= x_1 - \lambda_1 \\
x_2 &= \rho_2 \frac{\sigma_2^2}{\sigma_{2|1}^2 + \sigma_2^2} \mu_{2|1} + \rho_2 \frac{\sigma_{2|1}^2}{\sigma_{2|1}^2 + \sigma_2^2} s_2 + (1 - \rho_2) s_2 + \lambda_2 \\
&= \rho_2 \frac{\sigma_2^2}{\sigma_{2|1}^2 + \sigma_2^2} \mu_{2|1} + [1 - (1 - \frac{\sigma_{2|1}^2}{\sigma_{2|1}^2 + \sigma_2^2}) \rho_2] s_2 + \lambda_2 \\
s_2 &= \frac{1}{[1 - (1 - \frac{\sigma_{2|1}^2}{\sigma_{2|1}^2 + \sigma_2^2}) \rho_2]} [x_2 - \lambda_2 - \rho_2 \frac{\sigma_2^2}{\sigma_{2|1}^2 + \sigma_2^2} \mu_{2|1}] \\
&... \\
s_n &= \frac{1}{[1 - (1 - \frac{\sigma_{n|n-1}^2}{\sigma_{n|n-1}^2 + \sigma_n^2}) \rho_n]} [x_n - \lambda_n - \rho_n \frac{\sigma_n^2}{\sigma_{n|n-1}^2 + \sigma_n^2} \mu_{n|n-1}].
\end{aligned}
$$

Table 1: **Summary Statistics**

Variable	Mean	Med.	Min.	Max.	Std. Dev.	N.[a]
Restaurant Characteristics						
Reviews per Restaurant	32.85	14.00	1.00	698.00	50.20	4,101
Reviews per Day	0.16	0.03	0.00	5.00	0.33	4,101
Days between 1^{st} and 2^{nd} Review	154.75	79.00	0.00	1,544.00	199.95	3,651
Days between 11^{st} and 12^{nd} Review	33.96	20.00	0.00	519.00	41.71	2,199
Days between 21^{st} and 22^{nd} Review	20.63	13.00	0.00	234.00	25.27	1,649
Reviewer Characteristics						
Rating	3.74	4.00	1.00	5.00	1.14	134,730
by Elite	3.72	4.00	1.00	5.00	1.10	43,781
by Non-elite	3.75	4.00	1.00	5.00	1.18	90,949
Reviews per reviewer	7.18	2.00	1.00	453.00	17.25	18,778
by Elite	24.49	6.00	1.00	350.00	39.23	1,788
by Non-elite	5.35	2.00	1.00	453.00	11.49	16,990
Reviews per Day	0.12	0.17	0.00	1.52	0.07	18,778
by Elite	0.15	0.22	0.00	1.30	0.10	1,788
by Non-elite	0.12	0.16	0.00	1.52	0.07	16,990
Reviewer-Restaurant Matching Distance[b]	12.18	8.51	0.00	108.00	11.45	134,730
by Elite	11.26	7.47	0.00	108.00	10.77	43,781
by Non-elite	12.62	9.00	0.00	103.73	11.74	90,949
Reviewer Taste for Variety[c]	1.10	1.11	0.00	2.60	0.24	103,835
by Elite	1.11	1.12	0.00	2.60	0.17	40,521
by Non-elite	1.09	1.10	0.00	2.52	0.27	63,314

[a] There are 4,101 restaurants with a total of 134,730 reviews written by 18,778 reviewers in our sample.

[b] Reviewer-restaurant matching distance quantifies the quality of the match between a reviewer and a restaurant, calculated using information on restaurants the reviewer has previously rated. See page 7 for the formula and more thorough discussion.

[c] Reviewer taste for variety quantifies the reviewer's preference for variety in restaurants, calculated using information on restaurants the reviewer has previously rated. See page 7 for the formula and more thorough discussion.

Table 2: **What Explains the Variance of Yelp Ratings?**

Model	**Variance Explained (R^2)**
Reviewer FE	0.2329
Restaurant FE	0.2086
Reviewer FE & Restaurant FE	0.3595
Reviewer FE & Restaurant FE & Year FE	0.3595
Reviewer FE & Restaurant FE & Year FE & Matching & Variety	0.3749

Notes: [1] This table presents the R^2 from linear regressions of Yelp ratings on the fixed effects and variables indicated in each row.

[2] There are only a few observations in 2004 and 2010. Whenever the year fixed effect is added, 2005 year fixed effect was applied to the 2004 observations and 2009 year fixed effect was applied for the 2010 observations.

Table 3: **Does the Variability of Ratings Decline over Time?**

Model:^{a} $\widehat{\epsilon_{ri,yr}}^2 = \beta_0 + \beta_1 D_{ri,elite} + \beta_2 N_{ri} + \beta_3 N_{ri} \times D_{ri,elite} + \zeta_{ri,yr}$

$D_{ri}^{elite b}$	-12.000***	(0.940)
${N_{ri}}^c(100s)$	-0.021**	(0.007)
$D_{ri}^{elite} \times N_{ri}(100s)$	-0.009	(0.012)
constant	88.000***	(0.581)
N	134,730	

[a] $\widehat{\epsilon_{ri,yr}}$ is the residual from regression $Rating_{ri,year} = \mu_r + \alpha_i + \gamma_{year} + \epsilon_{ri,year}$

[b] D_{ri}^{elite} is the dummy variable for elite reviewer.

[c] N_{ri} indicates that reviewer i writes the N^{th} review on restaurant r.

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: **Restaurant Ratings are Serially Correlated**

Model:^{a} $\widehat{\epsilon_{ri,yr}} = \sum_{s=1}^{k} \beta_s \widehat{\epsilon_{r,i-s,yr}} + \eta_{ri,yr}$

	(1)	(2)	(3)	(4)
$\widehat{\epsilon_{r,i-1,yr}}$	0.0428***	0.0433***	0.0429***	0.0423***
	(0.0029)	(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-2,yr}}$	0.0299***	0.0300***	0.0299***	0.0311***
	(0.0029)	(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-3,yr}}$	0.0213***	0.0208***	0.0209***	0.0213***
	(0.0029)	(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-4,yr}}$	0.0151***	0.0146***	0.0145***	0.0148***
	(0.0029)	(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-5yr}}$	0.0126***	0.0117***	0.0111***	0.0110***
	(0.0029)	(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-5,yr}}$		0.0087**	0.0081**	0.0084**
		(0.0030)	(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-6,yr}}$			0.00991***	0.00996**
			(0.00300)	(0.00303)
$\widehat{\epsilon_{r,i-7,yr}}$				0.00312
				(0.00303)
Constant	-0.00629*	-0.00782**	-0.00856**	-0.00971***
	(0.00266)	(0.00269)	(0.00272)	(0.00275)
Observations	117,536	114,742	112,067	109,505

Notes: This table estimates the degree of serial correlation of ratings within a restaurant.

[a] $\widehat{\epsilon_{ri,yr}}$ is the residual from regressing $Rating_{ri,year} = \mu_r + \alpha_i + \gamma_{year} + \epsilon_{ri,year}$. To obtain sequential correlation of the residuals, we regress the residuals on their lags $\widehat{\epsilon_{ri-s,yr}}$, where s is the number of lag.

*** Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: **Does Matching Improve Over Time?**

	For Restaurants		For Reviewers	
	(1)	(2)	(3)	(4)
	Matching Distance[a]	Taste for Variety[b]	Matching Distance[a]	Taste for Variety[b]
Restaurant's n^{th} Review	0.00170	-0.00025***		
	(0.00117)	(0.00006)		
(Restaurant's n^{th} Review)^2	-0.00002	1.17e-6***		
	(0.00001)	(3.25e-7)		
(Restaurant's n^{th} Review)^3	1.06e-08	-1.54e-09***		
	(8.00e-09)	(4.01e-10)		
Reviewer's n^{th} Review			-0.06700***	0.00169***
			(0.00140)	(0.00006)
(Reviewer's n^{th} Review)^2			0.00045***	-0.00001***
			(0.00001)	(4.46e-7)
(Reviewer's n^{th} Review)^3			7.57e-7***	1.95e-08***
			(2.14e-08)	(8.35e-10)
Constant	12.13***	1.104***	12.57***	1.066***
	(0.03590)	(0.00157)	(0.02330)	(0.00103)
Observations	134,730	103,835	103,835	103,835

Notes: The sample sizes of regressions specified in column (2), (3) and (4) are smaller since we dropped the first reviews written by a reviewer. Because reviewer does not have any review history when she reviews for the first time, we set the two measures at their overall sample mean.

[a] Reviewer-restaurant matching distance quantifies the quality of the match between a reviewer and a restaurant, calculated using information on restaurants the reviewer has previously rated. See page 7 for the formula and more thorough discussion.

[b] Reviewer taste for variety quantifies the reviewer's preference for variety in restaurants, calculated using information on restaurants the reviewer has previously rated. See page 7 for the formula and more thorough discussion.

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6: **MLE Estimates: Signal Precision and Popularity Concerns**

	(1)	*(2)*	*(3)*	*(4)*
	Same	*Different*	*Quarterly*	*Full Model*
	σ, ρ	σ, ρ	*Quality Change*	*(Quarterly*
				Quality Change)
σ_e	1.2218***	1.1753***	0.9293***	0.9101***
	(0.0210)	(0.0210)	(0.0199)	(0.0196)
σ_{ne}		1.2350***	0.9850***	0.9546***
		0.0147	(0.0156)	(0.0151)
σ_ξ			0.1452***	0.1345***
			(0.0038)	(0.0039)
ρ_e	0.1718***	0.2430***	0.0454***	0.0230***
	(0.0007)	(0.0141)	(0.0215)	(0.0220)
ρ_{ne}		0.1362***	-0.0821^{***}	-0.1182^{***}
		(0.0110)	(0.0181)	(0.0187)
$(\lambda_e - \lambda_{ne})_0$		-0.0100	-0.0061	0.0007
		(0.0059)	(0.0059)	(0.0223)
$(\lambda_e - \lambda_{ne})_{Age}$				$3.8 \times 10^{-5***}$
				(1.3×10^{-5})
$(\lambda_e - \lambda_{ne})_{MatchD}$				-0.0002
				(0.0005)
$(\lambda_e - \lambda_{ne})_{TasteVar}$				-0.0025
				(0.0064)
$(\lambda_e - \lambda_{ne})_{FreqRev}$				0.0647**
				(0.0319)
$(\lambda_e - \lambda_{ne})_{NumRev}$				0.0006***
				(0.00015)
$(\mu + \lambda_{ne})_{Age}$				-0.0002^{***}
				(1.3×10^{-5})
$(\mu + \lambda_{ne})_{MatchD}$				0.0042***
				(0.0005)
$(\mu + \lambda_{ne})_{TasteVar}$				-0.0224^{***}
				(0.0029)
$(\mu + \lambda_{ne})_{FreqRev}$				-0.0348
				(0.0218)
$(\mu + \lambda_{ne})_{NumRev}$				-0.0007^{***}
				(0.0001)
Log Likelihood	-193,339	-192,538	-192,085	-191,829
N	122,473	122,473	122,473	122,473

Notes: 1. "e" and "ne" in the subscript indicate elite and non-elite status respectively.

2. Parameters in row (7) to (16) represent the effects of review-restaurant matching and reviewer attributes on sum of restaurant quality and non-elite stringency ($\lambda_{ne} + \mu$) and stringency difference between elite and non-elite reviewers. The subscripts that represent review-restaurant matching and reviewer attributes are reviewer-restaurant matching distance ($MatchD$), reviewer taste for variety ($TasteVar$), number of reviews written by the reviewer per day ($FreqRev$), and total number of reviews written by the reviewer ($NumRev$). Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7: **MLE Estimates with Changing Restaurant Quality**

	(1)	*(2)*	*(3)*
	Full Model	*Full Model*	*Full Model*
	(Quarterly	*(Half-yearly*	*(Monthly*
	Quality Change)	*Quality Change)*	*Quality Change)*
σ_e	0.9101^{***}	0.9365^{***}	0.8983^{***}
	(0.0196)	(0.0196)	(0.0196)
σ_{ne}	0.9546^{***}	0.9778^{***}	0.9440^{***}
	(0.0151)	(0.0150)	(0.0153)
σ_ξ	0.1345^{***}	0.1736^{***}	0.0810^{***}
	(0.0039)	(0.0051)	(0.0023)
ρ_e	0.0230^{***}	0.0499^{***}	0.0103^{***}
	(0.0220)	(0.0209)	(0.0227)
ρ_{ne}	-0.1182^{***}	-0.0922^{***}	-0.1307^{***}
	(0.0187)	(0.0176)	(0.0193)
$(\lambda_e - \lambda_{ne})_0$	0.0007	-0..00057	0.0007
	(0.0223)	(0.0223)	(0.0223)
$(\lambda_e - \lambda_{ne})_{Age}$	$3.8 \times 10^{-5***}$	$3.9 \times 10^{-5***}$	$3.8 \times 10^{-5***}$
	(1.3×10^{-5})	(1.3×10^{-5})	(1.3×10^{-5})
$(\lambda_e - \lambda_{ne})_{MatchD}$	-0.0002	-0.0001	-0.0002
	(0.0005)	(0.0005)	(0.0005)
$(\lambda_e - \lambda_{ne})_{TasteVar}$	-0.0025	-0.0025	-0.0024
	(0.0064)	(0.0064)	(0.0064)
$(\lambda_e - \lambda_{ne})_{FreqRev}$	0.0647^{**}	0.0657^{**}	0.0640^{**}
	(0.0319)	(0.0319)	(0.032)
$(\lambda_e - \lambda_{ne})_{NumRev}$	0.0006^{***}	0.0006	0.0006^{***}
	(0.00015)	(0.00055)	(0.00015)
$(\mu + \lambda_{ne})_{Age}$	-0.0002^{***}	-0.0002^{***}	-0.0003^{***}
	(1.3×10^{-5})	(1.3×10^{-5})	(1.4×10^{-5})
$(\mu + \lambda_{ne})_{MatchD}$	0.0042^{***}	0.0040^{***}	0.0040^{***}
	(0.0005)	(0.0005)	(0.0005)
$(\mu + \lambda_{ne})_{TasteVar}$	-0.0224^{***}	-0.0233^{***}	-0.0233^{***}
	(0.0029)	(0.0029)	(0.0029)
$(\mu + \lambda_{ne})_{FreqRev}$	-0.0348	-0.0347	-0.0353
	(0.0218)	(0.0218)	(0.0218)
$(\mu + \lambda_{ne})_{NumRev}$	-0.0007^{***}	-0.0007	-0.0007^{***}
	(0.0001)	(-0.0007)	(-0.0001)
Log Likelihood	-191,829	-191,870	-191,814
N	122,473	122,473	122,473

Notes: 1. Full model specification allows random walk of restaurant quality in different frequency, and both restaurant quality and reviewer stringency to vary flexibly with reviewer-restaurant matching matrices and reviewer attributes.

2. "e" and "ne" in the subscript indicate elite and non-elite status respectively.

3. Parameters in row (7) to (16) represent the effects of review-restaurant matching and reviewer attributes on sum of restaurant quality and non-elite stringency ($\lambda_{ne} + \mu$) and stringency difference between elite and non-elite reviewers. The subscripts that represent review-restaurant matching and reviewer attributes are reviewer-restaurant matching distance ($MatchD$), reviewer taste for variety ($TasteVar$), number of reviews written by the reviewer per day ($FreqRev$), and total number of reviews written by the reviewer ($NumRev$).

*** Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

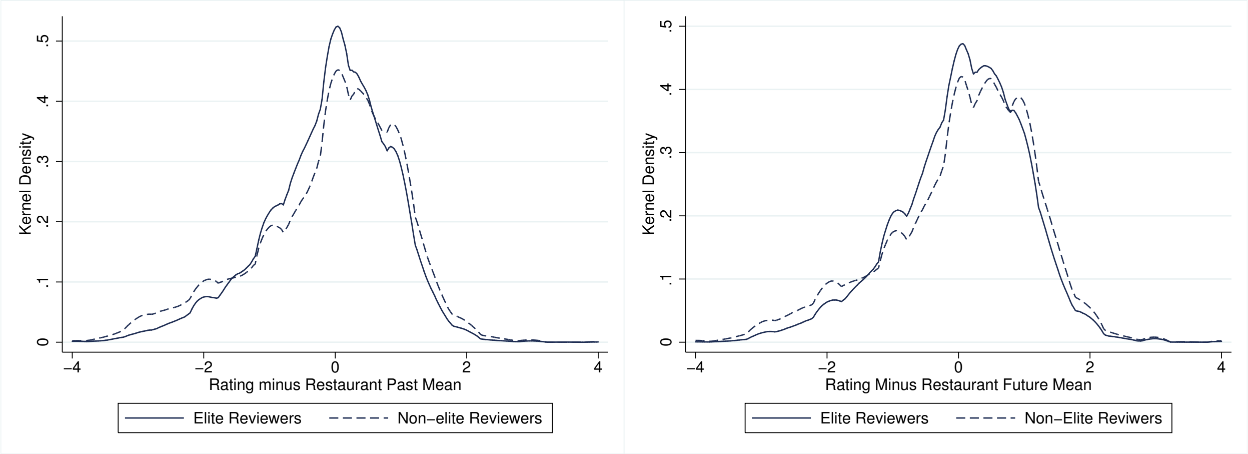
Table 8: **Optimal and Simple Average Algorithm Applied on the Real Data**

μ Update Frequency	% $\Delta\mu < -0.15$	% $\Delta\mu < -0.25$	% $\Delta\mu < -0.35$	% $\Delta\mu > 0.15$	% $\Delta\mu > 0.25$	% $\Delta\mu > 0.35$
(*i*) *Restaurant quality unaffected by restaurant age*						
Quarterly	54.08%	28.40%	13.42%	0.33%	0.06%	0.03%
Half-yearly	53.48%	26.43%	11.09%	0.18%	0.06%	0%
Monthly	54.17%	29.21%	13.69%	0.39%	0.09%	0.03%
(*ii*) *Restaurant quality affected only by restaurant age*						
Quarterly	2.57%	0.54%	0.09%	24.39%	9.00%	2.87%
Half-yearly	1.82%	0.39%	0%	23.02%	7.62%	2.30%
Monthly	2.99%	0.63%	0.15%	24.57%	9.72%	3.32%

Notes: 1. $\Delta\mu \equiv \mu^{simple} - \mu^{optimal}$ calculates the difference between simple average μ^{simple} and optimal average $\mu^{optimal}$ when a restaurant receives its last rating as of the end of our sample. $\%(\Delta\mu > x)$ calculates the percentage of restaurants having optimal and simple average difference greater than x.

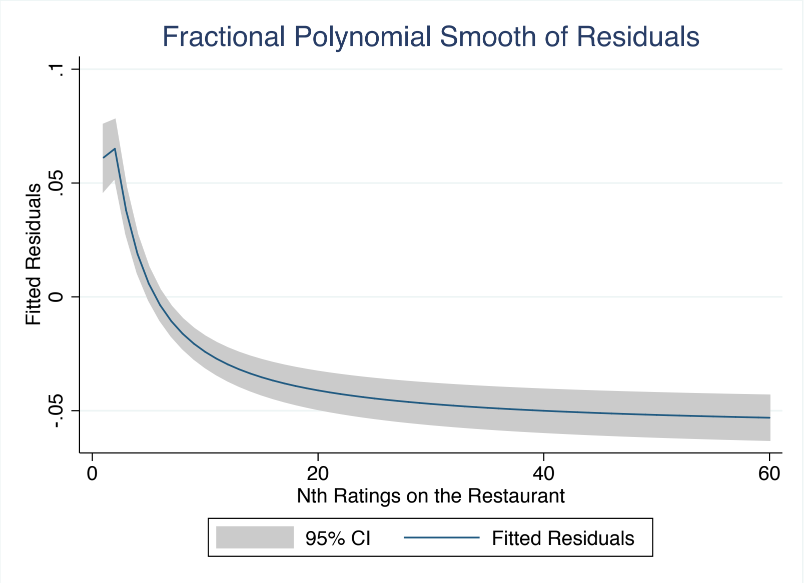
2. Given that in $\partial(\mu_{rn} + \lambda_{rn}^{ne})/\partial n$, we cannot separately identify the impact of restaurant age n on restaurant quality μ_{rn} and reviewer bias λ_{rn}^{ne}, we calculate $\mu^{simple} - \mu^{optimal}$ in the model assuming that restaurant age only affects λ_{rn}^{ne} (in panel (*i*)), and assuming that restaurant age only affects μ_{rn} (in panel (*ii*))

Figure 1: **The Distribution of Ratings by Elite Status**



Notes. 1. The figure on the left plots $Rating_{rn} - \overline{Rating}_{rn}^{BF}$, where $Rating_{rn}$ is the n^{th} rating on restaurant r, and $\overline{Rating}_{rn}^{BF}$ is the arithmetic mean of all past $n-1$ ratings on restaurant r before n. Similarly, figure on the right plots $Rating_{rn} - \overline{Rating}_{rn}^{AF}$, where $Rating_{rn}$ is the n^{th} rating on restaurant r, and $\overline{Rating}_{rn}^{AF}$ is the arithmetic mean of all future ratings on restaurant r until the end of the sample. 2. This figure shows the distribution of ratings relative to the restaurant mean. Ratings by elite reviewers tend to be closer to a restaurant's average rating, and have shorter tails.

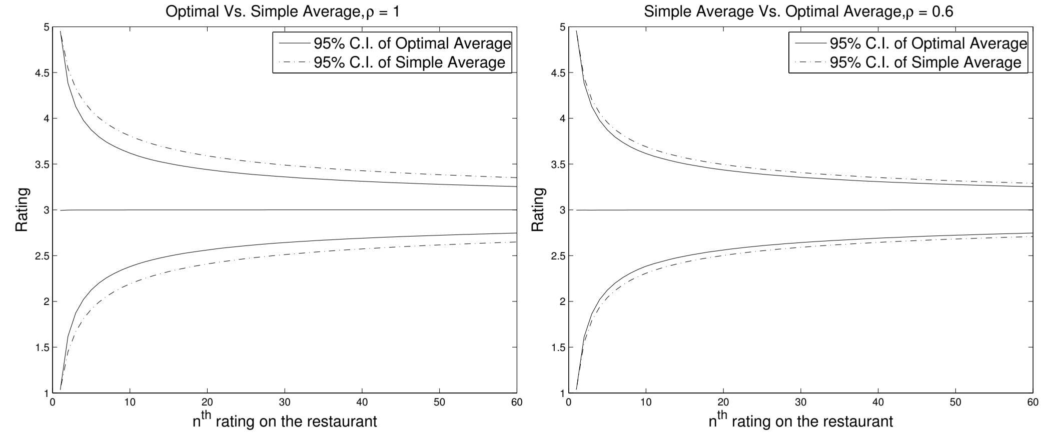
Figure 2: **Restaurants Experience a "Chilling Effect"**



Notes. This figure shows the trend of ratings within a restaurant over time. Restaurants begin with more favorable reviews, which deteriorate over time. It plots the fractional polynomial of the restaurant residual on the sequence of reviews. Residuals are obtained from regression $Rating_{rn,year} = \mu_r + \gamma_{year} + \epsilon_{rn,year}$.

Figure 3: **Comparing Optimal and Simple Averages: Reviewers with Popularity Concern**

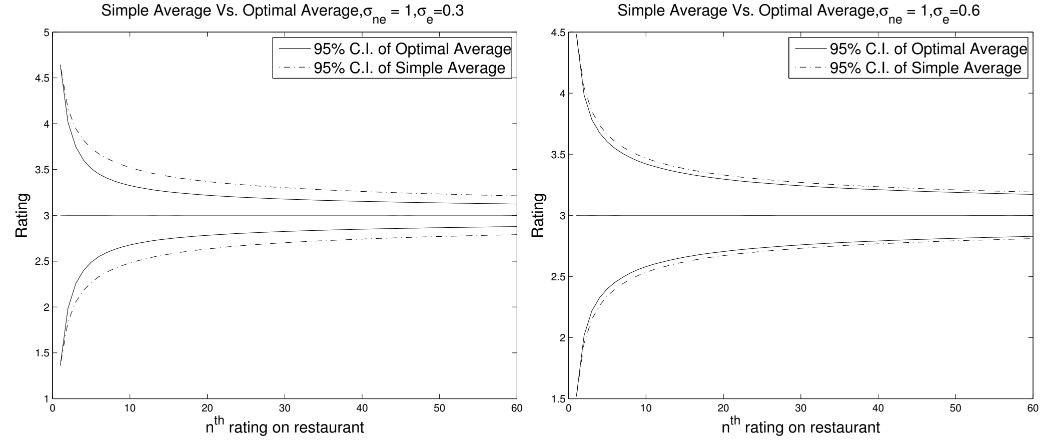
Parameters	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 1$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0.6$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$



Notes: The above figures plot the simulated 95% confidence interval for the average ratings that would occur for a restaurant at a given quality level. When reviewers have popularity concern, arithmetic and optimal averages are unbiased estimates for true quality. But relative to arithmetic average, optimal aggregation allows ratings to converge to the true quality more quickly. And the difference in converging speeds increases with elite reviewers' popularity concern relative to that of non-elite reviewers.

Figure 4: **Comparing Optimal and Simple Averages: Reviewers with Different Precisions**

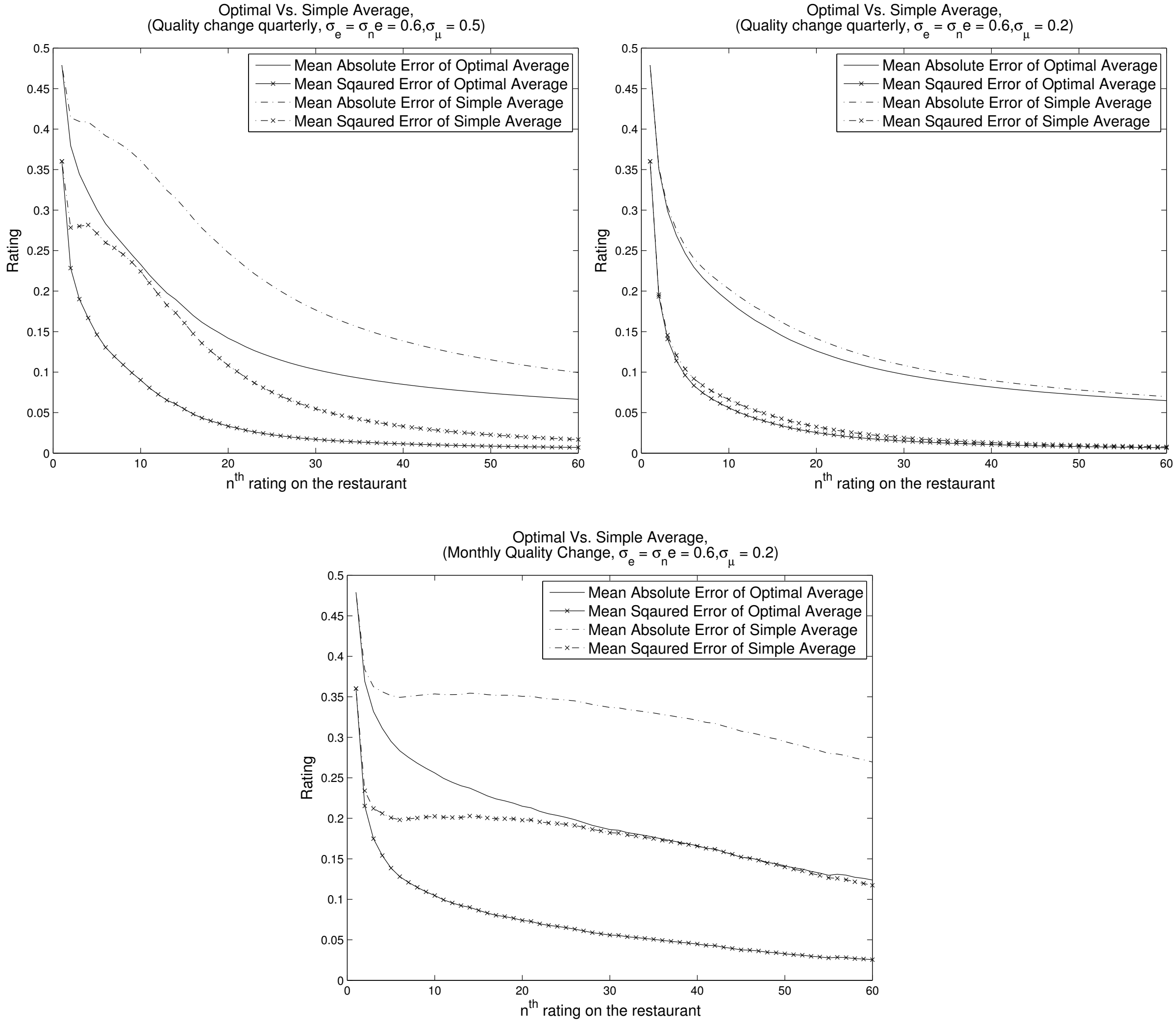
Parameter	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.6$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.3$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$



Notes: The above figures plot the simulated 95% confidence interval for the average ratings that would occur for a restaurant at a given quality level. When reviewers differ in precision, both arithmetic and optimal averages are unbiased estimates for true quality. But relative to arithmetic average, optimal aggregation allows ratings to converge to the true quality more quickly. The difference in converging speed increases with elite reviewers' precision relative to that of non-elite reviewers.

Figure 5: **Comparing Optimal and Simple Averages: Martingale Quality Change**

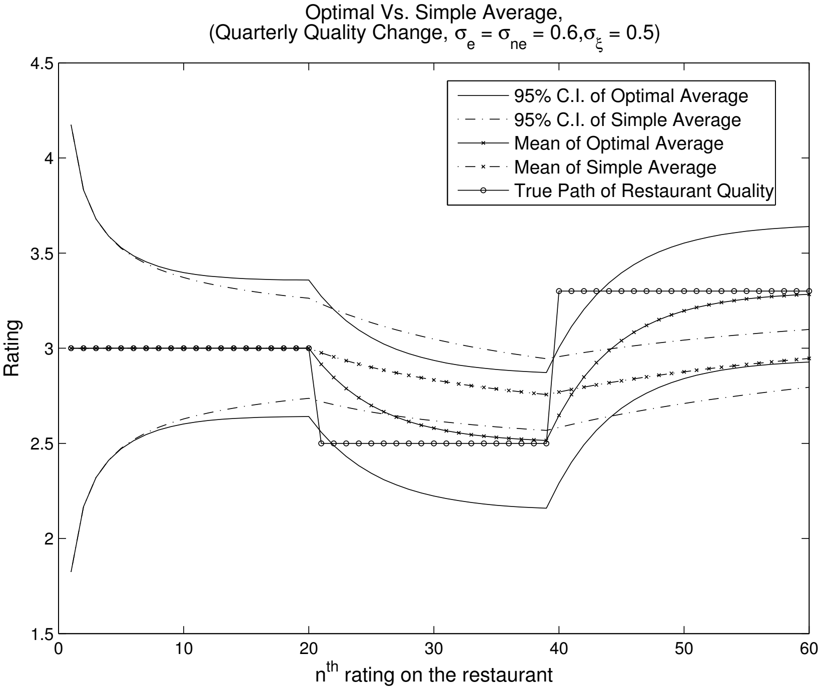
	ρ	σ	*Quality Update Frequency*	*Variance of* $\Delta_{Quality}$
(Top Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Top Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.2$
(*Bottom*)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Monthly	$\sigma_\xi = 0.2$



Notes: The above figures plot the mean absolute error and the mean squared error of optimal and simple average on restaurants with quality updates following a martingale process. The error on each simulated path is calculated as the difference between the averages and the true quality of the restaurant when the n^{th} review is written. The figures show that relative to arithmetic average, errors of optimal aggregation shrink faster. And the difference in this speed increases when qualities of restaurants change more often, and when the variance of incremental quality change is larger.

Figure 6: **How Quickly Do Ratings Adjust to Changes in Restaurant Quality?**

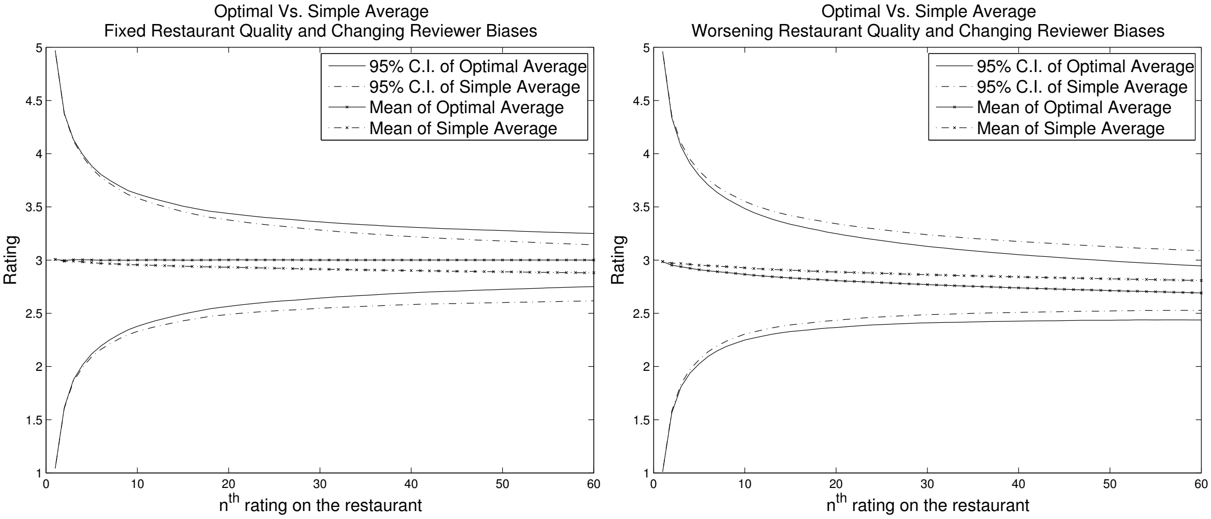
ρ	σ	*Quality Update Frequency*	*Variance of* $\Delta_{Quality}$
$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figures plot the simulated mean and 95% confidence interval for the average ratings that would occur for a single restaurant with its quality realized from the quarterly quality change martingale model. This quality path features restaurant's quality dropping before it receives its 20^{th} rating, and jumping before it receives its 40^{th} rating. Relative to arithmetic average, the optimal aggregation adapts to the change in restaurant's true quality more quickly. Since the optimal aggregation algorithm only gives weights to recent ratings, when fixing the review frequency, its standard error shrinks slower than arithmetic average that gives equal weights to all historical ratings.

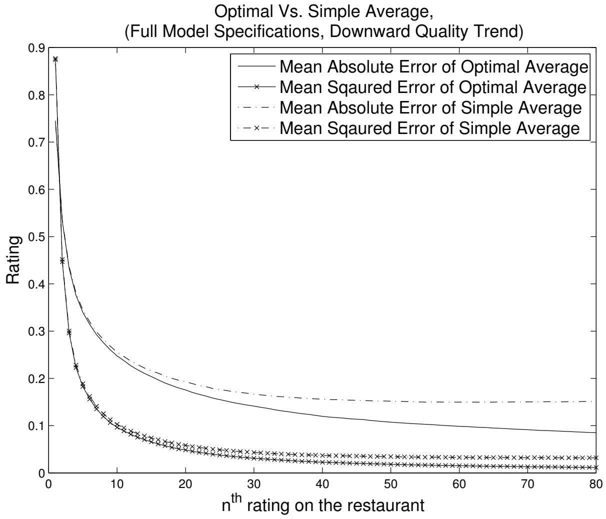
Figure 7: Simulated Ratings when Reviewers are Biased

Parameters	Value	Parameters	Value	Parameters	Value
$\rho_e,\ \rho_{ne}$	0	$\frac{\partial(\mu+\lambda_e)}{\partial Restaurat\ Age}$	-0.0002	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Review\ \#}$	-0.0007
$\sigma_e,\ \sigma_{ne}$	1	$\frac{\partial(\mu+\lambda_e)}{\partial Match\ Distance}$	0.0042	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Restaurat\ Age}$	0.00004
$Quality_0$	3	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Taste\ To\ Variety}$	-0.0224	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Reiviewer\ Frequency}$	0.0647
		$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Frequency}$	-0.0348	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Reviewer\ Review\ \#}$	0.0006



Notes: The above figures plot the simulated mean and 95% confidence interval for the average ratings that would occur for restaurants with biased reviewers. The figure on the left assumes that restaurants have fixed quality at 3, and reviewers' bias is trending downwards with restaurant age. The figure on the right assumes that the restaurants have quality trending downwards with restaurant age, and the reviewer bias is unaffected by restaurant age. In both cases, we assume that reviewers perfectly acknowledge other reviewers' biases and the common restaurant quality trend. So in both cases, optimal aggregation is an unbiased estimate for true quality while the arithmetic average is biased without correcting the review bias.

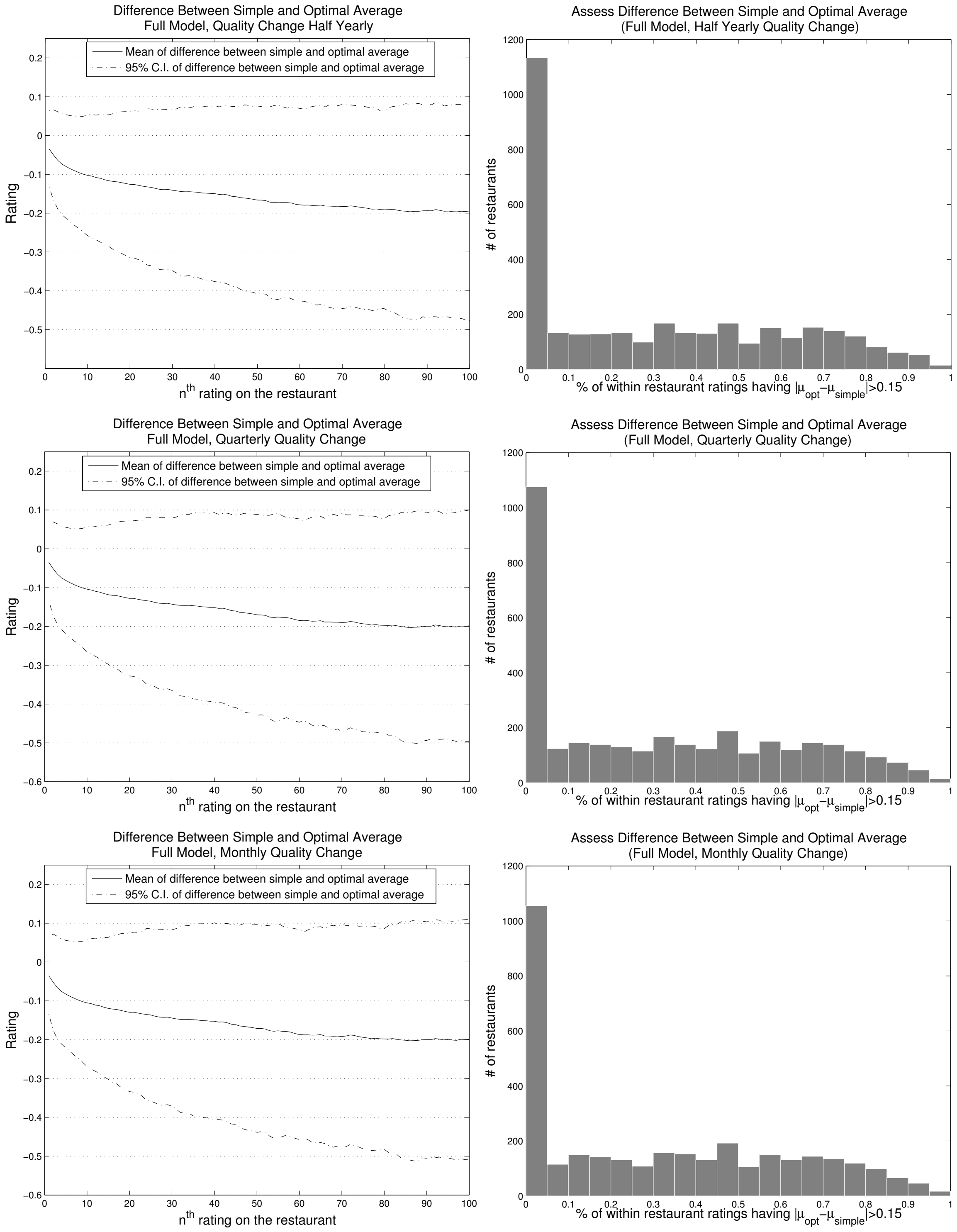
Figure 8: **Comparing Optimal and Simple Averages: Full Model Specification**



Notes: The above figure plots the mean absolute error and the mean squared error of optimal and simple average on restaurants with ratings generated from the full model with estimated parameters shown in Table 7. The error on each simulated path is calculated as the difference between the averages and the true quality of the restaurant when the n^{th} review is written. The figure shows that relative to arithmetic average, errors of optimal aggregation shrink faster. And this difference is the same if we assume the downward rating trend is due to reducing quality or biased consumers.

Figure 9: **Optimal and Simple Average Algorithm Applied on the Real Data**

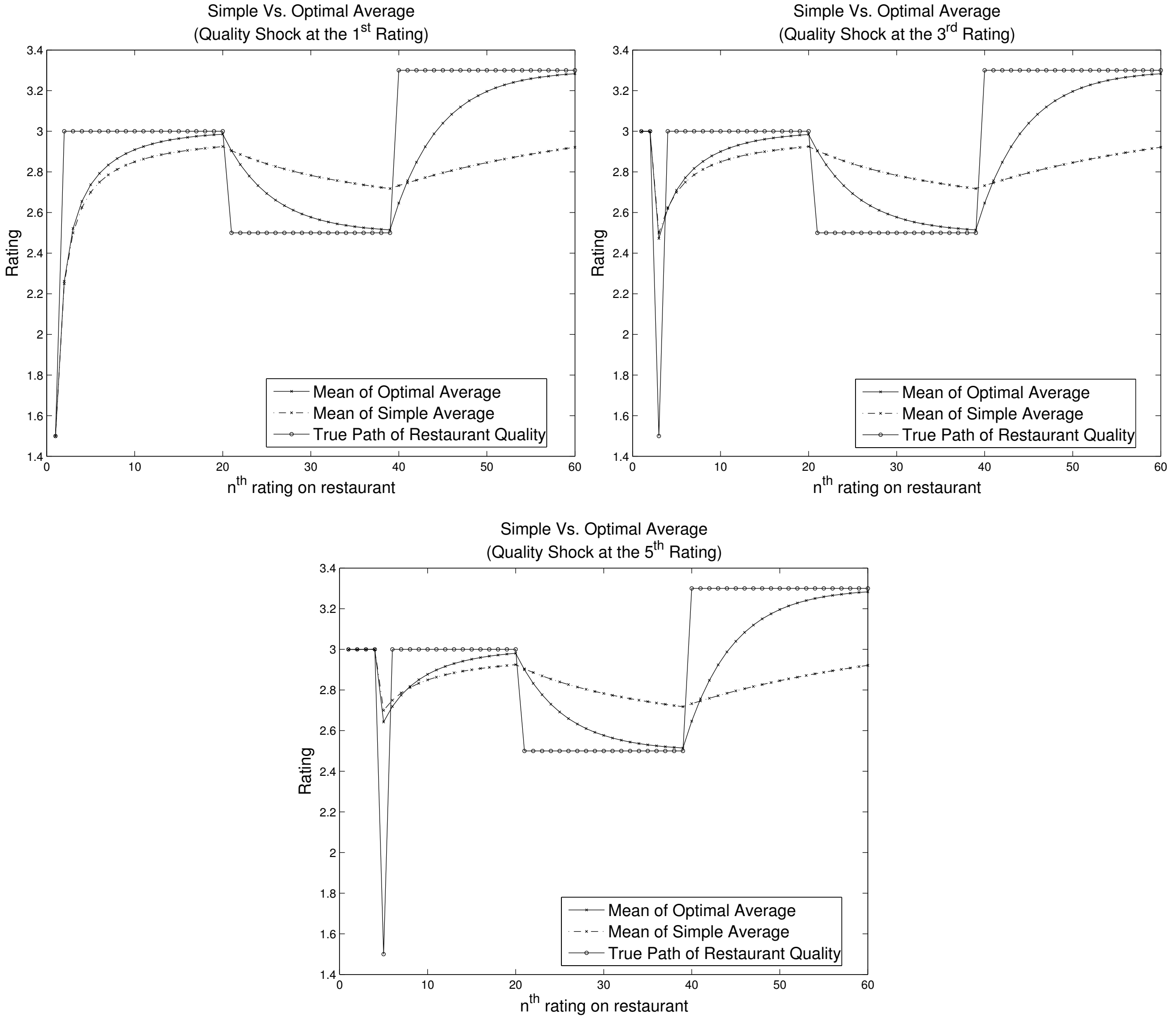
(Top)	*(Middle)*	*(Bottom)*
Quarterly Quality Change	Half-yearly Quality Change	Monthly Quality Change



Notes: 1. Figures on the left column plot the trend of mean and 95% confidence interval for $\mu_{rn}^{simple} - \mu_{rn}^{optimal}$. Figures on the right column plot the frequency of restaurants that have proportions of ratings satisfying $|\mu_{rn}^{optimal} - \mu_{rn}^{simple}| > 0.15$. 2. The model we present here assumes that there exists a reviewer "chilling" effect instead of downward trend of restaurant quality.

Figure 10: **Comparing Optimal and Simple Averages: Temporary Shocks to the Quality**

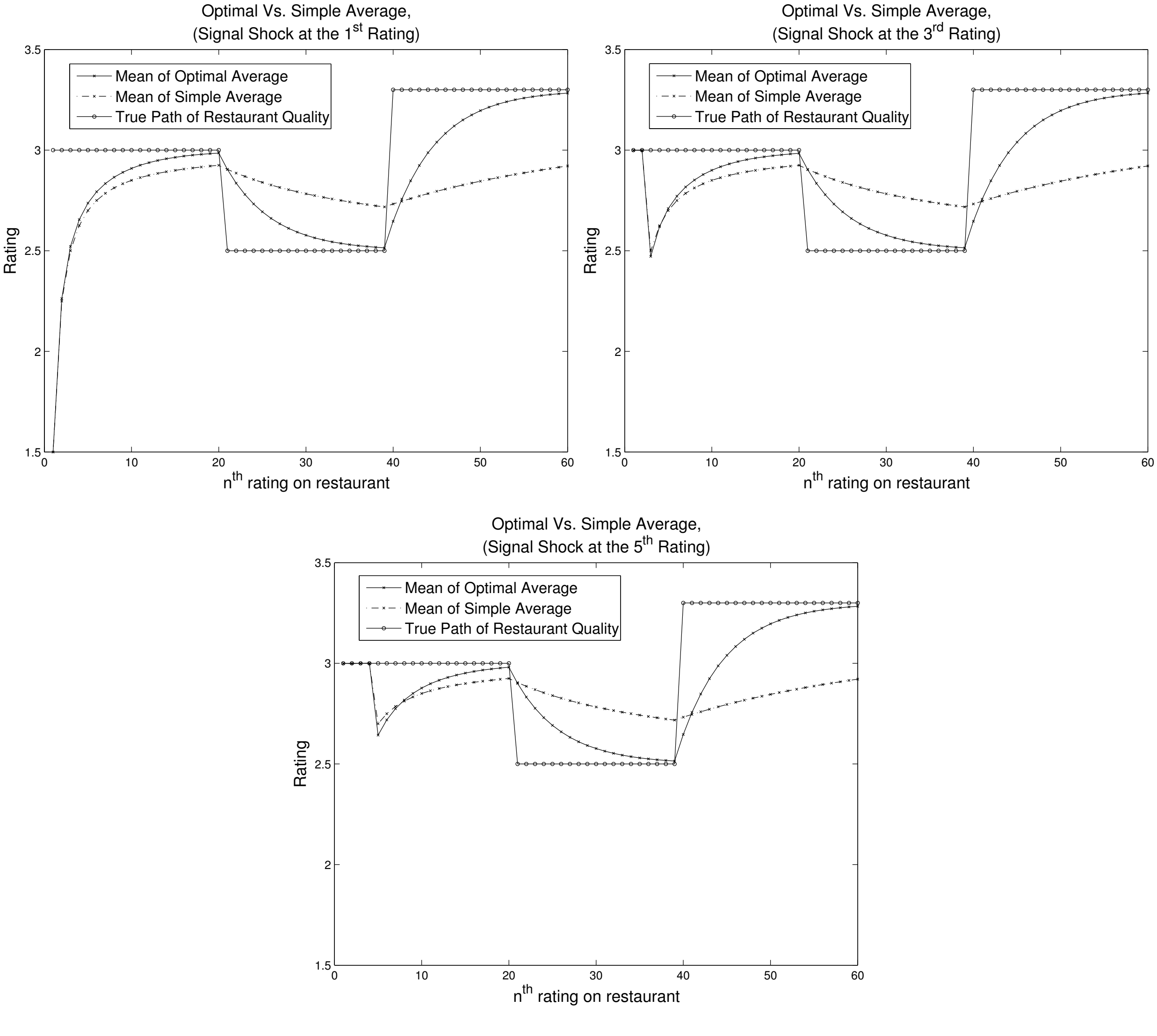
	Quality Shock	ρ	σ	*Quality Update Frequency*	*Update Noise*
(Left)	$\mu_1 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Right)	$\mu_3 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Bottom)	$\mu_5 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figures plot the simulated mean of the average ratings for a single restaurant that follows the martingale model of quarterly quality change, but experiences a temporary quality shock. Focusing on the temporary quality shock, the optimal aggregation is more responsive to temporary quality shock and converges back to the true quality in a faster rate. This is due to the fact that optimal aggregation gives higher weight to more recent reviews. In comparison, simple average smooths the shock by giving equal weight to all historical ratings.

Figure 11: **Comparing Optimal and Simple Averages: Temporary Shocks to the Signals**

	Signal Shock	ρ	σ	*Quality Update Frequency*	*Update Noise*
(Left)	$\mu_1 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Right)	$\mu_3 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Bottom)	$\mu_5 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figures plot the simulated mean of the average ratings for a single restaurant that follows the martingale model of quarterly quality change, but there is a temporary shock to the signal (We can think of this happening in the situation when a reviewer misreports her signal). Both aggregating algorithms weight past ratings, and are affected by the rating based on the shoecked signal. But compared with arithmetic mean, optimal aggregation "forgets" about earlier ratings and converges back to the true quality in a faster rate.