Optimal Aggregation of Consumer Ratings: An Application to Yelp.com *

Weijia Dai
University of Maryland

Ginger Jin
University of Maryland & NBER

Jungmin Lee
Sogang University & IZA

Michael Luca
Harvard Business School

October 2014

Abstract

Consumer review websites leverage the wisdom of the crowd, with each product being reviewed many times (some with more than 1,000 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 and accuracy, (2) reviewers to be influenced by existing reviews, and (3) product quality to change over time.

Applying this approach to restaurant reviews from Yelp.com, we construct optimal ratings for all restaurants and compare them to the arithmetic averages displayed by Yelp. Depending on how we interpret the downward trend of reviews within a restaurant, we find 19.1-41.38% of the simple average ratings are more than 0.15 stars away from optimal ratings, and 5.33-19.1% are more than 0.25 stars away at the end of our sample period. Moreover, the deviation grows significantly as a restaurant accumulates reviews over time. This suggests that large gains could be made by implementing optimal ratings, especially as Yelp grows. Our algorithm can be flexibly applied to many different review settings.

JEL: D8, L15, L86.

Keywords: crowdsourcing, social network, e-commerce, learning, Yelp.

*We are grateful to John Rust, Matthew Gentzkow, Connan Snider, Phillip Leslie, Yossi Spiegel, and participants at the 2012 UCLA Alumni Conference, the Fifth Workshop on the Economics of Advertising and Marketing, and the Yale Marketing-Industrial Organization Conference 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.

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 2013, 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.

It is easy to imagine situations in which this is not optimal. For example, consider two hypothetical restaurants. The first restaurant receives mainly 2 star ratings for its first few months and mainly 4 star ratings for the next few months. The second restaurant has the same set of ratings but in the opposite order: receiving 4s early, and then getting many 2s. Yelp would present the same average rating for these two restaurants. Yet, a careful reader would likely favor a restaurant with an upward trend. There are many other examples of situations in which one would want to make similar adjustments.

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 social concerns (some reviewers may prefer to conform to or 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 reviewer attributes 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 restaurant first appeared on Yelp, which could capture a linear or quadratic trend of reviewer stringency relative to the first review of the same restaurant, or a 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 , 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 all reviewers in the corresponding year), we find that, by the end of the sample, 41.38% of restaurants have their Yelp-style simple average ratings differ from the optimal by more than 0.15 stars, and 19.1% 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 18.95% of restaurants, and more than 0.25 stars for 5.33% of restaurants by the end of the data sample.

Aside from the "chilling" effect, 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. This also explains why the deviation of the simple average rating grows significantly as a restaurant accumulates reviews over time, no matter how we interpret the "chilling" effect. Reviewer's social concern, on the other hand, has little impact on the optimal average in the Yelp setting, even though social 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

[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.

that elite and non-elite reviewers have different signal precision and social 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 tendency to care about how Yelp readers follow and evaluate reviewers on the platform. These social 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 social concern explain much less of the optimal-vs-simple-average 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 (2012) 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 social concerns is also related to the vast literature on information cascade and observational learning (e.g. Banerjee 1992).

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]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 affects 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) argue that after the endogenous correlation among ratings, online user reviews have no significant impact on movies' box office revenues.

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 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 determine what to write in the review. 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 is what we refer to as social concerns. For example, a social-conscious reviewer may want to write a review that echoes the experience of all potential users so that she can receive favorable comments on her review, generate/satisfy followers, and maintain high status on Yelp. Because most Yelp users read but do not write reviews, the above social concern goes beyond a typical reputation game where earlier movers may strategically manipulate the behavior of later movers. Given the difficulty to model the mind of review readers (we have no data on them), we assume this social concern motivates a reviewer to be "right" about the true restaurant quality, perhaps out of a desire to contribute to a public good and with no strategic intention to influence future reviewers. In this sense the reviewer is seeking to express the truth. The way in which a reviewer incorporates previous reviews can take different directions: if everyone else rated a restaurant 4 stars, some reviewers may

[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.

downplay their own signal of 2 stars and conform to the crowd, while other reviewers may want to report 1 star in order to emphasize the difference and pull Yelp's reported average closer to their own experience. Whether conforming or differentiating, social concern implies some weight on prior reviews.

In particular, if reviewer i is motivated to best guess the true restaurant quality, we can model her choosing review x_{rt_n} in order to minimize an objective function:

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

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 $0 \leq \rho_i \leq 1$ is the weight that i puts on the importance of being "right" about the true quality in her report. The optimal review to minimize $F_{rn}^{(1)}$ is:

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

where $\lambda_{rn} = (1-\rho_i)\theta_{rn}$ represents the stringency or bias of reviewer i for restaurant r. The more reviewer i cares about being "right" about the true quality, the more positive is ρ_i.

Alternatively, if social concerns motivate reviewer i to deviate from prior reviews, we can model it as reviewer i choosing x_{rt_n} to minimize another objective:

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

where $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{n-1}})$ is the posterior belief of true quality given all the prior reviews (not counting i's own signal) and $w_i > 0$ is the marginal utility that reviewer i will get by reporting differently from prior reviews. By Bayes' Rule, $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{\{n-1\}}}, s_{rt_n})$ is a weighted average of $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{n-1}})$ and i's own signal s_{rt_n}, which we can write as, $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{\{n-1\}}}, s_{rt_n}) = \alpha \cdot s_{rt_n} + (1-\alpha) \cdot E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{n-1}})$. Combining this with the first order condition of $F_{rn}^{(2)}$, we have

$$\begin{aligned} x_{rt_n}^{(2)} &= \tfrac{1}{(1-w_i)} \quad \theta_{rn} + \frac{1-\alpha+w_i\alpha}{(1-w_i)(1-\alpha)} s_{rt_n} - \frac{w_i}{(1-w_i)(1-\alpha)} E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ..x_{rt_{n-1}}, s_{rt_n}) \\ &= \quad \lambda_{rn} + (1-\rho_i)s_{rt_n} + \rho_i E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ..x_{rt_{n-1}}, s_{rt_n}) \end{aligned}$$

if we redefine $\lambda_{rn} = \frac{1}{1-w_i}\theta_{rn}$ and $\rho_i = -\frac{w_i}{(1-w_i)(1-\alpha)}$. Note that the optimal reviews in the above two scenarios are written in exactly the same expression except that $\rho_i > 0$ if one puts some weights on best guessing the true restaurant quality in her report and $\rho_i < 0$ if one is motivated to deviate from prior reviews. The empirical estimate of ρ_i will inform us which scenario is more consistent with the data. In short, weight ρ_i is 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 social 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,

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

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:

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

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, social concern, and signal precision. Optimal information aggregation - in our definition - would correct for these differences. We rely on observable characteristics to capture reviewer heterogeneity, namely 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 social status 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, social concerns 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 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 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 and concave over time when we estimate it in quadratic terms – 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 of 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:

$$
\begin{aligned}
\rho_i &= Elite_i \cdot \rho_e + (1 - Elite_i) \cdot \rho_{ne} \\
\sigma_i^2 &= Elite_i \cdot \sigma_e^2 + (1 - Elite_i) \cdot \sigma_{ne}^2 \\
\lambda_{ri} &= Year_t \cdot \alpha_{yeart} + Age_{rt} \cdot \alpha_{age1} + Age_{rt}^2 \cdot \alpha_{age2} + NumRev_{it} \cdot \alpha_{numrev} + FreqRev_{it} \cdot \alpha_{freqrev} \\
&\quad + MatchD_{rit} \cdot \alpha_{matchd} + TasteVar_{it} \cdot \alpha_{tastevar} \\
&\quad + Elite_i \cdot [\lambda_{(e-ne)0} + Age_{rt} \cdot \beta_{age1} + Age_{rt}^2 \cdot \beta_{age2} + NumRev_{it} \cdot \beta_{numrev} + FreqRev_{it} \cdot \beta_{freqrev} \\
&\quad + MatchD_{rit} \cdot \beta_{matchd} + TasteVar_{it} \cdot \beta_{tastevar}]
\end{aligned}
$$

where $\{\rho_e, \rho_{ne}\}$ capture the social concern of elite and non-elite reviewers, $\{\sigma_e^2, \sigma_{ne}^2\}$ capture the signal precision of elite and non-elite reviewers, $\{\alpha_{age1}, \alpha_{age2}\}$ capture the catch-all trend in quality or stringency change,[5] $\{\alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}\}$ capture how restaurant and reviewer attributes change the stringency of non-elite reviewers, and $\{\lambda_{(e-ne)0}, \beta_{age1}, \beta_{age2}, \beta_{numrev}, \beta_{freqrev},$

[5] We define the raw age by calendar days since a restaurant's first review on Yelp and normalize the age variable in our estimation by (raw age-548)/10. We choose to normalize age relative to the 548th day because the downward trend of reviews is steeper in a restaurant's early reviews and flattens at roughly 1.5 years after the first review.

$\beta_{matchd}, \beta_{tastevar}\}$ capture how restaurant and reviewer attributes change the stringency difference between elite and non-elite reviewers. We include year fixed effects $\{\alpha_{yeart}\}$ in λ_{ri} to capture the possibility that reviewer stringency may vary by calendar year due to taste change in the general population.

As detailed below, our algorithm will incorporate restaurant fixed effects, which capture the initial restaurant quality at the time of the first review. This is why we do not include any time-invariant restaurant attributes in λ_{ri}. That being said, it is perceivable that different types of restaurants may differ in our key parameters regarding the precision of restaurant signals and the evolution of restaurant quality. To allow this possibility, one version of the model will estimate $\{\sigma_e^2, \sigma_{ne}^2, \alpha_{age1}, \alpha_{age2}, \sigma_\xi^2\}$ separately for ethnic and non-ethnic restaurants, where a restaurant is defined as "ethnic" if it offers cuisine from a specific country other than the US, according to Yelp classification. We have tried to allow ρ_i to vary by restaurant ethnicity 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, social concern, and signal precision, and a quadratic 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 gives weights to both her own experience and her social concerns. 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 de-bias the stringency of past reviewers. The reviewer also knows that the general

population of reviewers may change taste from year to year (captured in year fixed effects $\{\alpha_{yeart}\}$), and there is a quadratic trend in λ by restaurant age (captured in $\{\alpha_{age1}, \alpha_{age2}\}$). This trend could be driven by changes in reviewer stringency or restaurant quality and these two drivers are not distinguishable 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), social 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 a reviewer's posterior about restaurant quality is a weighted average of previous signals and her own signal, with the weight increasing with signal precision. This is complicated by the fact that restaurant quality evolves by a martingale process, 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, social 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 social concerns (ρ_e, ρ_{ne}), parameters affecting reviewer stringency ($\alpha_{yeart}, \alpha_{numrev}, \alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}$, $\lambda_{(e-ne)0}$, $\beta_{age1}, \beta_{age2}, \beta_{numrev}, \beta_{freqrev}, \beta_{matchd}, \beta_{tastevar}$), and the parameters for the catch-all quadratic trend by restaurant age on Yelp ($\alpha_{age1}, \alpha_{age2}$). In an extended model, we also allow $\{\sigma_e, \sigma_{ne}, \alpha_{age1}, \alpha_{age2}, \sigma_\xi\}$ to differ for ethnic and non-ethnic restaurants.

Note that consistent estimation of all other parameters depends on the consistency of $\{\mu_{r0}\}_{r=1}^{R}$, which requires that 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 social 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 social concerns ($\rho_n = 0$). When $\rho_n > 0$, $Var(x_{rn})$ declines with n. If the n^{th} reviewer cares about social concerns only ($\rho_n = 1$), we have the familiar form of $Var(x_{rn}) = \frac{\sigma_\epsilon^2}{n}$. In other words, the magnitude of a positive ρ determines the degree to which the variance of reviews shrinks over time, while σ_ϵ^2 determines the variance of the first review. When $\rho_n < 0$, $Var(x_{rn})$ increases with n. Thus the overtime variation of review variance can indicate the sign of social concerns, if other factors are not present.

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, social 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 social 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 puts 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 social 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 $\{\alpha_{yeart}, \alpha_{age1}, \alpha_{age2}, \alpha_{numrev}, \alpha_{freqrev}, \alpha_{matchd}, \alpha_{tastevar}\}$, $\{\beta_{day}, \beta_{numrev}, \beta_{freqrev}, \beta_{matchd}, \beta_{tastevar}\}$, 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 quadratic trend of ratings ($Age_{rt} \cdot \alpha_{age1} + Age_{rt}^2 \cdot \alpha_{age2}$ in λ_{rn}) as reviewer bias[6], 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 quadratic trend of ratings as changes in true quality, the optimal estimate of quality at t_n is $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, .., x_{rt_n}) + Age_{rt} \cdot \alpha_{age1} + Age_{rt}^2 \cdot \alpha_{age2}$. 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.[7] 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 at the data download time. 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

[6] This is relative to the review submitted 1.5 years after the first review, because age is normalized by (raw age - 548)/10.

[7] 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.

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 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 provides 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. Between the first and the last review dates of an average restaurant, the restaurant receives 0.16 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 11th and 12th 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 Variations 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 account for 23.3% of the total variations 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; its variations alone explain 20.86% of total variations 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 variations 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.

To check the difference between elite and non-elite reviewers on the same restaurant, we first obtain residual $\widehat{\epsilon_{ri,yr}}$ after regressing observed ratings 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 whether the review is written by an elite reviewer and the order of a review (N_{ri}) (i.e. $\widehat{\epsilon_{ri,yr}}^2 = \beta_0 + \beta_1 D_{ri,elite} + \beta_2 N_{ri} + \beta_3 N_{ri} \times D_{ri,elite} + \zeta_{ri}$). Results reported in Table 3 show 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 have more social motives to conform to the crowd 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 social concerns.

3.3 Dynamics of the Review Process

We now present reduce-form evidence for review dynamics, which will be used to inform our structural model.

As detailed in Section 2, identification of our model relies on the extent to which the variance and covariance of reviews change over time within a restaurant. If the true restaurant quality is a constant and reviewers tend to incorporate a restaurant's previous reviews in a positive way ($\rho > 0$), we should observe reviews to vary less and less over time around the restaurant's fixed effect, which is confirmed by the significantly negative coefficient on the order of review in Table 3.

Positive social concerns also imply positive serial correlation of ratings within a restaurant and such correlation should be stronger for close-by reviews. 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. That being said, positive social concern is not the only explanation for this pattern of serial correlations: a martingale evolution of restaurant quality could generate it as well.[8] It is up to the structural model to separate the effect of positive social concerns and quality evolution.

Furthermore, 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. More than one factor could contribute to this downward trend. For example, restaurant quality may decline over time. Alternatively, 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.

The first two columns of Table 5 regress reviewer-restaurant matching distance ($MatchD_{rit}$) and reviewer's taste for variety ($TasteVar_{rit}$) on the order of reviews within a restaurant. They show that, within a restaurant, later reviewers tend to have less diverse tastes but are not significantly different from earlier reviewers in matching distance. This suggests that later reviewers may be better sorted with the restaurant in terms of taste diversity, which does not explain why later reviewers tend to give worse ratings on the same restaurant unless less diverse diners are more critical. The last two columns of Table 5 examine variations of $MatchD_{rit}$ and $MatchD_{rit}$ 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.

Overall, reduce-form results yield five empirical observations related to review dynamics: ratings are less variable over time, ratings trend downward within a restaurant, ratings are serially correlated within a restaurant, restaurants tend to find reviewers with less diverse taste over time, and reviewers tend to find better matches of restaurants over time.

[8] Note that the martingale evolution of restaurant quality implies an increasing variance around the restaurant's fixed effect, while positive social concerns implies a decreasing variance.

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 the estimation results of our baseline structural model 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, social weight, and stringency. Note that social 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 optimal aggregation of all reviews.

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

4.1 Elite Reviewers

In Table 6 Column (2), we allow signal precision, social weight, and stringency to differ by reviewer's elite status. The estimates, as well as a likelihood ratio test between Columns (1) and (2), clearly suggest that elite and non-elite reviewers differ in both signal precision and social 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 σ_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

Table 6 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. One potential explanation is that non-elite reviewers deliberately give out an opinion that is different from previous reviews on the same restaurant because they either enjoy expressing a different

opinion from other reviewers or believe differentiation is the best way for them to contribute to the public good. Another possibility is that non-elite diners are more likely to leave a review on Yelp if their own experience is significantly different from the expectation they have had from reading previous reviews. With little information on the diners that choose not to leave any review on Yelp, it is difficult to distinguish these possibilities in the real data. Compared to the non-elite, elite reviewers are found to be more positively influenced by the past crowd, probably because their social concerns motivate them to be closer to the other reviewers. Although the quarterly noise in restaurant quality ($\sigma_\xi = 0.1452$) is estimated at 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 (1) the restaurant's tenure on Yelp (Age_{rt} and Age_{rt}^2), 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 who have written more reviews on Yelp tend to match better with a restaurant and have more diverse tastes. In comparison, an elite reviewer behaves similarly 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 uses the same model as in Table 6 Column 4, but allows restaurant quality to evolve 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' social 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 social preference because there are enormous variations in how closely sequential reviews arrive. Clearly, the more frequently we allow restaurant quality to vary, the smaller σ_ξ is (because it captures quality change in a smaller calendar window). By doing this, more of the variation in ratings is attributed to quality change, rather than simply noise in a reviewer's signal. 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.

Table 8 incorporates more restaurant heterogeneity in the specification of Table 6 Column 4. In particular, we allow ethnic and non-ethnic restaurants to differ in the noise of reviewer signal (σ_e, σ_{ne}), the noise of quality evolution (σ_ξ), and the influence of restaurant tenure (Age_{rt} and Age_{rt}^2). As reported in Table 8, the estimates of σ_e, σ_{ne}, σ_ξ and the coefficients on Age_{rt} and Age_{rt}^2 are indeed different between ethnic and non-ethnic restaurants. In light of these results, all of our real-data-based counterfactual simulations use estimates from the Table 8 specification, which assumes quarterly evolution of restaurant quality and allows heterogeneity between ethnic and non-ethnic restaurants.

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.[9] 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.[10] 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 to minimize the information loss.

[9] 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.

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

This suggests that the Bayesian model is more suitable for our data.

5 Counterfactual Simulations

This section presents two 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).

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 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 social weight or bias. To start, we take this condition as the benchmark. In order to highlight how much optimal average is superior to 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 a random variable in this model, we compare the mean absolute 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 non-zero weight on previous reviews. When social concern is the only deviation from the assumption that reviews are i.i.d., then the arithmetic average is unbiased but inefficient. If later reviews have already put positive weight on 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 socially concerned. However, the right graph suggests that efficiency gain over simple average is small even if the social 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. Similar logic applies if later reviews put negative weight on past reviews. In that case, optimal weighting should give less weight to early reviews. The simulated figure that compares optimal average with simple average is very similar to Figure 3, with optimal average being more efficient.

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, as the goal of average rating is to reflect the "current" quality of the restaurant at the date of last review. We present three graphs in Figure 5: the first two allow the variance to change quarterly 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 as one review per month. 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.

Moreover, as a restaurant accumulates Yelp reviews over time, the mean absolute error of the optimal rating becomes stabilized around 0.2 to 0.4, while the mean absolute error of the simple average keeps increasing over time, and could be as high as 1 after 60 reviews in the first graph of Figure 5. This is because the average rating is meant to the capture the "current" restaurant quality at the time of computing the average. The optimal rating does this well by giving more weights to recent reviews. In contrast, the simple average rating gives the same weight to every review; when there are N reviews, the weight to the most recent review $(1/N)$ actually decreases with N, which explains why the simple average rating is further away from the true "current" quality as N increases.

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 Table 8. 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. 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.[11]

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 8, the noise of quality update (σ_ξ) has a standard deviation of 0.1212 per quarter for ethnic restaurants and 0.1346 for non-ethnic restaurants, which amount to an average deviation of 0.49-0.54 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 0.92 and 0.98.

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$0.06) for both ethnic and non-ethnic restaurants, 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 social concerns, but the absolute magnitudes of ρ_e and ρ_{ne} never exceed 0.2, suggesting that the efficiency gain of optimal average due to social concerns is likely to be small as well.

Including all these elements, we compute simple and optimal average ratings at the time of every observed rating. This calculation is done for quarterly quality updates, according to the structural estimates in Table 8. We then calculate the difference between simple and optimal average, $\mu_{rn}^{simple} - \mu_{rn}^{optimal}$ for every observation and summarize it in Table 9.

If we interpret the coefficients on restaurant age as a change of reviewer stringency, the 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} and Age_{rt}^2, the second reviewer (if non-elite) will give a rating 0.13 stars higher for a non-ethnic restaurant and 0.14 stars higher for an ethnic restaurant, relative to the review coming

[11] 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 errors of the two aggregating methods. Optimal average always corrects any bias in reviewer bias, and simple average always reflects the sum of the changes in quality and reviewer bias.

1.5 years after the first review. In contrast, a review submitted six years from the first review of the restaurant will be -0.38 lower on a non-ethnic restaurant and -0.48 lower for an ethnic restaurant. Overall, we find that optimal and simple averages differ by at least 0.15 stars in 33.63% of observations, and differ by at least 0.25 stars in 13.38% of observations. If we round the two ratings before comparison, their difference is at least 0.5 stars for 25.39% of the observations. Interestingly, the deviation from simple average to optimal average is asymmetric: simple average is more likely to underreport than overreport, as compared to optimal average. We believe this is because optimal average puts more weight on late reviews, and late reviews entail a greater correction of bias than early reviews due to the chilling effect.

Alternatively, if we interpret the coefficients on restaurant age as a change of true restaurant quality, the two averages differ by at least 0.15 stars in 13.6% of observations, and are different by at least 0.25 stars in 2.91% of observations. If we round the two ratings before comparison, they are at least 0.5 stars apart for 14.44% of the observations. The asymmetry on the direction of deviation between the two averages also changes: simple average tends to overreport, as compared to optimal average when we interpret the restaurant age effect as true quality declining over time.

The remainder of Table 9 compares our optimal rating to 6-month, 12-month and 18-month moving average of reviewer ratings. No matter how we interpret the downward trend, these moving averages are even worse than simple averages, probably because many restaurants receive sparse reviews and the short window of moving averages excludes many reviews that could be useful for the aggregation.

Table 10 describes how the difference between simple and optimal averages vary over time. The first panel compares the two average ratings at each restaurant's last review in our sample. As before, the rating difference depends on our interpretation of the "chilling" effect. If this chilling effect is interpreted as reviewer bias only, we find that, by the end of the sample, 41.38% of restaurants have their Yelp-style simple average ratings differ from the optimal by more than 0.15 stars, and 19.1% 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 18.95% of restaurants, and more than 0.25 stars for 5.33% of restaurants by the end of the data sample.

Why are these numbers bigger than what we have presented in Table 9? This is because Table 9 summarizes the rating difference for all reviews of a restaurant rather than the last review in our sample. To see this more clearly, the next three panels of Table 10 calculate the rating difference for reviews 0-2 years, 2-4 years and >4 years since the first review of a restaurant. No matter how we interpret the chilling effect, the difference between simple and optimal ratings grows rapidly as a restaurant accumulates more reviews over time. As illustrated in Figure 5, when restaurant quality changes over time, it is important to adjust

weights towards recent reviews in order for an average rating to reflect the "current" restaurant quality. This factor is incorporated in our optimal rating but missing in the simple average rating.

The increasing divergence of simple vs. optimal ratings can be better shown in graph. Based on the above-estimated difference between simple and optimal average per observation, Figure 8 plots the mean and confidence interval of this difference by the order of review. Assuming the restaurant age effect as reviewers become more stringent over time (i.e. the chilling effect), the upper-left graph of Figure 8 shows that 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 upper right of Figure 8 plots the histogram of restaurants by this percent. For example, the second bar shows that roughly 200 restaurants (out of the total 3,345) have 5-10% of times with simple average ratings more than 0.15 stars away from the optimal. Overall, over 1,271 restaurants have simple average ratings more than 0.15 stars away from the optimal at least 30% of the time. This suggests that optimal average rating is likely to generate substantial improvement over simple average, especially as Yelp accumulates more and more reviews for each restaurant. The bottom two graphs of Figure 8 lead to a similar conclusion but of smaller magnitude when we interpret the restaurant age effect as true quality changes.

In Table 11, we summarize restaurant and reviewer attributes by whether the simple-to-optimal difference is <-0.15, -0.15~0.15, or >0.25 stars. Restaurant review frequency, reviewer review frequency and matching distance seem to differ most across the three groups, and how they differ is sensitive to our interpretation of the downward trend.

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 social concerns (ρ), reviewer's signal noise (σ_ϵ), or other terms in reviewer stringency (λ_{rt}). Because of the importance of restaurant quality updates (σ_ξ), the simple average is further away from the optimal average as each restaurant accumulates reviews over time.

6 Conclusion

As consumer reviews continue to proliferate, offering unprecedented amounts of information, this paper demonstrates 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.

We show that optimally aggregated information deviates significantly from arithmetic averages for a non-trivial fraction of restaurants. By law of large numbers, one might hope that a greater number of reviews will lessen the probelms of simple averaging of ratings over time. Yet this intuition can often turn out to be wrong - for example, if quality changes over time (a new chef, a different menu, etc). By moving toward an optimal aggregation approach, the market designer can detect such changes and deviations between true quality and the arithmetrical average.

In this section, we discuss the limits of our approach 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.

The information systems literature has documented bimodal distributions of reviews (with many one and five stars) 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 that Hu et al provide 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 trends and serial correlation of Yelp reviews within a restaurant, both of which could be byproducts of reviewer selection.

Two concurrent papers are currently investigating the selection process. Chan et al. (2010) use 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 connections 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) examine the determinants of 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 ratings. 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 reviews 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' social concern on Yelp and allow elite and non-elite reviewers to place different weight on social concerns. That being said, social concerns in our model can have multiple interpretations and we do not model one's incentive to manipulate reviews for popularity.[12]

6.3 Fake reviews

One potential problem for consumer review websites like Yelp is fake or promotional reviews. Mayzlin, Dover and Chevalier (forthcoming) have documented evidence for review manipulation on hotel booking platforms (Expedia.com and Tripadvisor.com). To minimize the presence of potentially non-authentic reviews, Yelp imposes a filter on all submitted reviews and only posts reviews that Yelp believes to be authentic or trustworthy. Accordingly, our data do not include the reviews that Yelp has filtered out. For an analysis of filtered reviews, see Luca and Zervas (working paper, 2013). While review filters can help to eliminate gaming, there are surely still erratic and fake reviews that get through the system. In Figure 9, we simulate the evolution of ratings in a situation in which an outlier review is posted. Figure 9 simulates two situations where an extremely low rating (1.5) occurs as either the first or the fifth review of a restaurant, while the true restaurant quality starts at 3 stars, jumps down to 2.5 stars at the time of the 20th review, and reverts back to 3.5 stars at the time of the 40th

[12] 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.

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 9 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 graph of Figure 9 suggests 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.

6.4 Transparency and Aggregation Decisions

Part of the motivation for this paper is that on 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). Even beyond literal aggregation of reviews, our method can inform the optimal ordering of businesses on a platform such as Yelp.

6.5 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 preferences. 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.

6.6 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 argue 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 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 easily incorporate text analysis methods into optimal information aggregation.

6.7 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 may not be an issue for fixed products such as books, movies, etc.; whereas reviewer heterogeneity may be much more important. 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

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.

Alevy, Jonathan E., Michael S. Haigh and John A. List (2007). "Information cascades: Evidence from a field experiment with financial market professionals." *The Journal of Finance,* 62(1): 151-180.

Archak, Nikolay, Anindya Ghose and Panagiotis G. Ipeirotis (2011). "Deriving the pricing power of product features by mining consumer reviews." *Management Science,* 57(8): 1485-1509.

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. (1992). "A simple model of herd behavior." *The Quarterly Journal of Economics,* 107(3): 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."

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.

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

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 (2008). "Do online reviews matter?—An empirical investigation of panel data." *Decision Support Systems,* 45(4): 1007-1016.

Ghose, Anindya and Panagiotis G. Ipeirotis (2011). "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.

Ghose, A., P. Ipeirotis, B. Li (2012). "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 (2009). "Firm-created word-of-mouth communication: Evidence from a field test." *Marketing Science*, 28(4):721–739.

Goldenberg, Jacob, Barak Libai and Eitan Muller (2010). "The chilling effects of network externalities." *International Journal of Research in Marketing,* 27(1): 4-15.

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

Hu, N.; Liu, L. and Zhang, J. (2008). "Do Online Reviews Affect Product Sales? The Role of Reviewer Characteristics and Temporal Effects." *Information Technology Management* 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 (2011). "Reviews, Reputation, and Revenue: The Case of Yelp.com." *Harvard Business School working paper*.

Luca, Michael and Jonathan Smith (2013). "Salience in Quality Disclosure: Evidence from The US News College Rankings." *Journal of Economics & Management Strategy*.

Luca, Michael and Georgios Zervas (2013). "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." Working Paper.

Mayzlin, Dina (2006). "Promotional chat on the Internet." *Marketing Science,* 25(2): 155-163.

Mayzlin, Dina, Y. Dover and Judy A. Chevalier (forthcoming). "Promotional Reviews: An Empirical Investigation of Online Review Manipulation." *American Economic Review.*

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*, 28(6): 1154-1165.

Wang, Qingliang, Khim Yong Goh and 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 (2010). "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*.

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,

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

$E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n}) = \mu_{n|n}$ 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:

$$
\begin{aligned}
x_1 &= s_1 + \lambda_1 \\
s_1 &= x_1 - \lambda_1 \\
x_2 &= \rho_2 \frac{\sigma^2_2}{\sigma^2_{2|1} + \sigma^2_2} \mu_{2|1} + \rho_2 \frac{\sigma^2_{2|1}}{\sigma^2_{2|1} + \sigma^2_2} s_2 + (1 - \rho_2) s_2 + \lambda_2 \\
&= \rho_2 \frac{\sigma^2_2}{\sigma^2_{2|1} + \sigma^2_2} \mu_{2|1} + [1 - (1 - \frac{\sigma^2_{2|1}}{\sigma^2_{2|1} + \sigma^2_2})\rho_2] s_2 + \lambda_2 \\
s_2 &= \frac{1}{[1 - (1 - \frac{\sigma^2_{2|1}}{\sigma^2_{2|1} + \sigma^2_2})\rho_2]} [x_2 - \lambda_2 - \rho_2 \frac{\sigma^2_2}{\sigma^2_{2|1} + \sigma^2_2} \mu_{2|1}] \\
&... \\
s_n &= \frac{1}{[1 - (1 - \frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1} + \sigma^2_n})\rho_n]} [x_n - \lambda_n - \rho_n \frac{\sigma^2_n}{\sigma^2_{n|n-1} + \sigma^2_n} \mu_{n|n-1}].
\end{aligned}
$$

Tables

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] Our sample includes 134,730 reviews written on 4,101 restaurants in Seattle. They are written by 18,778 unique reviewers.

[b] The reviewer-restaurant matching distance variable measures the match quality between a reviewer and a restaurant. It is calculated as the Euclidean distance between characteristics of a particular restaurant and the mean characteristics of all restaurants a reviewer has reviewed before.

[c] Reviewer taste for variety measures how much a reviewer enjoys restaurant variety. It is calculated as the variation in characteristics among all restaurants a reviewer has reviewed before.

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 Distance & Taste to Variety	0.3749

Notes: 1. This table presents R^2 of the linear regression in which Yelp ratings is the dependent variable, and fixed effects and matching variables indicated in each row are independent variables. 2. There are only a few observations in 2004 and 2010, so we use fixed effect of 2005 for 2004, and fixed effect of 2009 for 2010.

Table 3: **Variability of Ratings Declines 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] $\widehat{\epsilon_{ri,yr}}$ are residuals from regression $Rating_{ri,year} = \mu_r + \alpha_i + \gamma_{year} + \epsilon_{ri,year}$

[b] D_{ri}^{elite} equals to one if reviewer i is an elite reviewer.

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

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

Table 4: **Examine Serial Correlation in Restaurant Ratings**

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)
$\widehat{\epsilon_{r,i-6,yr}}$			0.0099***	0.0100**
			(0.0030)	(0.0030)
$\widehat{\epsilon_{r,i-7,yr}}$				0.0031
				(0.0030)
Constant	-0.0063*	-0.0078**	-0.0086**	-0.0097***
	(0.0027)	(0.0027)	(0.0027)	(0.0028)
Observations	117,536	114,742	112,067	109,505

Notes: This table estimates the degree of serial correlations 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 residuals, we regress 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	0.0017	-0.0003***		
	(0.0012)	(0.0001)		
(Restaurant's n^{th} Review)^2	-2e-5	1.17e-6***		
	(1e-5)	(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			-0.0670***	0.0017***
			(0.0014)	(0.0001)
(Reviewer's n^{th} Review)^2			0.0005***	-1e-5***
			(1e-5)	(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)	(0.0233)	(0.0010)
Observations	134,730	103,835	103,835	103,835

Notes: The sample sizes of regressions specified in columns (2)-(4) are smaller since we dropped the first review written by a reviewer. It is dropped in columns (2) and (4) since we do not have a measure of taste for variety when a reviewer has only written one review. It is dropped in column (3) since we cannot calculate reviewer's match distance with the restaurant when a reviewer has no review history. In column (1), we assume that the match distance for a reviewer when she writes the first review is the same as the mean distance in sample.

[b] The reviewer-restaurant matching distance variable measures the match quality between a reviewer and a restaurant. It is calculated as the Euclidean distance between characteristics of a particular restaurant and the mean characteristics of all restaurants a reviewer has reviewed before.

[c] Reviewer taste for variety measures how much a reviewer enjoys restaurant variety. It is calculated as the variation in characteristics among all restaurants a reviewer has reviewed before.

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

Table 6: **MLE: Signal Precision, Popularity Concern and Quality Change**

Panel A. Common Parameters in model (1) - (4)

	(1) *Same* σ, ρ	*(2)* *Different* σ, ρ	*(3)* *Quarterly Quality Change*	*(4)* *Full w Quarterly Quality Change*
σ_e	1.2218***	1.1753***	0.9293***	0.9004***
	(0.0210)	(0.0210)	(0.0199)	(0.0193)
σ_{ne}		1.2350***	0.9850***	0.9514***
		0.0147	(0.0156)	(0.0150)
σ_ξ			0.1452***	0.1323***
			(0.0038)	(0.0038)
ρ_e	0.1718***	0.2430***	0.0454***	0.0122***
	(0.0007)	(0.0141)	(0.0215)	(0.0222)
ρ_{ne}		0.1362***	−0.0821***	−0.1221***
		(0.0110)	(0.0181)	(0.0186)
$(\lambda_e - \lambda_{ne})_0$		−0.0100	-0.0061	0.0161
		(0.0059)	(0.0059)	(0.0233)
Log Likelihood	-193,339	-192,538	-192,085	-191,770.2
N	133,688	133,688	133,688	133,688

Panel B. Bias parameters in model (4)

	(4)		*(4)*
$(\mu + \lambda_{ne})_{Age}$	−0.0032***	$(\lambda_e - \lambda_{ne})_{Age}$	−0.0002
	(0.0003)		(0.0002)
$(\mu + \lambda_{ne})_{Age^2}$	4×10^{-6}	$(\lambda_e - \lambda_{ne})_{Age^2}$	$1 \times 10^{-5***}$
	(2.5×10^{-5})		(2×10^{-6})
$(\mu + \lambda_{ne})_{MatchD}$	0.0367***	$(\lambda_e - \lambda_{ne})_{MatchD}$	−0.0028
	(0.0040)		(0.0053)
$(\mu + \lambda_{ne})_{TasteVar}$	−0.2453***	$(\lambda_e - \lambda_{ne})_{TasteVar}$	−0.0266
	(0.0354)		(0.0768)
$(\mu + \lambda_{ne})_{NumRev}$	−0.0062**	$(\lambda_e - \lambda_{ne})_{NumRev}$	0.0041***
	(0.0010)		(0.0014)
$(\mu + \lambda_{ne})_{FreqRev}$	0.0256***	$(\lambda_e - \lambda_{ne})_{FreqRev}$	−0.0556***
	(0.03997)		(0.0535)

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

Notes: 1. The columns in this table show estimates from models that gradually add review heterogeneity. Model in column (1) assumes that reviewers have common precision, popularity concern, and biases in judging restaurants' quality. Model in column (2) allows reviewers' precision, popularity concern, and bias to differ by elite status. "e" and "ne" in the subscripts indicate reviewer's elite and non-elite status respectively. Model in column (3) allows stochastic restaurant quality evolving in a random walk process. Column (4) further allows reviewer bias to depend on reviewer characteristics and her match with the restaurant. We also add a common year dummy in bias to capture time trend in ratings besides the trend relative to restaurant's own history. 2. The lower panel shows how reviewer characteristics and her match with restaurant affect her biases. 3. Since we estimate the model based on first differences in reviews, we are not able to identify true quality of the restaurants, but we can identify the effect of review characteristics on the change in review biases. We use non-elite reviewers as baseline and the estimates are shown in the left column of panel B. The elite versus non-elite relative differences in bias are shown in the right column. The subscripts are in turn age (Age), age square (Age^2) of the restaurant, the reviewer-restaurant match distance ($MatchD$), and reviewer taste for variety ($TasteVar$), number of reviews written by the reviewer per day ($FreqRev$), and total number of reviews written by the reviewer ($NumRev$). 4. Variables that influence reviewer bias are scaled down by ten.

Table 7: **MLE for Baseline Model with Changing Restaurant Quality**

	(1) *Quarterly Quality Change*	*(2)* *Half-yearly Quality Change*	*(3)* *Monthly Quality Change*
σ_e	0.9004^{***}	0.9251^{***}	0.8889^{***}
	(0.0193)	(0.0194)	(0.0194)
σ_{ne}	0.9514^{***}	0.9725^{***}	0.9400^{***}
	(0.0150)	(0.0149)	(0.0151)
σ_ξ	0.1323^{***}	0.1706^{***}	0.0795^{***}
	(0.0038)	(0.0051)	(0.0023)
ρ_e	0.0122^{***}	0.0377^{***}	-0.0007^{***}
	(0.0222)	(0.0212)	(0.0229)
ρ_{ne}	-0.1221^{***}	-0.0985^{***}	-0.1359^{***}
	(0.0186)	(0.0177)	(0.0193)
$(\mu+\lambda_{ne})_{Age}$	-0.0032^{***}	-0.0032^{***}	-0.0032^{***}
	(0.0003)	(0.0003)	(0.0003)
$(\mu+\lambda_{ne})_{Age^2}$	4×10^{-6}	4×10^{-6}	5×10^{-6}
	(2.5×10^{-5})	(2.4×10^{-6})	(2.5×10^{-6})
$(\mu+\lambda_{ne})_{MatchD}$	0.0367^{***}	0.0372^{***}	0.0367^{***}
	(0.0040)	(0.0040)	(0.0040)
$(\mu+\lambda_{ne})_{TasteVar}$	-0.2453^{***}	-0.2554^{***}	-0.2551^{***}
	(0.0354)	(0.0355)	(0.0354)
$(\mu+\lambda_{ne})_{NumRev}$	-0.0062^{**}	-0.0060^{**}	-0.0061^{**}
	(0.0010)	(0.0010)	(0.0010)
$(\mu+\lambda_{ne})_{FreqRev}$	0.0256^{***}	0.0244^{***}	0.0237^{***}
	(0.03997)	(0.0397)	(0.0396)
$(\lambda_e-\lambda_{ne})_0$	0.0161	0.0157	0.0161
	(0.0233)	(0.0233)	(0.0233)
$(\lambda_e-\lambda_{ne})_{Age}$	-0.0002	-0.0003	-0.0003
	(0.0002)	(0.0001)	(0.0002)
$(\lambda_e-\lambda_{ne})_{Age^2}$	$1\times10^{-5***}$	$1\times10^{-5***}$	$1\times10^{-5***}$
	(2×10^{-6})	(3×10^{-6})	(2×10^{-6})
$(\lambda_e-\lambda_{ne})_{MatchD}$	-0.0028	-0.0029	-0.0031
	(0.0053)	(0.0053)	(0.0053)
$(\lambda_e-\lambda_{ne})_{TasteVar}$	-0.0266	-0.0238	-0.0232
	(0.0768)	(0.0768)	(0.0768)
$(\lambda_e-\lambda_{ne})_{NumRev}$	0.0041^{***}	0.0041^{***}	0.0041^{***}
	(0.0014)	(0.0014)	(0.0014)
$(\lambda_e-\lambda_{ne})_{FreqRev}$	-0.0556^{***}	-0.0589^{***}	-0.0554^{***}
	(0.0535)	(0.0536)	(0.0535)
Log Likelihood	-191,770.2	-191,810.8	-191,756.3
N	133,688	133,688	133,688

Notes: Estimates in the above tables are the same as the model shown in Table 6 column (4). Column (1), (2), (3) represents models in which restaurants get a new draw of quality every quarter, every half-year, or every month.

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

Table 8: **MLE with Changing Restaurant Quality And Ethnic Restaurant Types**

	Restaurant Types	
	Non-Ethnic	Ethnic
σ_e	0.8959^{***}	0.9181^{***}
	(0.0194)	(0.0211)
σ_{ne}	0.9778^{***}	0.9678^{***}
	(0.0150)	(0.0160)
σ_ξ	0.1346^{***}	0.1212^{***}
	(0.0042)	(0.0085)
ρ_e	0.0112^{***}	
	(0.0224)	
ρ_{ne}	-0.1245^{***}	
	(0.0187)	
$(\mu + \lambda_{ne})_{Age}$	-0.0032^{***}	-0.0034^{***}
	(0.0003)	(0.0004)
$(\mu + \lambda_{ne})_{Age^2}$	$5.28 \times 10^{-6**}$	2.61×10^{-6}
	(2.69×10^{-6})	(4.71×10^{-6})
$(\mu + \lambda_{ne})_{MatchD}$	0.0370^{***}	
	(0.0040)	
$(\mu + \lambda_{ne})_{TasteVar}$	-0.2551^{***}	
	(0.0354)	
$(\mu + \lambda_{ne})_{NumRev}$	-0.0061^{***}	
	(0.0010)	
$(\mu + \lambda_{ne})_{FreqRev}$	0.0246	
	(0.0397)	
$(\lambda_e - \lambda_{ne})_0$	0.0171	
	(0.0233)	
$(\lambda_e - \lambda_{ne})_{Age}$	-0.0003	
	(0.0002)	
$(\lambda_e - \lambda_{ne})_{Age^2}$	$1.02 \times 10^{-5***}$	
	(2.81×10^{-6})	
$(\lambda_e - \lambda_{ne})_{MatchD}$	-0.0031	
	(0.0051)	
$(\lambda_e - \lambda_{ne})_{TasteVar}$	-0.0252	
	(0.0767)	
$(\lambda_e - \lambda_{ne})_{NumRev}$	0.0041^{***}	
	(0.0014)	
$(\lambda_e - \lambda_{ne})_{FreqRev}$	-0.0575	
	(0.0535)	
Log Likelihood	-191,758.9	
N	133,688	

Notes: 1. Estimated model in this table adds to the baseline model shown in Table 7 to allow quality signal noise, quality shock, and restaurant rating time trends to differ by restaurant ethnic status. 2. The cuisine type information is reported by Yelp. We classify a restaurant as ethnic if its Yelp cuisine category contains words indicating Chinese, Thai, Vietnamese, Asian, Korean, Indian, Ethiopian, Mediterranean, Peruvian, Russian, or Moroccan food.

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

Table 9: Simple, Moving and Optimal Averages Comparison

A. Distribution of Δ. ($\Delta = \hat{\mu}_{other} - \hat{\mu}_{optimal}$)				
	$\Delta < -0.15$	$\Delta > 0.15$	$\Delta < -0.25$	$\Delta > 0.25$
Chilling Model (interpret time trend as rating inflation/deflation)				
Simple average	29.42%	4.22%	13.16%	0.22%
6-month moving average	43.74%	11.61%	26.57%	5.88%
12-month moving average	41.55%	8.01%	23.30%	3.19%
18-month moving average	39.38%	6.73%	21.23%	2.31%
Non-chilling Model (interpret time trend as quality change)				
Simple average	5.04%	8.56%	0.85%	2.06%
6-month moving average	20.80%	13.10%	11.01%	7.23%
12-month moving average	15.14%	8.92%	6.5%	4.05%
18-month moving average	12.23%	7.50%	4.71%	3.02%

B. Distribution of rounded Δ. ($\Delta = round(\hat{\mu}_{simple}) - round(\hat{\mu}_{optimal})$)				
	$\Delta \leq -0.5$	$\Delta \geq 0.5$	$\Delta \leq -1$	$\Delta \geq 1$
Chilling Model	19.37%	6.02%	0.06%	0%
Non-chilling Model	6.19%	8.25%	0%	0.01%

Notes: 1. The above table shows the percentage of reviews with the differences between other averaging methods and optimal average exceeding 0.15 and 0.25. The calculation of optimal ratings is based on the model differentiating ethnic and non-ethnic restaurants, and with quality change every quarter. 2. The optimal ratings are for every review written on restaurants with at least 3 reviews in total in the sample. This covers 3,345 restaurants and 133,668 ratings. 3. In the upper panel that calculates optimal rating in the chilling model, we assume that the decreasing rating trends with time is due to rating inflation by early reviews and rating deflation by late reviews, and hence we correct by adjusting down early reviews and adjusting up late reviews to reflect the preference of an average consumer. In the lower panel that calculates optimal rating in the non-chilling model, we assume that decreasing rating trend is reflecting the decrease in actual restaurant quality. Hence, we did not correct for the time trends. The distribution of the averages gap in chilling model is skewed to the left due to the domination of rating deflation in old restaurants, and the optimal rating adjusted up to compensate the deflation. 4. Panel B rounds the simple average and optimal averages to every 0.5 points. 5. Panel D examines differences between simple and optimal ratings based on the time the reviews is left on the restaurants.

Table 10: Simple and Optimal Averages Comparison for Early and Late Restaurant Reviews

Distribution of Δ. ($\Delta = \hat{\mu}_{simple} - \hat{\mu}_{optimal}$)				
	$\Delta < -0.15$	$\Delta > 0.15$	$\Delta < -0.25$	$\Delta > 0.25$
Sample last review on each restaurant (3,345 reviews)				
Chilling model	38.09%	3.29%	18.86%	0.24%
Non-chilling model	5.05%	13.90%	1.14%	4.19%
0-2 Years (57,688 reviews)				
Chilling model	1.85%	8.58%	0.26%	0.34%
Non-chilling model	2.74%	1.56%	0.28%	0.10%
2-4 Years (65,366 reviews)				
Chilling model	46.79%	0.95%	19.33%	0.14%
Non-chilling model	6.67%	12.32%	1.21%	2.87%
>4 Years (10,614 reviews)				
Chilling model	72.24%	0.62%	45.30%	0.01%
Non-chilling model	7.57%	22.9%%	1.75%	7.74%

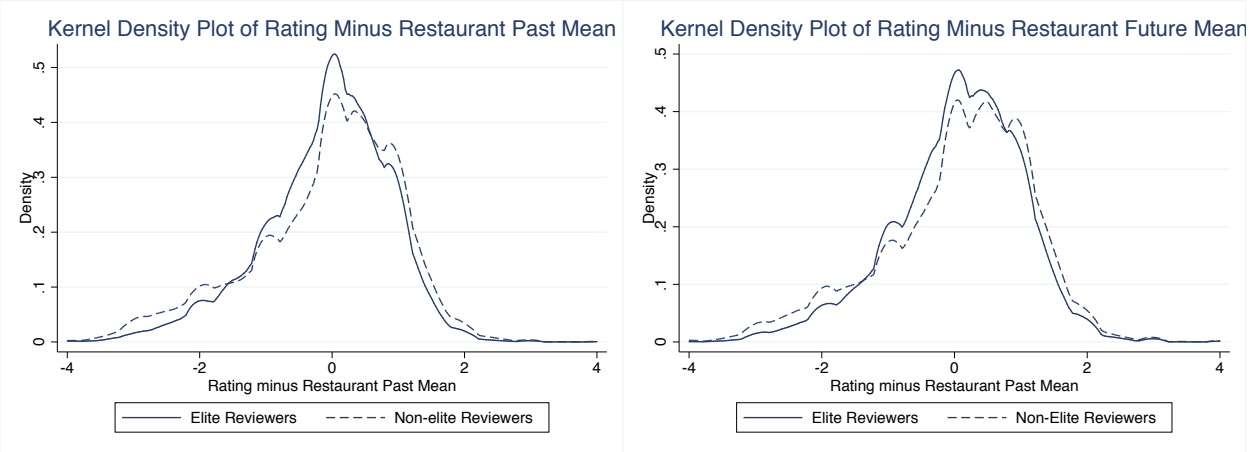
Notes: This table shows the differences between simple and optimal averages in early and late restaurant reviews. Pooling the last review written on each restaurant in our sample, the average number of days the last review is written since the restaurant's first review is 1,030 days, median is 1,099 days and the standard deviation is 485 days.

Table 11: Characteristics of Review with Different Simple and Optimal Averages Gaps

	$\Delta < -0.15$	$-0.15 \leq \Delta < 0.15$	$\Delta > 0.15$
Chilling Model (interpret time trend as rating inflation/deflation)			
# of Days since Restaurant's 1st Review	40.44	21.49	9.27
Restaurant Review Frequency	0.07	0.08	0.13
Matching Distance	12.75	13.23	20.95
Reviewer Taste Variance	2.68	2.78	2.83
# of Reviews Each Reviewer Written	22.35	22.86	20.86
Reviewer Review Frequency	0.33	0.42	0.48
Non-chilling Model (interpret time trend as quality change)			
# of Days since Restaurant's 1st Review	32.74	24.92	39.29
Restaurant Review Frequency	0.1	0.08	0.07
Matching Distance	10.79	13.17	17.42
Reviewer Taste Variance	2.7	2.76	2.75
# of Reviews Each Reviewer Written	28.05	22.55	20.15
Reviewer Review Frequency	0.4	0.4	0.35

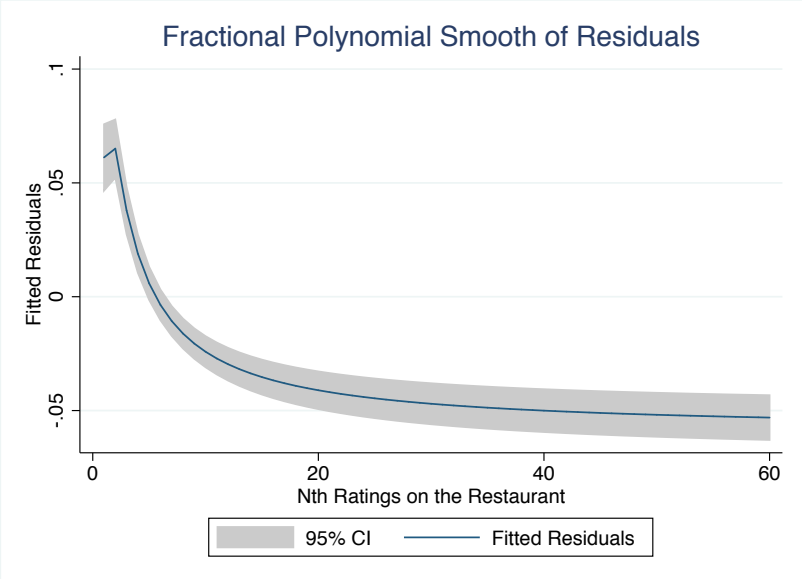
Notes: This table shows the mean characteristics of restaurants and reviewers with the gap between simple and optimal ratings greater and smaller than 0.15.

Figure 1: **Distribution of Ratings Relative to Restaurant Mean by Elite Status**



Notes: 1. The figure on the left plots the distribution $Rating_{rn} - \overline{Rating}_{rn}^{BF}$, where $Rating_{rn}$ is the n^{th} rating on restaurant r, and $\overline{Rating}_{rn}^{BF}$ is the arithmetic mean of past $n-1$ ratings on restaurant r before n. Similarly, the figure on the right plots the distribution of $Rating_{rn} - \overline{Rating}_{rn}^{AF}$, where $Rating_{rn}$ is the n^{th} rating on restaurant r, and $\overline{Rating}_{rn}^{AF}$ is the arithmetic mean of future ratings on restaurant r until the end of our sample. 2. These figures show that ratings by elite reviewers are closer to a restaurant's average rating.

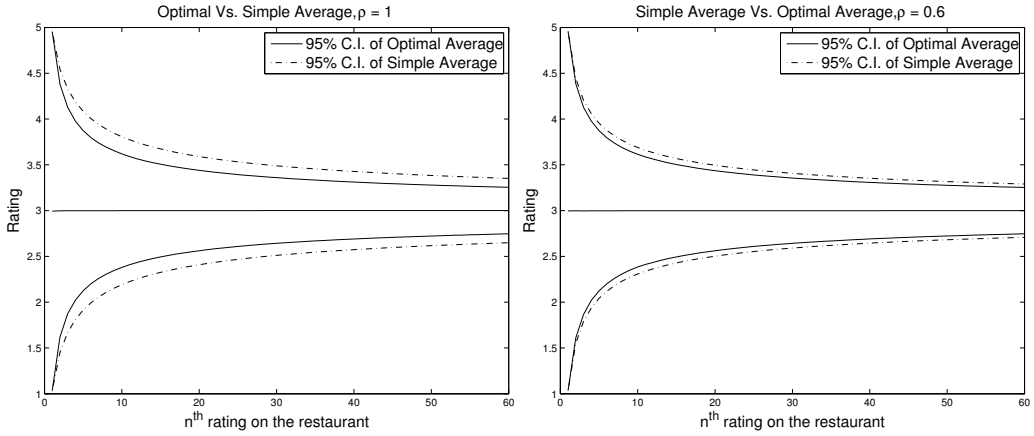
Figure 2: **Restaurants Experience a "Chilling Effect"**



Notes: This figure shows the rating trend within a restaurant over time. Ratings are on average more favorable to restaurants in the beginning and decline over time. We plot the fractional polynomial of the restaurant residual on the sequence of reviews. Residuals $\epsilon_{rn,year}$ are obtained from regression $Rating_{rn,year} = \mu_r + \gamma_{year} + \epsilon_{rn,year}$.

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

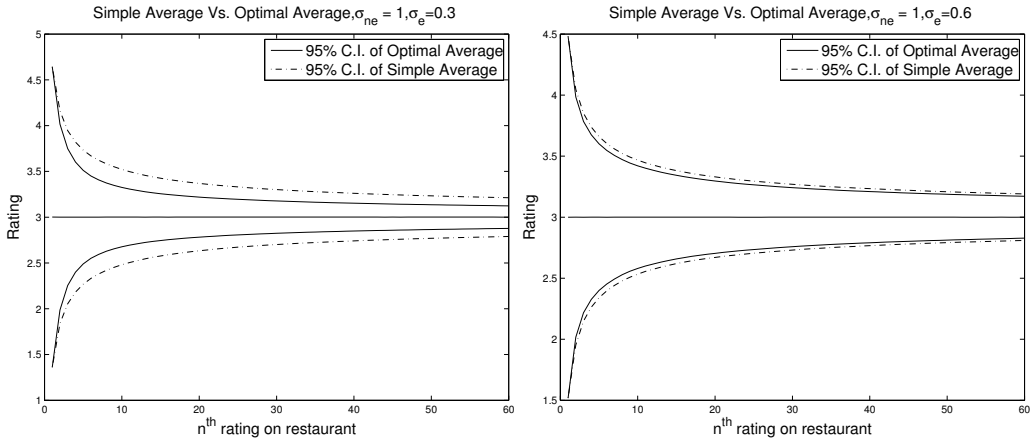
Parameters	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 1$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0.6$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$



Notes: The above figures simulated 95% confidence interval for optimal and simple average ratings in predicting true restaurant quality. When reviewers have popularity concerns, arithmetic and optimal averages are both unbiased estimates for true quality. But, relative to arithmetic average, optimal aggregation converges faster to the true quality, and the relative efficiency of optimal average is greater when reviewers' social incentive is larger.

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

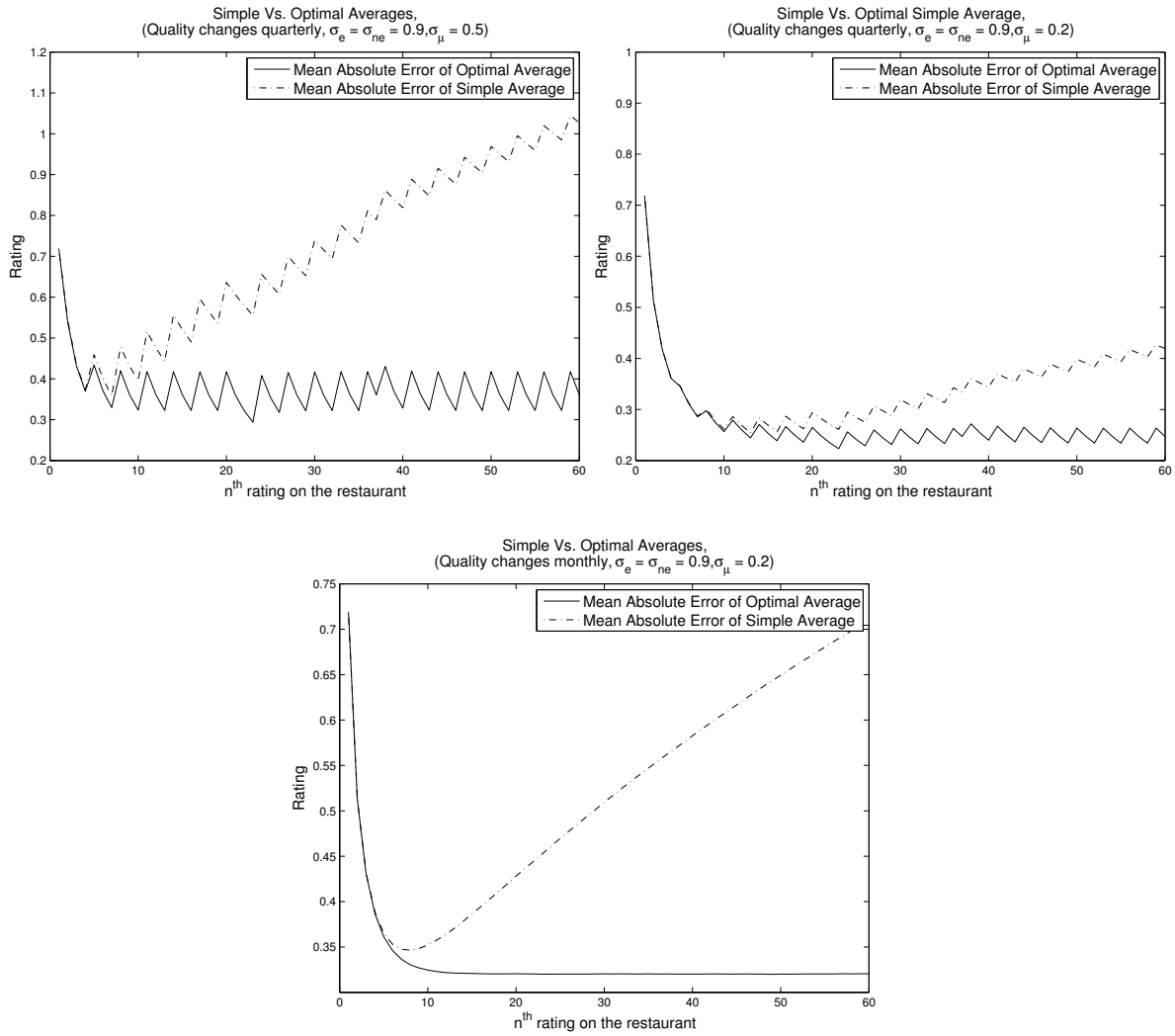
Parameter	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.6$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.3$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$



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 average converges faster to true quality. The difference in converging speed increases when elite reviewers' precision relative to that of non-elite reviewers is larger.

Figure 5: **Comparing Optimal and Simple Averages: Restaurants with Quality Random Walk**

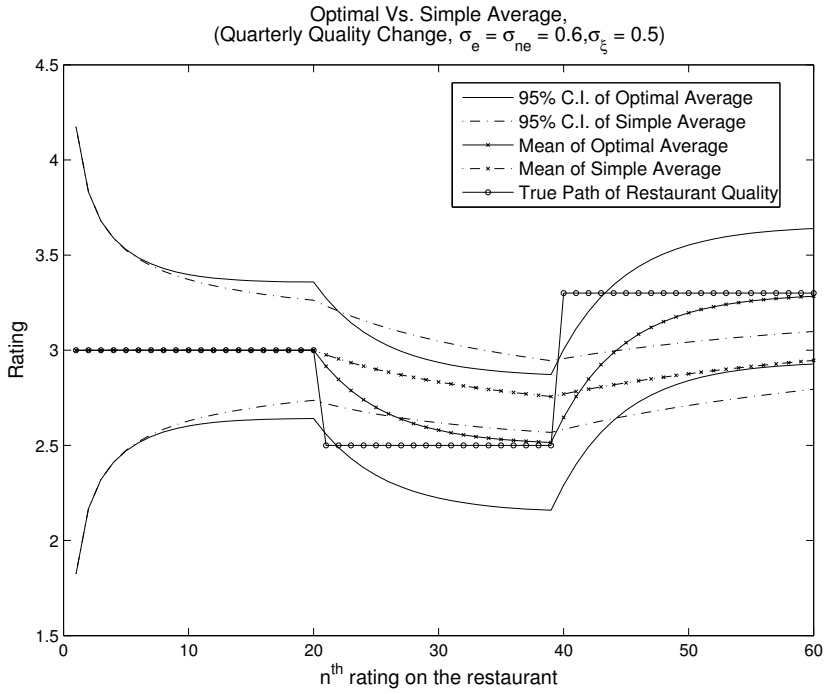
	ρ	σ	*Quality Update Frequency*	*StdDev of* $\Delta_{Quality}$
(Top Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.9$	Quarterly	$\sigma_\xi = 0.5$
(Top Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.9$	Quarterly	$\sigma_\xi = 0.2$
(Bottom)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.9$	Monthly	$\sigma_\xi = 0.2$



Notes. 1. The above figures plot the mean absolute errors of optimal and simple average ratings in estimating quality when quality evolves in a random walk process. To isolate randomness in review frequency on restaurants, we fix the frequency of reviews on restaurants to be one per month. We simulate a history of 60 reviews, or a time span of 5 years. 2. The figures show that simple averages become more erroneous in representing the true quality over time while the optimal average keeps the same level of mean absolute error. The error of simple average is greater compared with optimal average if a restaurant has larger variance in quality, and changes quality more frequently.

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

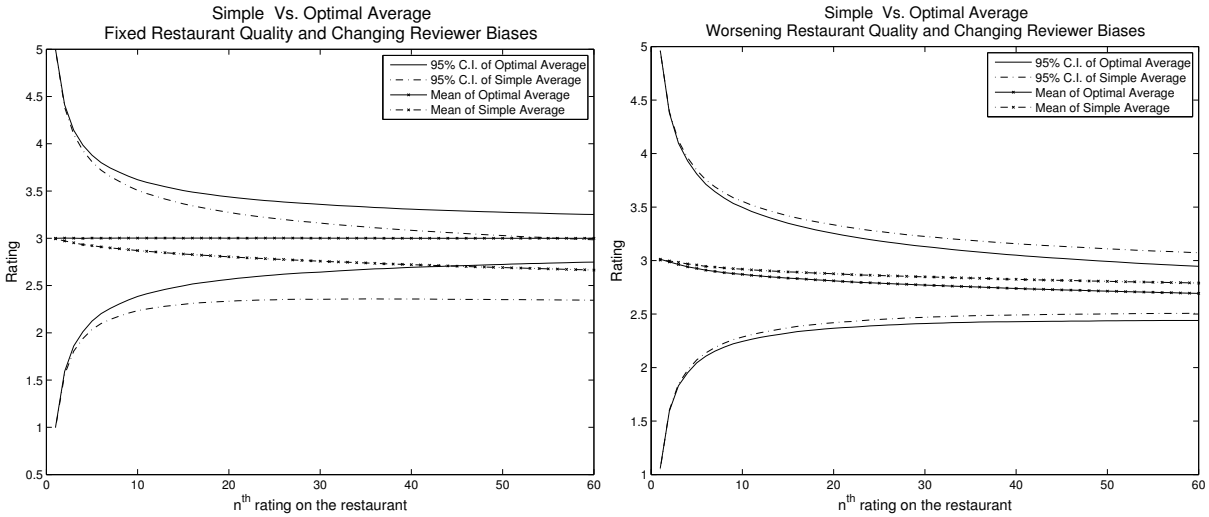
ρ	σ	*Quality Update Frequency*	*StdDev of* $\Delta_{Quality}$
$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figures plot the simulated mean and 95% confidence interval for the average ratings for one hypothetical restaurant quality path realized in a random walk process. The quality drops to 2.5 before restaurant receives its 20^{th} rating, and rises to 3.25 before it receives its 40^{th} rating. Optimal aggregation adapts to changes in restaurant's true quality, while simple average becomes more biased in representing restaurant's true quality. Since the optimal aggregation algorithm only gives weights to recent ratings and simple average gives equal weights to all past ratings, standard error of optimal average shrinks slower than simple average.

Figure 7: Simulated Ratings when Reviewers are Biased

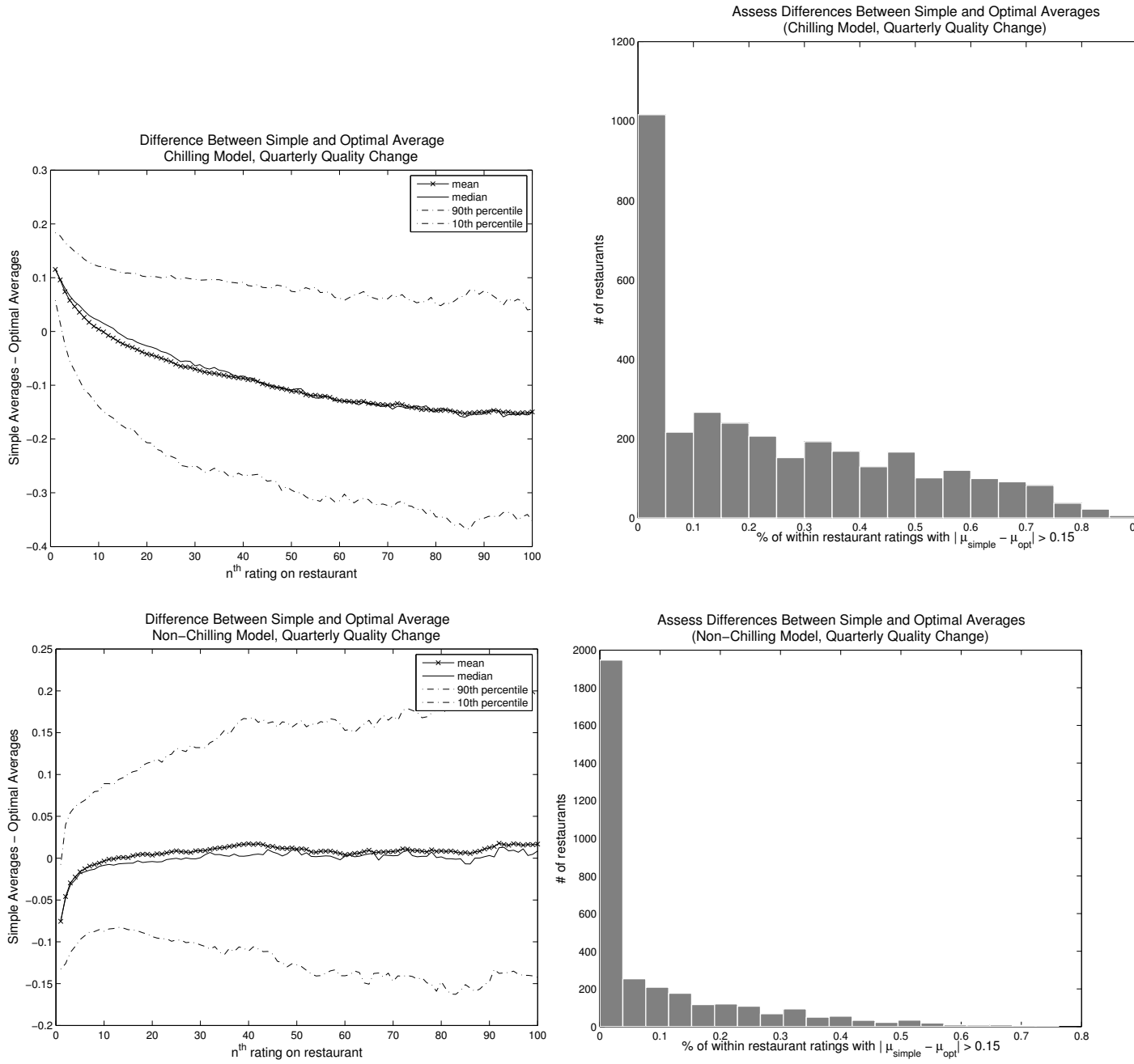
Parameters	Value	Parameters	Value	Parameters	Value
$\rho_e,\ \rho_{ne}$	0	$\frac{\partial(\mu+\lambda_e)}{\partial Restaurant\ Age^2}$	4×10^{-6}	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Frequency}$	0.0256
$\sigma_e,\ \sigma_{ne}$	1	$\frac{\partial(\mu+\lambda_e)}{\partial Match\ Distance}$	0.0367	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Restaurant\ Age^2}$	0.00001
$Quality_0$	3	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Taste\ To\ Variety}$	-0.2453	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Reviewer\ Review\ \#}$	0.0041
$\frac{\partial(\mu+\lambda_e)}{\partial Restaurant\ Age}$	-0.0032	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Review\ \#}$	-0.0062	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Reviewer\ Frequency}$	-0.0556



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 simple average is biased without correcting the review bias.

Figure 8: Optimal and Simple Average Algorithm Applied on Sample Data

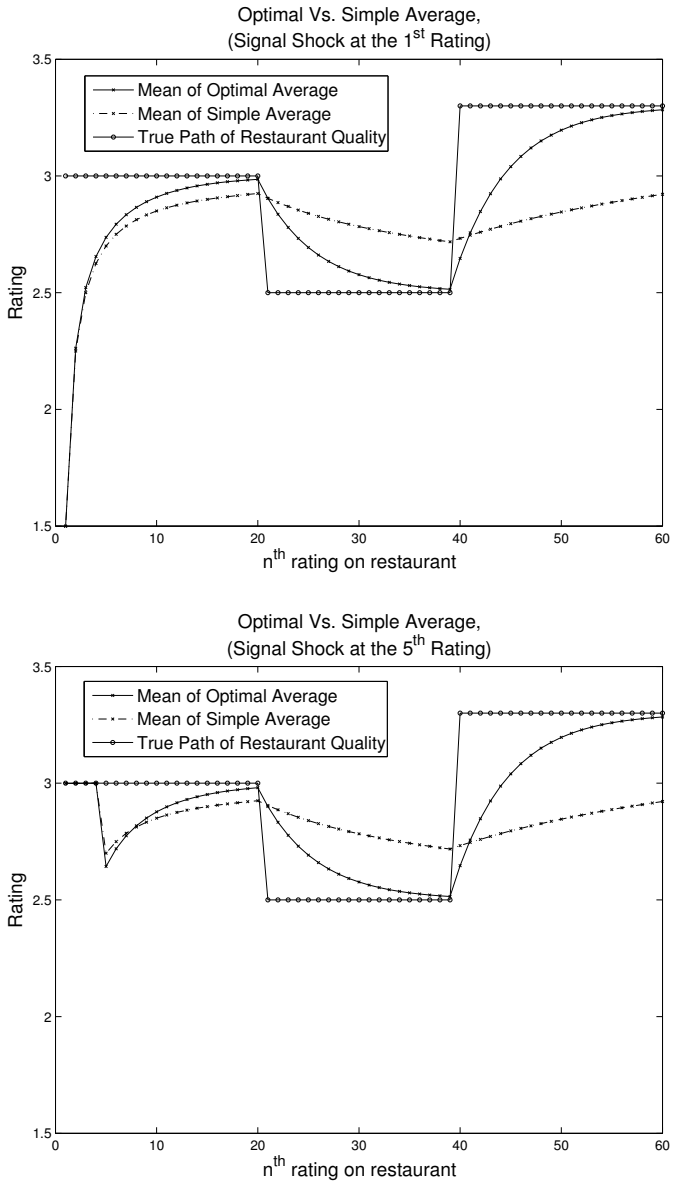
(Top)	*(Bottom)*
Chilling Model	Non-Chilling Model



Notes: 1. Figures on the left plot the trend of mean and 95% confidence interval for $\mu_{rn}^{simple} - \mu_{rn}^{optimal}$. Figures on the right plot the frequency of restaurants that have proportions of ratings satisfying $|\mu_{rn}^{optimal} - \mu_{rn}^{simple}| > 0.15$. 2. The upper panel assumes that the rating trend over time comes from reviewer bias, and the lower panel assume that the rating trend over time is the change in restaurants' true quality.

Figure 9: **Comparing Optimal and Simple Averages: "Fake" Review**

	Fixed Signal	ρ	σ	*Quality Update Frequency*	*StdDev of* $\Delta_{Quality}$
(Top)	$s_1 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$
(Bottom)	$s_5 = 1.5$	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figures plot the simulated mean of the average ratings for a single restaurant whose quality follows the random walk process. A "fake" review that is fixed at 1.5 appears as the first or the fifth review on the restaurant. We consider the review "fake" if a reviewer misreports her signal. Both aggregating algorithms weight past ratings, and are affected by the "fake" rating. But compared with arithmetic mean, optimal aggregation "forgets" about earlier ratings and converges back to the true quality in a faster rate.