NBER WORKING PAPER SERIES

AGGREGATION OF CONSUMER RATINGS: AN APPLICATION TO YELP.COM

Weijia Dai
Ginger Z. Jin
Jungmin Lee
Michael Luca

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

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2012, Revised February 2018

Previously circulated as "Optimal Aggregation of Consumer Ratings: An Application to Yelp.com." 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Aggregation of Consumer Ratings: An Application to Yelp.com
Weijia Dai, Ginger Z. Jin, Jungmin Lee, and Michael Luca
NBER Working Paper No. 18567
November 2012, Revised February 2018
JEL No. D8,L15,L86

ABSTRACT

Because consumer reviews leverage the wisdom of the crowd, the way in which they are aggregated is a central decision faced by platforms. We explore this "rating aggregation problem" and offer a structural approach to solving it, 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 to restaurant reviews from Yelp.com, we construct an adjusted average rating and show that even a simple algorithm can lead to large information efficiency gains relative to the arithmetic average.

Weijia Dai
University of Maryland
Department of Economics
3114 Tydings Hall
College Park, MD 20742
daisy.w.dai@gmail.com

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

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

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

1 Introduction

The digital age has transformed the way that consumers learn about product quality. Websites ranging from Yelp and TripAdvisor to eBay and Amazon use crowdsourcing to generate product ratings and reviews. This has dramatically increased the amount of information consumers have when making a decision. Intuitively, 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 analyze ratings on a restaurant review website (Yelp.com) and develop a method to systematically aggregate individual ratings into one adjusted average rating.

Why focus on one single metric? In principle, the platform could simply present all reviews and allow consumers to decide for themselves how 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). Many platforms, including Yelp, highlight the arithmetic mean of consumer ratings as one aggregation of product quality. Both data analysis and online survey suggest that Yelp consumers pay more attention to the aggregate rating.[1] There is also evidence that only a tiny fraction of consumers actively create content while the vast majority read reviews without responding.[2] Because most review users are inattentive, the method chosen to aggregate ratings is of first-order importance.

Using the simple average to aggregate ratings imposes restrictive assumptions on the informational content of ratings. If each rating is an unbiased, i.i.d. signal of the constant, true quality, the simple average is statistically efficient. However, it is easy to imagine situations in which this is not the case. For example, consider two hypothetical restaurants. The first one receives 2-star ratings for 12 months and then 4-star ratings for the next 12 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 after 24 months. However, a reader would likely favor a restaurant with an upward trend than a downward trend.

[1] Luca (2011) shows that Yelp consumers respond directly to the average rating even though it is coarser than the underlying information. The importance of the average rating is also supported by an online survey we conducted for this study. In this survey, we ask subjects to report their general use and understanding of restaurant ratings, without mentioning Yelp. Out of the 239 respondents, 93.7% use the average rating to choose a restaurant, but a much lower percentage of respondents said that they pay attention to other review information such as the number of reviews, rating trends, or reviewer profile. More details of the survey are presented in Section 6.

[2] Yelp had 167 million unique visitors per month in the first quarter of 2016 (source: http://www.yelp.com/factsheet), but according to a 2011 blog post of Yelp, journalist Susan Kuchinskas estimated that "only 1 percent of users will actively create content. Another 9 percent, the editors, will participate by commenting, rating, or sharing the content. The other 90 percent watch, look, and read without responding." (https://www.yelpblog.com/2011/06/yelp-and-the-1990-rule, accessed on June 5, 2016).

The goal of this paper is to create a systematic aggregation of consumer ratings, as a proxy of the concurrent vertical quality of a restaurant. To do so, we create an adjusted average rating that satisfies two criteria: first, observable preferences and biases of different types of reviewers must be separated from a reviewer's signal of vertical quality; second, ratings must be weighted to account for their informativeness, with more weight assigned to ratings containing more information. Our hope is to shed light on the rating aggregation problem, and to move toward optimal aggregation of consumer ratings – where the definition of optimal depends on the objective of the platform.

To derive an aggregation algorithm that meets the above two criteria, we develop a structural model. The model focuses on a reviewer's rating decision after she visits a restaurant, observes a vertical quality signal of the restaurant, and decides to review it on Yelp. We allow reviewers to vary in accuracy (some reviewers are more precise than others), stringency (some reviewers leave systematically lower ratings)[3], and social incentives (some reviewers may prefer to conform to or deviate from prior ratings), conditional on observed attributes of reviewers and the timing of reviews. We also account for the fact that a restaurant's quality can change over time. The model is estimated using the entire Yelp rating history for 4,101 restaurants in Seattle, including the reviewer identity of each rating.

Our model is subject to an important caveat: because our data do not contain any information on the consumers who choose not to patronize a restaurant, or the consumers who patronize the restaurant but do not leave a review, we cannot explicitly model reviewer selection and identify it in the real data. That being said, we control for as many reviewer characteristics as possible, including the elite status of the reviewer, the number and frequency of her reviews, and the type of restaurants she has reviewed previoåusly. Our model also allows restaurant ratings to follow a time trend since the first review on the same restaurant. Li and Hitt (2008) have demonstrated a downward trend of ratings on Amazon book reviews, which they interpret as a "chilling effect" driven by the fact that an enthusiastic consumer is likely to purchase and review the book earlier than a general consumer. In our case, the trend could reflect reviewer selection or a decline of real restaurant quality. We do not have enough data to tease out these two, but if reviewer selection is one reason driving the chilling effect, it is implicitly controlled for in the time trend.

With that caveat in mind, we are able to estimate the model using rating variation across reviewers, restaurants, and time. Assuming reviewers can observe all previous ratings and unpack their information contents (in rational expectation), we back out consumers' quality signals and use the Bayesian method to construct the posterior belief of restaurant quality given all the ratings available at any particular time. This posterior belief is defined as our "adjusted

[3]We assume that horizontal preferences of reviewers affects reviewer stringency. Hence, when we present the vertical quality to general readers of Yelp reviews, we should benchmark the adjusted average to the stringency of one type of reviewers. We choose to benchmark it to a reviewer with average attributes. The construction of reviewer horizontal preference is presented in Section 2.1.

average".

Empirical estimation sheds light on a few interesting findings. For instance, we find that elite reviewers observe quality signal with a higher precision than non-elite reviewers, suggesting that elite reviews should carry more weight in the aggregated rating.[4] Elite and non-elite reviewers also differ in their social incentives: assuming restaurant quality changes by quarter, we find that elite reviewers put a small, positive weight on past ratings when they report their own signal, but non-elite reviewers place a negative weight on past ratings. This implies that when a non-elite reviewer draws a quality signal that is different from her belief based on past ratings, she tends to overweight her own signal, as if to emphasize how her signal differs from previous ratings. In contrast, elite reviewers tend to (slightly) herd with previous ratings, probably because they care about their social status on Yelp and therefore have stronger social incentives. Moreover, we find that reviewer stringency differs by number and frequency of restaurant reviews, the variety of restaurants the reviewer has been to in the past, and whether the reviewer is reviewing a restaurant similar in cuisine type to the ones she has often reviewed before. Above all, these findings suggest that reviewer history and attributes should be accounted for when we aggregate ratings into one adjusted average.

The empirical estimates also highlight the importance of quality change and time trend. We show that much of the adjusted-vs-simple-average difference is driven by the evolution of restaurant quality. This is because the simple average weights every rating equally despite quality changes but the adjusted average increases the weight assigned to more recent ratings and hence adapts to changes in quality. More importantly, our results suggests that simple average deviates more from true quality as the number of ratings grows over time. This is intuitive, as the most recent rating should be the most informative about concurrent quality, but the simple average tends to give less and less weight to the latest rating as ratings accumulate. Consistent with Li and Hitt (2008), we find a significant downward trend of reviews within a restaurant. Depending on how we interpret this trend, we find 19.1-41.38% of the simple average ratings are more than 0.15 stars away from our adjusted ratings, and 5.33-19.1% are more than 0.25 stars away at the end of our sample period. The deviation grows significantly over time, suggesting that large differences could be made by implementing our aggregated ratings, especially as Yelp grows.

In addition to empirical estimation, we use simulations to demonstrate the advantage of our algorithm. After simulating true quality and reviewer ratings, we compute the adjusted and simple averages, and contrast them to the true quality at the time of aggregation. In theory, based on our Bayesian model assumptions, the Kalman filter used by the adjusted average is the best for linear models with Gaussian errors (in term of consistency and minimizing mean square error) and thus should be better than the simple average (Ljungqvist and Sargent 2012; Welch and Bishop 2001). Our simulation confirms this argument.

In another validity check, we conducted an online survey and asked subjects directly how

[4]The "Elite" status is a badge displayed next to the reviewer name, and is rewarded by Yelp to prolific reviewers who write high quality reviews.

they use restaurant ratings (without mentioning Yelp). The key rationale behind our algorithm – namely more weight should be given to more recent reviews and more precise reviews – are consistent with the preference reported by the survey subjects. Because this survey is completely independent of our model and data, its consistency extends support to our model.

We contribute to the growing literatures on information aggregation and consumer reviews. Like Li and Hitt (2008), we find a downward trend of ratings within the same product. Li and Hitt attribute it to reviewer selection, but we are open to the idea that it may capture reviewer selection or real quality decline. Our algorithm generates different adjusted averages, depending on how we interpret the downward trend. Researchers have tried other ways to improve aggregate ratings. For example, Glazer et al. (2008) have considered alternatives to simple average ratings in the context of health plan report cards, but their work is largely theoretical. Ghose, Ipeirotis and Li (2012) have aggregated ratings via demand estimation.[5] Two other concurrent papers have independently studied bias in online reviews. First, Nosko and Tadelis (2015) were concerned about bias in seller reputations on online platforms. Using eBay transaction data, they construct a new quality measure that takes into account the non-response from buyers. By running a controlled experiment that promotes the search ranking of better quality sellers using the the adjusted quality measure, they find that the new ranking algorithm increases retention of buyers. Second, Fradkin et al. (2017) studied how the design of the feedback system of Airbnb affect the informativeness of ratings and reviews. These three papers all rely on observing the transaction data in conjunction with the consumer reviews. In comparison, we design the review aggregation algorithm without complementary data on how consumers use such reviews when they choose a product. This form of data constraint is faced by many opinion generation websites that offer consumer ratings but do not sell the rated products.[6] Finally, our model of social incentives is related to the vast literature on information cascade and observational learning such as Banerjee (1992) and Bikhchandani et al. (1992), and more recently, Alevy et al. (2007), and Eyster and Rabin (2010). The simple average presented by the platform is similar to the naive individual in Eyster and Rabin (2010) who believes that each previous person's action reflects solely that person's private information and ignores the fact that each previous person's action already embeds early movers' signals. As a result, the model predicts that the crowd will herd on incorrect actions with positive probability even in rich-information settings. Our paper complements this literature by empirically estimating the

[5]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."

[6]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. 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.

degree of herding (or deviation) on a real consumer review website. Though our estimation finds herding and deviation to be statistically significant, their magnitudes are small and their impact on the aggregate ratings is less than that of quality change and time trend.

2 Model and Estimation

Consider a consumer review website that has already gathered many consumer ratings on many products over a period of time. Our goal is to systematically summarize these ratings into a single metric of concurrent quality for each product. In this section, we present a reviewer rating model in which the reviewer chooses how to rate a restaurant in her review after she visits the restaurant, observes a quality signal at the restaurant, and decides to review it on Yelp. As detailed below, because our data consist of reviewer ratings only, we have to focus on the rating stage and abstract away from reviewer selection that may occur before that stage.

2.1 Model Setup

Consider reviewer i who writes a review for restaurant r at calendar time t_n.[7] 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.[8]

The second incentive is the reviewer's social incentive that may generate a positive or negative correlation across ratings ((Muchnik et al., 2013). 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 (Chen et al., 2010). Because most Yelp users read but do not write reviews, the above social incentive 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 users who read reviews only (we have no data on them), we assume this social incentive motivates a

[7] 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.

[8] 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.

reviewer to be "right" about the true restaurant quality. Alternatively, we can also interpret such motivation by 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. While the above social incentive generates positive herding, there might also be incentives for a reviewer to deviate from previous ratings. For example, if one reviewer expects a restaurant to be a 4-star experience after reading others' reviews, her actual 3-star experience may motivate her to give the restaurant 2-stars, not only out of disappointment but also out of an intention to pull Yelp-reported simple average closer to her own experience. Whether conforming or differentiating, social incentives imply that reviewers put some weight on prior ratings. We identify such weight from reviewers' rating behavior, but do not attempt to distinguish the psychological factors behind social incentives.

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 a loss function:

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

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}, θ_{rn} is the subjective stringency of reviewer i, 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 rating that minimizes $F_{rn}^{(1)}$ is:

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

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. In Appendix A, we show an alternative model to capture reviewer incentive to differentiate from prior ratings. It gives rise to exactly the same equation except for $\rho_i < 0$. For this reason, our empirical estimation does not impose a sign on ρ_i.

It is worth noting that the above model assumes each reviewer is fully rational thus has perfect information about the other reviewers' observable attributes. This knowledge allows her to first back out each reviewer's signal before her, and then compute the posterior belief of quality $E(\mu_{rt_n}|x_{rt_1}, x_{rt_2}, ...x_{rt_{\{n-1\}}}, s_{rt_n})$. In the empirical section, we will present an alternative model of limited attention and show that it does not fit the data as well as our Bayesian model.

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 adjusted aggregation can only improve efficiency by giving more weight to more precise reviewers or reviewers with greater social incentives.

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 random walk 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, which increases with the timing of the earlier date (t) but is independent of the time between t and t'.[9] Recall that x_{rt_n} is the n^{th} review written at time t_n since r was first reviewed. We can express the n^{th} reviewer's signal as $s_{rt_n} = \mu_{rt_n} + \epsilon_{rn}$, where $\mu_{rt_n} = \mu_{rt_{n-1}} + \xi_{t_{n-1}+1} + \xi_{t_{n-1}+2} + ... + \xi_{t_n}$.[10]

Reviewer Heterogeneity and Reviewer-Restaurant Match In addition to random changes in restaurant quality and random noise in reviewer signal, reviewers may differ in stringency, social incentives, and signal precision. Systematic information aggregation should account for these differences.

One observable reviewer heterogeneity is elite status. We 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 $\sigma_e^2 < \sigma_{ne}^2$. We also allow elite and non-elite reviewers to have different stringencies, λ_e and λ_{ne}.

The second reviewer attribute we use is the number of reviews that reviewer i has submitted for Seattle restaurants before writing a new review for restaurant r at time t. We denote it as $NumRev_{it}$. Another reviewer attribute 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 $FreqRev_{it}$.

[9] The correlation structure is detailed in Appendix B.3.

[10] 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 $\{s_{rt_1}, ..., s_{rt_{n-1}}\}$. 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)).

We also attempt to capture the heterogeneity in reviewer taste by reviewer-restaurant match. In reality, reviewers may have differentiated 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 comparing characteristics of the restaurants with the average restaurants the reviewer has written reviews for in the past. In particular, we use 14 cuisine type indicators and 5 price categories defined by Yelp[11] and decompose them into 8 orthogonal factors $F_r = [f_{r,1}, ..., f_{r,8}]$.[12] We use F_{il} to denote the vector of factors of the l^{th} restaurant that reviewer i visited, and $\overline{f}_{il,q} = \frac{1}{m-1}\sum_{l=1}^{m-1} f_{il,q}$ to denote the mean in factor q among the $m-1$ restaurants that i has visited. We then collapse a reviewer history into two metrics: $C_{it}(= \frac{1}{m-1}\sum_{l=1}^{m-1} F_{il},)$ measures the average restaurant that this reviewer has written reviews for before she writes her m^{th} review at time t; and $TasteVar_{it}(= \sqrt{\sum_{q=1}^{8} \frac{1}{m-2}\sum_{l=1}^{m-1}(f_{il,q} - \overline{f}_{il,q})^2})$ measures the variety of $m-1$ restaurants that she has written reviews for before her m^{th} review at time t.[13] 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 as $MatchD_{rit} = (C_{it} - F_r)'(C_{it} - F_r)$.The shorter the matching distance, 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 ($TasteVar_{it}$). In the empirical model, we allow elite status to affect ρ in equation (2) and allow all other characteristics to affect only λ.[14]

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. This is on top of year fixed effects, ($Year_t$), which already captures the overall stringency or taste change of the Seattle population from one year to another.[15] By definition, this trend – which turns out to be negative and concave over time when we estimate it in quadratic terms – is subject to

[11]The cuisine indicators describe whether a restaurant is traditional American, new American, European, Mediterranean, Latin American, Asian, Japanese, seafood, fast food, lounge, bar, bakery/coffee, vegetarian, or others. They are not mutually exclusive. The five price categories are (1,2,3,4) as defined by Yelp plus a missing price category (which we code as 0).

[12]By construction, the sample mean of each factor is normalized to 0 and sample variance normalized to 1.

[13]If reviewer i has not reviewed any restaurant yet, we set her taste equal to the mean characteristics of restaurants ($C_{it} = 0$).

[14]We have tried to estimate the model that allows ρ_i to vary by reviewer attributes other than elite status, but none of other attributes significantly affect ρ based on the likelihood ratio test.

[15]Our model of time trend in addition to year fixed effects is consistent with Godes and Silva (2012), who suggest that the negative temporal trend may be due to the fact that that reviewers are becoming more critical and more negative in general, and they find that after conditioning on the year a review was written, ratings increase over time. In our case, the time trend since the first review of a restaurant remains negative after we control for year fixed effects.

multiple interpretations. It is possible that true restaurant quality declines over time for every restaurant.[16] 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) and we are unable to distinguish the two. When we calculate the adjusted average, we use the two extreme interpretations separately, in order to bound the true adjusted average.

Summary Above all, we assume: $\{\rho_e, \rho_{ne}\}$ capture the social incentives 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 restaurant quality or reviewer stringency change,[17] $\{\alpha_{freqrev}$, α_{matchd}, $\alpha_{tastevar}\}$ capture how restaurant and reviewer attributes change the stringency of non-elite reviewers, and $\{\lambda_{(e-ne)0}$, β_{age1}, β_{age2}, β_{numrev}, $\beta_{freqrev}$, β_{matchd}, $\beta_{tastevar}\}$ capture how restaurant and reviewer attributes change the stringency difference between elite and non-elite reviewers. We also 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.[18]

Note that our model allows 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}. To incorporate the possibility that different types of restaurants may differ in our key parameters regarding the precision of restaurant signals and the evolution of restaurant quality, we also estimate a model that allows $\{\sigma_e^2, \sigma_{ne}^2, \alpha_{age1}, \alpha_{age2}, \sigma_\xi^2\}$ to differ by 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.

2.2 Model Estimation and Identification

Maximum Likelihood Estimation According to the derivation of $E(\mu_{rt_n}|s_{rt_1}, ...s_{rt_n})$ illustrated in Appendix B.2, 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 $logL = \sum_{r=1}^{R} logL(x_{rt_1}, x_{rt_2}, ..x_{rt_{N_r}})$.[19]

Consistent estimation of all other parameters depends on estimating restaurants' initial quality $\{\mu_{r0}\}_{r=1}^{R}$ consistently, 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

[16] Note that this decline is in addition to the random walk evolution of restaurant quality because the martingale deviation is assumed to have a mean of zero.

[17] 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.

[18] A summary of the statistical data generating process is available in Appendix B.1.

[19] The parameters to be estimated are $\{\mu_{r0}\}_{r=1}^{R}$,σ_ξ, (σ_e, σ_{ne}), (ρ_e, ρ_{ne}), $(\alpha_{yeart}$, α_{numrev}, $\alpha_{freqrev}$, α_{matchd}, $\alpha_{tastevar}$, $\lambda_{(e-ne)0}$, β_{age1}, β_{age2}, β_{numrev}, $\beta_{freqrev}$, β_{matchd}, $\beta_{tastevar})$, and $(\alpha_{age1}, \alpha_{age2})$. In an extended model, we also allow $\{\sigma_e$, σ_{ne}, α_{age1}, α_{age2}, $\sigma_\xi\}$ to differ for ethnic and non-ethnic restaurants.

14. When we use simulated data to test the MLE estimation of observed reviews, we find that the 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. This way, the initial restaurant qualities $\{\mu_{r0}\}_{r=1}^{R}$ are cancelled out. The details to derive $f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}})$ are shown in Appendix C.

Estimating 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,[20] the Bayesian 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 $E(\mu_{rt_n}|s_{rt_1}, s_{rt_2}, .., s_{rt_n})$.[21] If we interpret the quadratic trend of ratings as changes in true quality, the Bayesian 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 infer restaurant quality using both interpretations in Section 5.

The Bayesian estimate of restaurant quality at t_n is also our adjusted average rating. Because we use the Kalman filter in our Bayesian calculation and the Kalman filter has been established to be the best for linear models with Gaussian errors (in term of consistency and minimizing mean square error), our adjusted rating should be better than the simple average (Ljungqvist and Sargent 2012; Welch and Bishop 2001). We will later confirm this via simulation and online survey.

Model 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, consider a simple case where restaurant quality is stable (i.e. $\sigma_\xi^2 = 0$). If everyone 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 incentives ($\rho_n = 0$). When $\rho_n > 0$, $Var(x_{rn})$ declines with n. If the n^{th} reviewer cares about social incentives 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 incentives, 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

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

[21] The estimation of $E(\mu_{rt_n}|s_{rt_1}, s_{rt_2}, .., s_{rt_n})$ is detailed in Appendix B.2.

each review equally according to the Bayesian principle.

Nevertheless, social incentive 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 social incentive 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, 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 incentive 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 stringency 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 capture the effect of reviewer attributes, restaurant characteristics, and reviewer-restaurant match on 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 ratings vary by restaurant age, reviewer attributes at time t, reviewer-restaurant match, and their interaction with elite status. It is important to note that we can only use observed variations to identify relative differences in stringency. Moreover, because we assume stringency to be additive to the true restaurant quality, we cannot separately identify the absolute magnitude of stringency and true quality. This does not hamper our ability to compute the adjusted average ratings. Though each Yelp review user may have his own stringency, we benchmark the stringency of the aggregated rating to the average stringency of all sampled reviewers.

2.3 Selection of Reviews

One limitation of our model is that we do not explicitly model the selection of consumers who decide to leave a review. In practice, reviewers select to purchase a product and conditional on purchase, select to leave a review. In principle, selection into a product tends to skew ratings upward (you are more likely to eat at a restaurant that you like). The decision to review has an ambiguous effect, depending on whether people are more likely to write a review after a good experience or a bad one.

A growing literature has attempted to measure the selection process by imposing struc-

tural assumptions. Moe and Trusov (2011) conjectured that consumer ratings reflect not only consumers' socially unbiased product evaluation but also the social dynamics in the product's ratings environment. They estimate the net effect of social influence on ratings and their subsequent effect on product sales. They find that social dynamics can have a direct impact on sales through rating valence. Moe and Scheweidel (2012) presented a model in which reviewers choose whether to rate and what rating to give jointly. They find that valence in the rating environment affects the likelihood of review, and active and inactive reviewers differ in their tendency to review restaurants depending on whether the restaurant's past ratings are in consensus or in controversy with their own. Also, several other papers have discussed the importance of taking into account the influence of selected purchase on ratings, for example, in Li and Hitt (2008), Hu et al. (2009) and Godes and Silva (2012).[22]

We do not model reviewer selection explicitly because our data do not tell us what type of consumers choose not to patronize a restaurant, or what type of consumers patronize but choose not to leave a review. Without such information, any structural assumption we put on the selection process is not testable in the real data. These structural assumptions can be arbitrary and may lead to spurious estimation results.

That being said, we believe our model has at least partially controlled for review selection based on observables. In particular, our model of rating heterogeneities is conditional on observed reviewer attributes such as elite status, the number of reviews written, the frequency of reviewing, and whether she visits a wide variety of restaurants. We have also conditioned on the timing of her review relative to the restaurant rating history and the match distance between a reviewer and the restaurant she is reviewing. If one type of reviewer is more stringent, and they prefer to review the restaurant later rather than earlier, we control for this selection by allowing reviewer stringency to vary by observed type and by incorporating time of review in the Bayesian updating structure and time trend. If there is a chilling effect due to reviewer selection (as Li and Hitt (2008) have argued), our time trend captures it in linear and quadratic terms. We acknowledge that our controls do not absorb all sorts of reviewer selection or reviewer heterogeneity. For example, if more stringent reviewers are only stringent when they review a 5-star restaurant but are not as stringent when they review a 3-star restaurant, that is not captured in our model. If an individual reviewer changes her selection rule from 2004 to 2008 and this change is not the same as the population taste change, our model will miss it too.

Our paper also complements two other concurrent papers investigating the selection process. Wu et al. (2015) use a Bayesian learning model on data from a Chinese restaurant review website similar to Yelp.com in order to measure the value of online reviews for consumers and firms. They study how consumers learn from different reviewers based on the the perceived correlation between their own tastes and the reviewers' tastes, but they do not consider reviewers' strategic

[22]In Hu et al. (2009), ratings are found to follow bimodal distributions on Amazon (with many one and five stars) and the paper attributed this to the tendency to review when opinions are extreme. We do not find the bimodal distribution pattern on Yelp that Hu et al. (2009) provide as evidence of significant reviewer selection.

reporting behavior as well as quality change. Our objective is to systematically aggregate existing ratings into one measure 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. Indeed, we find that reviewers with a wider variety of reviewing experience are relatively more stringent.

3 Yelp Data and Reduced Form Results

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.[23] 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. Yelp grants elite status to reviewers who review often and write high quality reviews. We do not observe the change of elite status within each reviewer; we observe only whether a reviewer has been granted elite status by the end of our data. For our purposes, we take elite status as fixed.[24] Another data limitation is that our data contain only star ratings given in each review (one to five), but do not include the text; hence, our analysis focuses on ratings. 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 rich within-reviewer variation.

Table 1 summarizes the main variables. In the top 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 day a restaurant receives reviews, the restaurant receives on average 0.16 reviews per day. The review frequency is highly heterogeneous – 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 arrives 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 two or three reviews far apart in time, while a small fraction of restaurants are reviewed frequently.

The bottom panel of Table 1 summarizes reviewer statistics. Although less than 10% of reviewers are elite, an average elite reviewer writes five times more reviews than a non-elite reviewer (i.e. 24 versus 5). As a result, elite reviewers writes about 32.5% of all reviews. Comparing elite and non-elite reviewers, they are similar in average rating per review (both

[23] Reviews identified by Yelp as fake reviews are removed from the Yelp pages. We do not observe these reviews and do not consider them in our analysis.

[24] We can potentially predict the elite status using past activities on Yelp of a reviewer, but since we do not observe how many rating the sample reviewers have left outside Seattle, we cannot reliably predict elite status.

around 3.7 stars), but elite reviewers have higher review frequency, a closer match with the restaurants they review, and slightly higher variety in taste.

Variance decomposition in Appendix Table D.1 shows that restaurant fixed effects alone account for 20.9% of the total variations in Yelp ratings, leaving 79.1% explained by within restaurant variations. Furthermore, a regression that incorporates both reviewer and restaurant fixed effects 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 some degree of match between reviewers and restaurants.

To check the difference between elite and non-elite reviewers, we first obtain residual $\widehat{\epsilon_{ri,yr}}$ after regressing observed ratings on reviewer, restaurant, and year fixed effects (i.e. $x_{ri,year} = \mu_r + \alpha_i + \gamma_{year} + \epsilon_{ri,year}$), and then associate $\widehat{\epsilon_{ri,yr}}^2$ with whether the review is written by an elite reviewer and the order of a review (N_{ri}) (i.e. $\widehat{\epsilon_{ri,yr}}^2 = \beta_0 + \beta_1 D_{ri,elite} + \beta_2 N_{ri} + \beta_3 N_{ri} \times D_{ri,elite} + \zeta_{ri}$). As shown in Appendix Table D.2, the coefficient of the elite dummy is significantly negative, suggesting that elite reviews deviate less from the long-run average rating of the restaurant, probably because elite reviewers have more precise signals ($\sigma_e^2 < \sigma_{ne}^2$) or have more social motives to conform to the crowd on Yelp ($\rho_e > \rho_{ne}$). We also examine the kernel density of a rating's deviation from the restaurant's average rating beforehand and afterward for elite and non-elite reviewers separately. We find that an elite reviewer tends to give a rating closer to the restaurant's average ratings before or after her, one phenomenon that is to be expected if elite reviewers have either more precise signal or correlate their ratings more positively to previous ratings.

We now present reduce-form evidence for review dynamics. 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 constant and reviewers incorporate a restaurant's previous reviews in a positive way ($\rho > 0$), we should observe reviews to vary less over time around the restaurant's fixed effect. This is confirmed by the negative coefficient on the order of review in Appendix Table D.2.

Positive social incentives 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 Appendix Table D.3, the residuals are positively correlated 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 incentive is not the only explanation for this pattern: a martingale evolution of restaurant quality could generate it as well.[25] It is up to the structural model to separate the effect of positive social incentives and quality evolution.

Furthermore, our data show that ratings within a restaurant tend to decline over time.

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

Appendix Figure E.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. More than one factor could contribute to this downward trend. Restaurant quality may decline over time, or 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, as suggested by Li and Hitt (2008).

If selection were the primary driver of the result in our setting, we would expect later reviewers to be a worse fit for the restaurant. To check this, we regress reviewer-restaurant matching distance ($MatchD_{rit}$) and reviewer's taste for variety ($TasteVar_{rit}$) on the order of reviews within a restaurant. As shown in Appendix Table D.4, within a restaurant, later reviewers tend to have less diverse tastes but are not significantly different from earlier reviewers in matching distance. While this suggests that later reviewers may be better sorted with the restaurant in terms of taste diversity, it does not explain why later reviewers tend to give worse ratings, unless less diverse diners are more critical. The last two columns of this table examine variations of $MatchD_{rit}$ and $TasteVar_{rit}$ within a reviewer, which turn out to be quite different from variations 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.

One may wonder whether different types of restaurants are subject to different review dynamics. We explore this potential distinction between ethnic and non-ethnic restaurants. A restaurant is defined as ethnic if its cuisine type is explicitly linked to a specific country or area outside of the US. Out of 4,101 restaurants in our sample, 742 are classified as ethnic according to Yelp classification of cuisine. Appendix Figure E.3 shows the polynomial smooth plot of within restaurant rating trends by ethnic and non-ethnic restaurants separately, after we take out restaurant fixed effects. It suggests faster decline and slower recovery in ratings for ethnic restaurants than non-ethnic restaurants. In light of this, one version of our model allows the within restaurant rating trend and the variance of new quality draws to differ for these two restaurant types.

Overall, reduce-form results yield six 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 attract reviewers with less diverse taste over time, reviewers tend to find restaurants better matched to them over time, and the dynamics differ between ethnic and non-ethnic restaurants.

4 Results from Structural Estimation

4.1 Main Results

As described in Section 2, 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 match between the reviewer and the restaurant. We allow these parameters to vary between elite and non-elite reviewers because elite reviewers are a central part of the review system, as documented in Section 3.

Table 2 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. The social weight estimate is statistically different from zero, suggesting that reviewers incorporate the content of previous reviews. As shown in the simulation section, this will cause later reviews to receive more weight than early reviews in the adjusted average.

In 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 differences between elite and non-elite reviewers.

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 significantly negative, suggesting that a non-elite reviewer tends to deviate from the mean perspective of the crowd before her, after we allow positive autocorrelation across reviews from restaurant quality change. One potential explanation is that non-elite reviewers deliberately differentiate from previous reviews because they enjoy expressing a different opinion or believe differentiation is the best way 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. Without further information, it is difficult to distinguish these possibilities. Compared to the non-elite reviewers, elite reviewers are found to be more positively influenced by the past crowd. 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 (2) and Column (3) favors the inclusion

of restaurant quality change.

In addition to restaurant quality change, Column (4) allows reviewer stringency to vary by Age_{rt}(in linear and quadratic terms), $MatchD_{rit}$, $TasteVar_{ri}$, $NumRev_{it}$, $RevFreq_{it}$, 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 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 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 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, we estimate models that allow restaurant quality to evolve by month, quarter, and half-year. As shown in Appendix Table D.5, the main changes occur in the estimates for σ_e, σ_{ne}, σ_ξ, ρ_e, and ρ_{ne}. This is not surprising because they are all identified by the variance-covariance structure of reviews within a restaurant. Nevertheless, we can 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. While likelihood suggests that the raw data are better explained by more frequent changes of restaurant quality, the difference between elite and non-elite reviewers remains similar across the three columns in Appendix Table D.5.

Table 3 follows the specification of Table 2 Column (4), but allows ethnic and non-ethnic restaurants to differ in σ_e, σ_{ne}, σ_ξ, Age_{rt} , and Age_{rt}^2. As reported in Table 3, the estimates of σ_ξ are statistically different from each other. Though the estimates of σ_e, σ_{ne}, and the coefficients are not statistically different, the point estimates show that the downward rating trend on a non-ethnic restaurant flattens out more quickly than on an ethnic restaurant. This is consistent with the raw data. According to the Akaike information criterion (Akaike 1974), the model with the ethnic-non-ethnic distinction fits the data better. In light of these results, all of our real-data-based counterfactual simulations use estimates from Table 3.

4.2 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. To address the concern, we estimate an alternative model in which we

assume that reviewers have limited attention 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 has limited attention, the optimal rating 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 Appendix Table D.6, we compare the MLE result and log likelihood of the Bayesian and limited attention models, while allowing restaurant quality to update by quarter. The two models generate similar results: signals are less noisy for elite reviewers than for non-elite reviewers, elite reviewers demonstrate more positive social incentives, there is a significant noise in quality change per quarter, individual ratings trend downwards within a restaurant, and reviewer heterogeneity has a significant influence on ratings. While the signs and statistical significance of all coefficients are the same and the magnitudes of most coefficients are similar across the two models, the social weight reviewers put on past ratingsdiffer in magnitudes. In particular, the positive social weight of elite reviewers is higher in limited attension model (0.0721 vs. 0.0122) and the negative social weight of non-elite reviewers is much closer to zero (-0.0072 vs. -0.1221). This is not surprising because the limited attention model changes the way that past ratings enter into a reviewer's optimal choice of rating. The log likelihood is lower in the limited attention model, suggesting that our Bayesian model fits the data better. According to the Akaike information criterion (Akaike 1974), the Bayesian model is $46,630$ times as probable as the limited attention model to minimize the information loss, if we assume quality updates by quarter.[26]

5 Counterfactual Simulations

This section presents two sets of counterfactual simulations. The first set highlights the role of each modeling element in the adjusted aggregation of simulated ratings. Using real (instead of simulated) ratings, the second set compares the adjusted average ratings, as determined by our algorithm, to the arithmetic average ratings currently presented on Yelp.

[26]Specifically, we have $exp((AIC_{Bayesian} - AIC_{LimitedAttention})) = exp((logL_{Bayesian} - logL_{LimitedAttention})/2) = 46,630$.

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? We analyze this question through a series of counterfactual simulations.

Recall that the simple average is an unbiased and efficient summary of restaurant quality, if reviewer signals are i.i.d., restaurant quality is stable, and there is no reviewer social weight or bias. We take this condition as the benchmark, and then add each variation separately to the benchmark. Our simulation compares the adjusted and simple averages against the simulated true quality.

The first model variation allows reviewers to put non-zero weight on previous reviews. When social incentive is the only deviation from the i.i.d. assumption, 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 adjusted average should give more weight to later reviews. Appendix Figure E.4 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 $\sigma_\epsilon = 1$.[27] In both cases, adjusted average is more efficient than simple average, and the efficiency improvement is greater if reviewers are more socially concerned. However, efficiency gain over simple average is small even if the social weight is as large as $\rho = 0.6$. Recall that our structural estimate of ρ never exceeds 0.25, which suggests that the efficiency gain from accounting for ρ in the adjusted average is likely to be small in the real data. Similar logic applies if later reviews put negative weight on past reviews. In that case, adjusted weighting should give less weight to early reviews. The simulated figure that compares adjusted average with simple average is very similar, with adjusted 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 disallow reviewers to differ in stringency or restaurants to change quality, an arithmetic average is going to be unbiased but inefficient. As shown in Appendix Figure E.5, the more precise the elite reviewers' signals are relative to other reviewers, the larger the efficiency gain is for adjusted average versus simple average. This is achieved by the adjusted aggregation assigning more weight to elite reviews.

The third model variation adds restaurant quality evolution to the benchmark. Unlike the first two deviations, 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 Appendix Figure E.6: the first two allow different standard deviation in the noise of quarterly quality update, while the third one allows restaurant quality to update monthly with the same σ_ξ as in the second graph. Review

[27] We create these figures by simulating a large number of ratings according to the underlying model, and then computing adjusted versus simple average of ratings at each time of review.

frequency is simulated as one review per month. Comparison across the three graphs suggests that the adjusted average rating 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 adjusted rating becomes stabilized at 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, as shown in the top-left graph of Appendix Figure E.6. This is because the average rating is meant to the capture the "current" restaurant quality at the time of aggregation. The adjusted 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 is further away from the true "current" quality as N increases.

To illustrate the magnitude of bias of adjusted and simple average in one realized path of quality, Figure 1 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, we fix review frequency at 4.5 days per review. As shown in the figure, adjusted average tracks the actual quality change better than simple average.

Appendix Figure E.7 highlights the importance of reviewer stringency (and its heterogeneity). Compared to the benchmark, 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 Appendix Figure E.7. Reviewer and restaurant characteristics are simulated using their empirical distribution as observed in the raw data. The first graph of Appendix Figure E.7 assumes that the reviewer bias changes with restaurant age, but the restaurant quality does not. The second graph assumes that the reviewer bias does not change with restaurant age, and only the restaurant quality does. Both graphs show that adjusted 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.[28]

5.2 Adjusted Versus Simple Average for Real Yelp Data

We now compare adjusted and simple average based on real Yelp ratings as observed in our data. According to Table 3, the noise of quality update (σ_ξ) has a standard deviation of 0.12 per quarter for ethnic restaurants and 0.13 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

[28] 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. Adjusted average always corrects any bias in reviewer bias, and simple average always reflects the sum of the changes in quality and reviewer bias.

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 adjusted average ratings: quality update implies that adjusted average needs to give more weight to recent reviews. 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 adjusted 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 incentives, but the absolute magnitudes of ρ_e and ρ_{ne} never exceed 0.25, suggesting that the efficiency gain from social incentives is likely to be small as well.

Including all these elements, we use the structural estimates to compute simple and adjusted average ratings at the time of every observed rating. We then calculate the difference between simple and adjusted average, $\mu_{rn}^{simple} - \mu_{rn}^{optimal}$ for every observation and summarize it in the first row of Table 4.

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 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 for a non-ethnic restaurant and -0.48 lower for an ethnic restaurant. Overall, we find that adjusted 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 adjusted average is asymmetric: simple average is more likely to underreport than overreport, as compared to adjusted average. We believe this is because adjusted average puts more weight on later reviews, and later reviews entail a greater correction of bias than earlier 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 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 adjusted average when we interpret the restaurant age effect as true quality declining over time. In reality, we believe that the trends are a combination of chilling effect and true quality change, so the simulations from our two model interpretations are likely to bound the comparisons.

The remainder of Table 4 compares our adjusted rating to 6-month, 12-month, and 18-month

moving average of reviewer ratings. No matter how we interpret the downward trend, the moving averages perform worse than our adjusted aggregating rating. This is probably because many restaurants receive sparse reviews and the short window of moving averages excludes many reviews that could be useful for the aggregation. This is mainly due to the fact that the median restaurant in our sample receives only one review permonth, and the average time gap between the first and second review is 154 days. For restaurants that are reviewed infrequently, the moving average only averages the few most recent ratings but throws out information embodied in all past ratings. This drastically reduces the information used, and hence is even further away from our adjusted average than the simple average.

Table 5 describes how the difference between simple and adjusted averages varies 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 it is interpreted as reviewer bias only, we find that, by the end of the sample, Yelp's simple average ratings differ from our adjusted average by more than 0.15 stars for 41.38% of restaurants and by more than 0.25 stars for 19.1% of restaurants. . If the above chilling effect is interpreted as changes in true quality, the absolute difference between simple and adjusted 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 4? This is because Table 4 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 4 calculate the rating difference for reviews 0-2 years, 2-4 years, and more than 4 years from the first review of a restaurant. No matter how we interpret the chilling effect, the difference between simple and adjusted ratings grows rapidly as a restaurant accumulates more reviews over time. As illustrated in Figure 1 and Appendix Figure E.6, 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 adjusted rating but missing in the simple average rating.

The increasing divergence of simple versus adjusted ratings can be better shown graphically. Based on the above-estimated difference between simple and adjusted average per observation, Figure 2 plots the mean and the 10th and 90th percentile of this difference by the order of reviews. Assuming the restaurant age effect as reviewers become more stringent over time (i.e. the chilling effect), the upper-left graph shows that the the restaurant quality is overestimated in the beginning. When we assume away chilling effect, the lower-left graph shows that the simple average rating is on average close to adjusted average, but the absolute values of the 10th and 90th percentile differences grow 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 adjusted average rating. The bar chart on the upper right plots the

histogram of restaurants by this percent. For example, the second bar shows that roughly 200 restaurants (out of 3,345 restaurants that have more than two reviews) have 5-10% of the time when its simple average ratings is more than 0.15 stars away from the adjusted ratings. Overall, over 1,271 restaurants have over 30% of the time when the simple average ratings are more than 0.15 stars away from the adjusted average. This suggests that adjusted average rating is likely to generate substantial improvement over simple average, especially as Yelp accumulates more reviews for each restaurant. The bottom two graphs of Figure 2 lead to a similar conclusion, but of smaller magnitude, when we interpret the restaurant age effect as true quality changes.

In Appendix Table D.7, we examine what restaurant and reviewer attributes lead to a greater difference between the adjusted and simple averages. In particular, we compare the mean attributes of restaurants and reviewers by whether the simple-vs-adjusted difference is greater or smaller than 0.15 stars. We find that restaurant review frequency, reviewer review frequency, and matching distance matter the most.

Overall, in the Yelp setting, the difference between adjusted and simple average is mostly driven by restaurant quality updates (σ_ξ) and time trend ($Age_{rt} \cdot \alpha_{age}$), and less by social incentives (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 adjusted average as each restaurant accumulates reviews over time.

6 Results from the Online Survey

The counterfactual simulation presented above is conditional on our structural model, thus one may argue that simulation alone does not prove that the advantage of our adjusted average is independent of model assumptions. To address this concern, we conducted an online survey using Amazon Mturk ("Restaurant Reviews Beliefs Survey") on February 1, 2016. In this survey, we asked how respondents use and comprehend restaurant ratings online (without mentioning Yelp in the survey). In total, 239 Mturk workers responded to our survey. The exact questionnaire is presented in Appendix F and the survey answers are summarized in Table 6.

Results show that the rationale of our adjusted aggregation is consistent with reported user preferences. In particular, nearly 60% of respondents use restaurant reviews at least once a week and 81.2% report that they rely on online reviews "frequently" or "sometimes" when choosing restaurants. When they use restaurant reviews, 93.7% pay attention to the average rating, but only 56.9% look at the number of reviews and 33.7% look at changes in the ratings. This confirms our motivation to generate one informative aggregate rating. When asked whether to take into account the fact that some reviews are older than others, 86.6% prefer more weights on more recent reviews. In comparison, when asked about whether to account for the completeness of the reviewer profile, 69.9% do not take reviewer profile into account, while 26.4% put more weight on reviewers with a complete profile. As shown in the second column of Table 6, results

are similar if we restrict the sample to only those that rely on restaurant reviews "frequently" or "sometimes" when they choose restaurants.

Overall, the survey results are consistent with the rationale that more weights should be given to more recent reviews and reviews written by more seasoned reviewers. Since users tend to pay much more attention to the average rating, it is important to construct the aggregated rating in a way that systematically reflects the informativeness of various reviews. This is exactly why our model incorporates a list of cross-sectional and dynamic factors, allowing Yelp review data to identify the relative importance of each. The external-validity check of the survey is completely independent of our structural model and our Yelp data, thus its consistency extends support to our model.

7 Conclusion

As consumer reviews continue to proliferate, 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 variations 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 our adjusted average 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 problems of simple averaging of ratings over time. Yet, this intuition can be wrong especially when quality changes over time (e.g. a new chef, a different menu, etc.) By moving toward more systematic aggregation, the market designer can detect such changes and be wary of the deviation between true quality and the arithmetical average.

As acknowledged before, one major caveat of our paper is its inability to model reviewer selection explicitly, mostly due to data limits. That being said, Section 2 has elaborated on how our model partially controls for the selection, so we will not repeat it here. Nor will we repeat how robust our findings are to the assumption of reviewer rationality, which has been addressed in Section 4.2. Besides these two points, we now discuss the remaining limitations of our approach, the potential uses of our algorithm, and a few directions for future research.

Incentives to Write Reviews Our paper has focused on taking an existing set of reviews and adjusting the aggregating method to better reflect the quality of a product. An alternative mechanism to achieve this goal is to use social image to encourage people to leave more representative reviews. There is a large theoretical literature studying social image (Akerlof 1980, Bénabou and Tirole 2006). Theoretically modeling a crowdsourced setting, Miller, Resnick and Zeckhauser (2005) argue 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. Our finding is consistent with Wang (2010), who 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 reviews, reviewers are less likely to give extreme ratings, and more prolific Yelp reviewers have more friends, receive more anonymous review votes per review, and display more compliment letters per review. Comparing the same restaurants listed on both platforms, he finds that restaurants are less likely to receive extreme ratings on Yelp. These findings motivate us to explicitly allow elite and non-elite reviewers to place different weight on social incentives. That being said, social incentives in our model can have multiple interpretations and we do not model one's incentive to manipulate reviews for popularity.[29]

Note that our model does not completely ignore strategic incentives in rating. For example, elite reviewers may care more about their own reputation on the rating platform and therefore have an incentive to submit a review that better reflects their prediction of readers' taste rather than their own taste. In contrast, non-elite reviewers may prefer to vent out their own experience with little regard for how readers may react to their reviews. We have incorporated these considerations in the model and let the data identify the extent to which elite and non-elite reviewers incorporate their own experience. Similarly, reviewers that have reviewed different restaurants may have different incentives or tastes to rate the current restaurant more or less favorably. Although we cannot separate incentives from tastes, we attempt to control for the influence of a reviewer's review history on her rating behavior, which should in turn capture part of the strategic incentives that may arise from differential review history.

Fake Reviews One potential problem for consumer review websites is fake or promotional reviews. Mayzlin, Dover and Chevalier (2014) have documented evidence for review manipulation on hotel booking platforms (i.e. 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 (2016). While review filters can help to eliminate gaming, there are surely still erratic and fake reviews that get through the system. In Appendix Figure E.8, we simulate the evolution of ratings in 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

[29]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.

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 review. All the other reviews are simulated in a large number according to the underlying model. We then plot true quality, simple average, and adjusted average by order of review. The top graph shows that, if the outlier review is the first review, over time adjusted average has a better ability to shed the influence of this outlier review, because it gives more weight to recent reviews. The bottom graph suggests that adjusted average is not always the best; because it gives more weight to recent reviews, so it gives more weight to the outlier review right after it has been submitted (the dip of the adjusted average after the fifth review is greater than the simple average). However, for the same reason, adjusted average also forgets about the outlier review faster than simple average, and better reflects true quality afterward.

Transparency and Aggregation Decisions Part of the motivation for this paper is that on almost every consumer review website, reviews are aggregated. In practice, the most common way to aggregate reviews is using an arithmetic average, which is done by 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. After this paper became public, some platforms switched to alternative approaches to aggregating ratings. For example, Amazon now uses machine learning to aggregate ratings rather than relying on a simple arithmetic average, in an attempt to capture more information from the ratings.

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 adjusted 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 adjusted aggregation against other incentives that may move them away from this standard.

Potential Uses of Our Approach The aggregation of ratings has implications beyond shaping what metric is shown to consumers. For example, Yelp uses the average rating, among other factors, to determine the order in which businesses are shown. Our adjusted average could help to improve this ordering. The adjusted average can also be used to form personalized restaurant recommendations to consumers. In addition, it can be presented to business owners and managers to help them understand how much of the rating reflects vertical quality, and how much is influenced by reviewer type and reviewer behavior.

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 variations, 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.

Ariely, Dan, Anat Bracha and Stephan Meier (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.

Wu, C., Che, H., Chan, T.Y., Lu, X. (2015) The Economic Value of Online Reviews. Marketing Science, 34(5),739-754.

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.

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.

Eyster, Erik., and Rabin, M. (2010). Naïve Herding in Rich-Information Settings. American Economic Journal: Microeconomics, 2(4): 221-243.

Fradkin, Andrey, Elena Grewal and David Holtz (2017). "he Determinants of Online Review Informativeness: Evidence from Field Experiments on Airbnb." working paper.

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

Godes, David, and José C. Silva (2012). "Sequential and temporal dynamics of online opinion." *Marketing Science* 31, no. 3: 448-473.

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.

Ljungqvist, Lars and Thomas J. Sargent (2012). *Recursive Macroeconomic Theory.* MIT Press, 3 edition, 2012.

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 (2016). "Fake it till you make it: Reputation, competition, and Yelp review fraud." *Management Science*.

Mayzlin, Dina, Y. Dover and Judy A. Chevalier (2014). "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.

Moe, Wendy W. and Michael Trusov (2011). "The value of social dynamics in online product ratings forums." *Journal of Marketing Research* 48, no. 3: 444-456.

Moe, Wendy W. and David A. Schweidel (2012). "Online product opinions: Incidence, evaluation, and evolution." *Marketing Science* 31, no. 3: 372-386.

Muchnik, Lev; Sinan Aral and Sean J. Taylor (2013) "Social Influence Bias: A Randomized Experiment", *Science*, 9 August 2013: Vol. 341 no. 6146 pp. 647-651.

Nosko, Chris and Steven Tadelis (2015). "The Limits of Reputation in Platform Markets:An Empirical Analysis and Field Experiment." *NBER Working Paper* No. 20930, January 2015.

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.*

Welch, Greg, and Gary Bishop (2001). "An introduction to the Kalman filter." *Proceedings of the Siggraph Course, Los Angeles.*

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. The reviews 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: **MLE: Signal Precision, Social Incentives 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
N	133,688	133,688	133,688	133,688

Panel B. Stringency 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}	$(\lambda_e-\lambda_{ne})_{Age^2}$	1×10^{-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)

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, social incentives, and biases in judging restaurants' quality. Model in column (2) allows reviewers' precision, popularity concerns, 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 3: **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)
$(\mu+\lambda_{ne})_{Age^2}$	5.28×10^{-6}	2.61×10^{-6}
	(2.69×10^{-6})	(4.71×10^{-6})
Other parameters		
$(\mu+\lambda_{ne})_{MatchD}$	0.0370	$(\lambda_e-\lambda_{ne})_{Age^2}$ 1.02×10^{-5}
	(0.0040)	(2.81×10^{-6})
$(\mu+\lambda_{ne})_{TasteVar}$	-0.2551	$(\lambda_e-\lambda_{ne})_{MatchD}$ -0.0031
	(0.0354)	(0.0051)
$(\mu+\lambda_{ne})_{NumRev}$	-0.0061	$(\lambda_e-\lambda_{ne})_{TasteVar}$ -0.0252
	(0.0010)	(0.0767)
$(\mu+\lambda_{ne})_{FreqRev}$	0.0246	$(\lambda_e-\lambda_{ne})_{NumRev}$ 0.0041
	(0.0397)	(0.0014)
$(\lambda_e-\lambda_{ne})_0$	0.0171	$(\lambda_e-\lambda_{ne})_{FreqRev}$ -0.0575
	(0.0233)	(0.0535)
$(\lambda_e-\lambda_{ne})_{Age}$	-0.0003	
	(0.0002)	
Log Likelihood	-191,758.9	
N	133,688	

Notes: 1. Estimated model in this table adds to the baseline model shown in Table 2 Column(4) 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.

Table 4: Simple, Moving and Adjusted 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 adjusted average exceeding 0.15 and 0.25. The calculation of adjusted ratings is based on the model differentiating ethnic and non-ethnic restaurants, and with quality change every quarter. 2. The adjusted rating is calculated for every review written on restaurants with at least 3 reviews. This covers 3,345 restaurants and 133,668 ratings. 3. Panel B rounds the simple average and adjusted averages to every 0.5 points.

Table 5: Simple and Adjusted 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 adjusted 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.

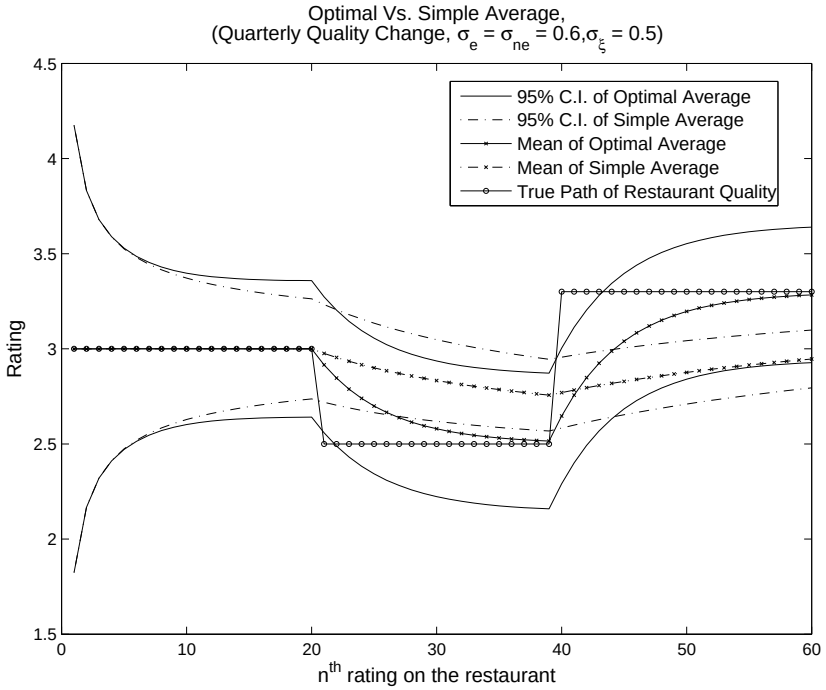
Table 6: Survey Results

	All Respondents	Review Users[a]
Q1 How often do you go to restaurants?		
Less than once per month	8.00%	5.70%
About once per month	33.10%	32.50%
About once per week	48.10%	51.00%
Multiple times per week	10.90%	10.80%
Q2 When choosing restaurants, do you rely on online reviews?		
Frequently	19.70%	24.20%
Sometimes	61.50%	75.80%
Rarely	17.60%	–
Never	1.30%	–
Q3 When looking at reviews to choose a restaurant, what factors do you take into account? *(Choose all that apply)*		
The number of reviews	56.90%	58.20%
The average rating	93.70%	95.90%
Changes in the rating (improvements or declines over time)	34.70%	37.60%
Others	12.60%	12.90%
Q4 When looking at reviews, do you take into account the fact that some reviews are older than others?		
I put more weight on recent reviews	86.60%	89.20%
I put more weight on older reviews	2.50%	2.60%
I don't take the age of the review into account	10.90%	8.30%
Q5 When looking at reviews, do you take into account the completeness of the reviewer profile?		
I put more weight on reviews by reviewers with more complete profiles	26.40%	29.40%
I put less weight on reviews by reviewers with more complete profiles	3.80%	4.10%
I don't take the completeness of the reviewer's profile into account	69.90%	66.50%

Notes: a. The second column summarizes results from "Review Users" - those who chooses "Frequently" or "Sometimes" in Q2 "*When choosing restaurants, do you rely on online reviews?*".

Figure 1: **How Quickly Do Average Ratings Adjust to Changes in Restaurant Quality?**

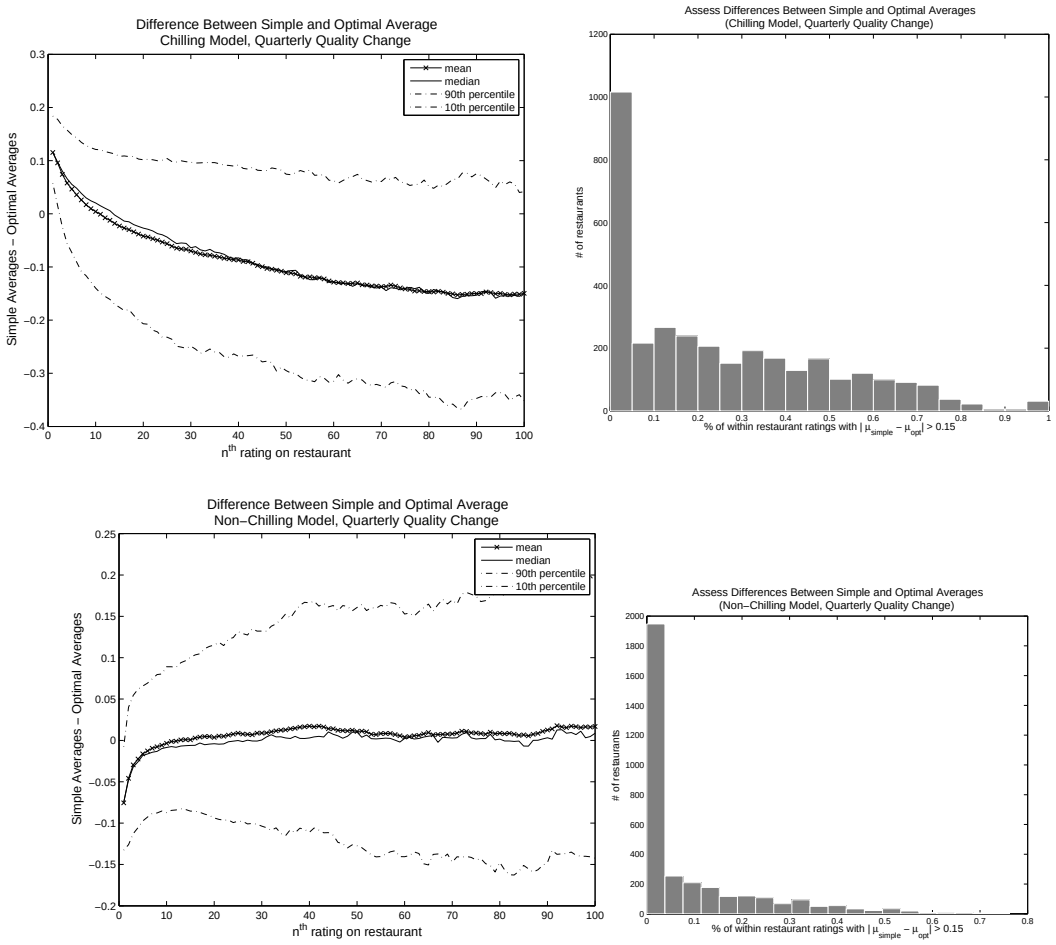
ρ	σ	*Quality Update Frequency*	*StdDev of* $\Delta_{Quality}$
$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.6$	Quarterly	$\sigma_\xi = 0.5$



Notes: The above figure plots 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. Adjusted aggregation adapts to changes in restaurant's true quality, while simple average becomes more biased in representing restaurant's true quality. Since the adjusted aggregation algorithm only gives weights to recent ratings and simple average gives equal weights to all past ratings, standard error of adjusted average shrinks slower than simple average.

Figure 2: Adjusted and Simple Average Algorithms 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.

Appendices for Aggregation of Consumer Ratings: An Application to Yelp.com

Appendix A: Model of Reviewer Incentives to Deviate From Prior Reviews

In this appendix, we show an alternative model to capture reviewer incentive to differentiate from prior ratings corresponding to our baseline model in Section 2.1. It gives rise to exactly the same equation except for $\rho_i < 0$.

If social incentives motivate reviewer i to deviate from prior reviews, we can model it as reviewer i choosing to report x_{rt_n} to minimize a slightly different 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 ratings (not counting i's own signal) and $w_i > 0$ is the marginal utility that reviewer i will get by reporting differently from prior ratings. 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 ratings in the above two scenarios are written in exactly the same expression except that $\rho_i > 0$ if one tries to be close to the best guess of the true restaurant quality in her report and $\rho_i < 0$ if one is motivated to deviate from prior ratings. 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 rating correlates with past ratings. As long as later ratings contain information from past ratings, aggregation needs to weigh early and late reviews differently.

Appendix B: Notes on the Data Generating Process

B.1 Data Generating Process

The model presented in section 2.1 includes random change in restaurant quality, random noise in reviewer signal, reviewer heterogeneity in stringency, social incentives, 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 random walk 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 rating that gives weights to both her own experience and her social incentives. The optimal rating 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 B.2.

B.2 Deriving $E(\mu_{rt}|s_{rt_1},...s_{rt_n})$

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

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

and we assume that the first reviewer uses an uninformative prior

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

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

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

Hence

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

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

In the above formula, the prior belief of μ_{rt_n} given the realization of $\{s_{rt_1},...,s_{rt_{n-1}}\}$, or $\pi_{n|n-1}(\mu_{rt_n})$, depends on the posterior belief of $\mu_{rt_{n-1}}$, $h_{n-1|n-1}(\mu_{rt_{n-1}})$ and the evolution process from $\mu_{rt_{n-1}}$ to μ_{rt_n}, denoted as $g(\mu_n|\mu_{n-1})$. Hence,

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

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

$h_{n|n}(\mu_{rt_n})$, the mean and variance evolves according to the following rule:

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

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

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

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

B.3 The Correlation of Ratings Induced by Quality Change

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 $\xi_t \sim i.i.d\ N(0, \sigma^2_\xi)$. 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 $\{s_{rt_1}, ..., s_{rt_{n-1}}\}$. These two features greatly facilitate reviewer n's Bayesian estimate of restaurant quality.

Appendix C: Deriving the Likelihood Function

C.1 Deriving the Likelihood Function $f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}})$

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 $f(s_{rt_2} - s_{rt_1}, ..., s_{rt_{N_r}} - s_{rt_{N_r-1}})$,

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

The derivation of $f(x_{rt_2} - x_{rt_1}, ..., x_{rt_{N_r}} - x_{rt_{N_r-1}})$ is shown as the following,

- 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 B.2), we can solve $\{s_{rt_1}, ...s_{rt_n}\}$ from $\{x_{rt_1}, ...x_{rt_n}\}$ according to the recursive formula in Appendix C.2.

- 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}.

C.2 Solving $\{s_{rt_1}, ...s_{rt_n}\}$ from Observed Ratings

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

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

Appendix D: Tables

Table D1: **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 D2: **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)
$D_{ri}^{elite} \times N_{ri}$(100$s$)	-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 D3: **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 D4: **Does matching improve over time?**

	For Restaurants		For Reviewers	
	(1) Matching Distance[a]	(2) Taste for Variety[b]	(3) Matching Distance[a]	(4) 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 D5: **Baseline Model with Different Quality Update Frequency Assumptions**

	(1)	*(2)*	*(3)*
Quality Change	*Quarterly*		
Quality Change	*Half-yearly*		
Quality Change	*Monthly*		
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 2 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 D6: **Estimation Results: Limited Attention Model Vs. Bayesian Rational Model**

	(1) *Limited Attention* *Quarterly Quality Change*	*(2)* *Fully Rational Model* *Quarterly Quality Change*
σ_e	0.9617^{***}	0.9004^{***}
	(0.0196)	(0.0193)
σ_{ne}	1.0524^{***}	0.9514^{***}
	(0.0169)	(0.0150)
σ_ξ	0.1272^{***}	0.1323^{***}
	(0.0039)	(0.0038)
ρ_e	0.0721^{***}	0.0122^{***}
	(0.0181)	(0.0222)
ρ_{ne}	-0.0072^{***}	-0.1221^{***}
	(0.0157)	(0.0186)
$(\mu + \lambda_{ne})_{Age}$	-0.0031^{***}	-0.0032^{***}
	(0.0003)	(0.0003)
$(\mu + \lambda_{ne})_{Age^2}$	3.8×10^{-6}	4×10^{-6}
	(2.5×10^{-5})	(2.5×10^{-5})
$(\mu + \lambda_{ne})_{MatchD}$	0.0372^{***}	0.0367^{***}
	(0.0041)	(0.0040)
$(\mu + \lambda_{ne})_{TasteVar}$	-0.2547^{***}	-0.2453^{***}
	(0.0355)	(0.0354)
$(\mu + \lambda_{ne})_{NumRev}$	-0.0060^{***}	-0.0062^{**}
	(0.0010)	(0.0010)
$(\mu + \lambda_{ne})_{FreqRev}$	0.0241^{***}	0.0256^{***}
	(0.0398)	(0.03997)
$(\lambda_e - \lambda_{ne})_0$	0.0148	0.0161
	(0.0230)	(0.0233)
$(\lambda_e - \lambda_{ne})_{Age}$	-0.0004^{**}	-0.0002
	(0.0002)	(0.0002)
$(\lambda_e - \lambda_{ne})_{Age^2}$	$1 \times 10^{-5***}$	$1 \times 10^{-5***}$
	(2.7×10^{-6})	(2×10^{-6})
$(\lambda_e - \lambda_{ne})_{MatchD}$	-0.0015	-0.0028
	(0.0052)	(0.0053)
$(\lambda_e - \lambda_{ne})_{TasteVar}$	-0.0420	-0.0266
	(0.0757)	(0.0768)
$(\lambda_e - \lambda_{ne})_{NumRev}$	0.0035^{***}	0.0041^{***}
	(0.0013)	(0.0014)
$(\lambda_e - \lambda_{ne})_{FreqRev}$	-0.0576^{***}	-0.0556^{***}
	(0.0516)	(0.0535)
Log Likelihood	-191,791.7	-191,770.2
N	133,688	133,688

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

Notes: 1. Column (1) shows the estimation results of the limited attention model presented in Section 4.2, Column (2) shows the results of the baseline Bayesian belief model.

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

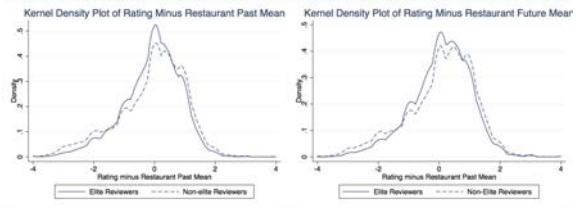
Table D7: Characteristics of Review with Different Simple and Adjusted Averages Gaps

	$\Delta < -0.15$	$-0.15 \leq \Delta < 0.15$	$\Delta > 0.15$
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 adjusted ratings greater and smaller than 0.15.

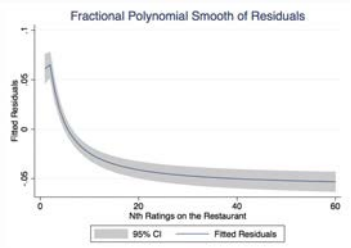
Appendix E: Figures

Figure E1: **Distribution of Ratings Relative to Restaurant Mean by Elite Status**



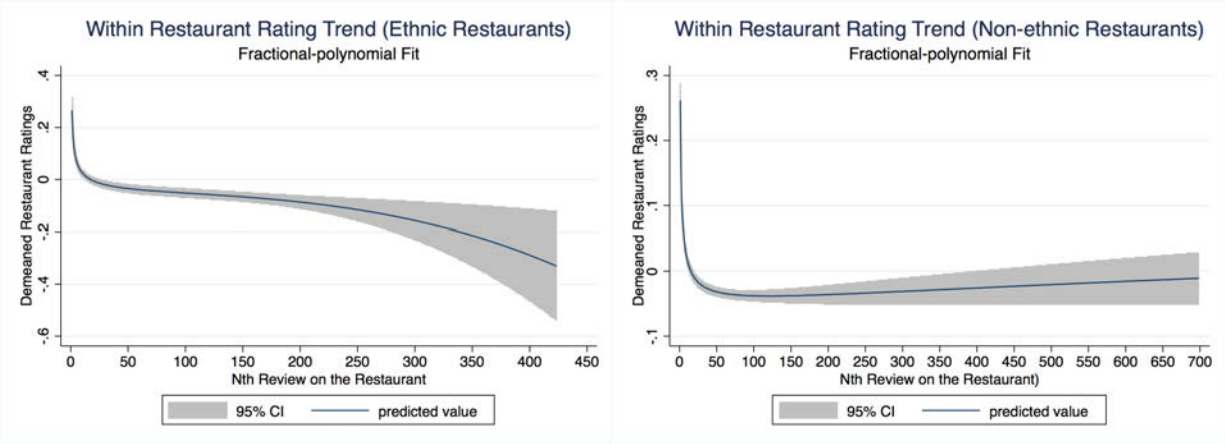
Notes: 1. The figure on the left plots the distribution $Rating_{rn} - \overline{Rating}_{rn}^{BF}$, where $Rating_{rn}$ is the n^{th} rating on restaurant r, and $\overline{Rating}_{rn}^{BF}$ is the arithmetic mean of past $n-1$ ratings on restaurant r before n. Similarly, the figure on the right plots 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 E2: **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 $Rating_{rn,year} = \mu_r + \gamma_{year} + \epsilon_{rn,year}$.

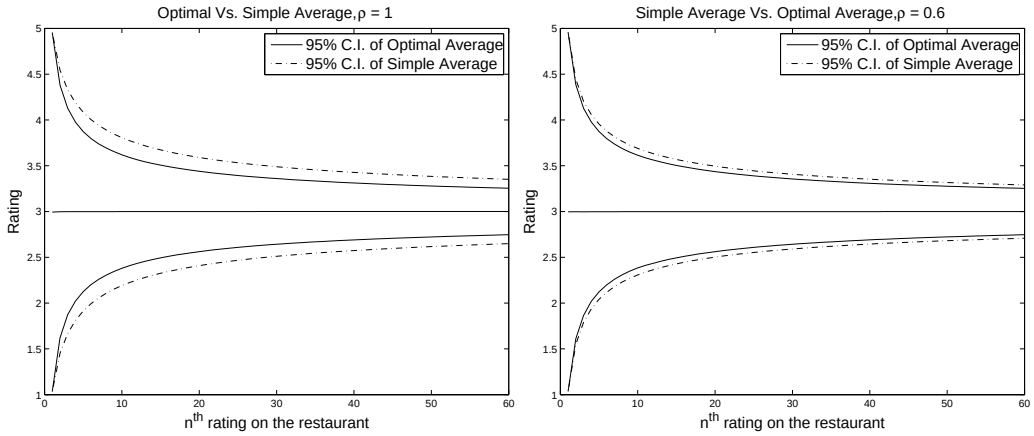
Figure E3: **Fractional-polynomial Fit of Within Restaurant Rating Trend (Ethnic Vs. Non-ethnic Restaurants)**



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 adjusted averages are unbiased estimates for true quality. But, relative to arithmetic average, adjusted 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 E4: **Adjusted and Simple Averages Comparison: Reviewers with Different Social Incentives**

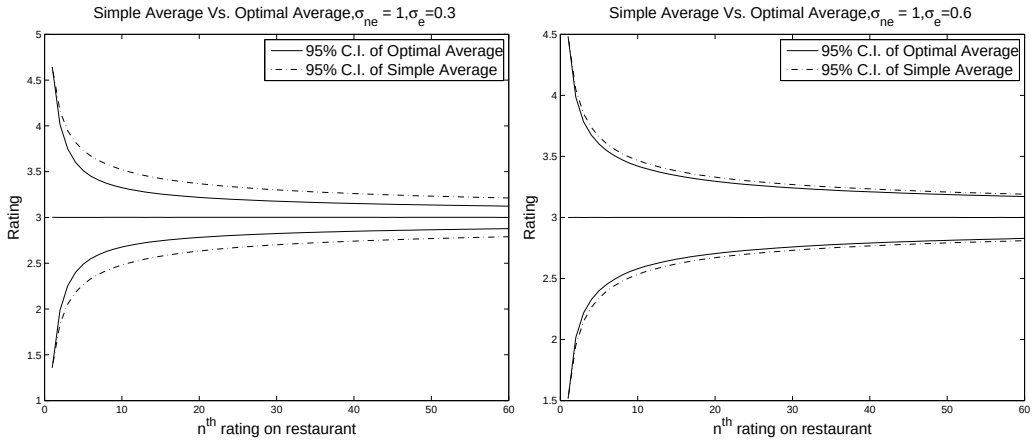
Parameters	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 1$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0.6$	$\sigma_e = \sigma_{ne} = 1$	Quality fixed at $\mu = 3$



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

Figure E5: **Adjusted and Simple Averages Comparison: Reviewers with Different Precisions**

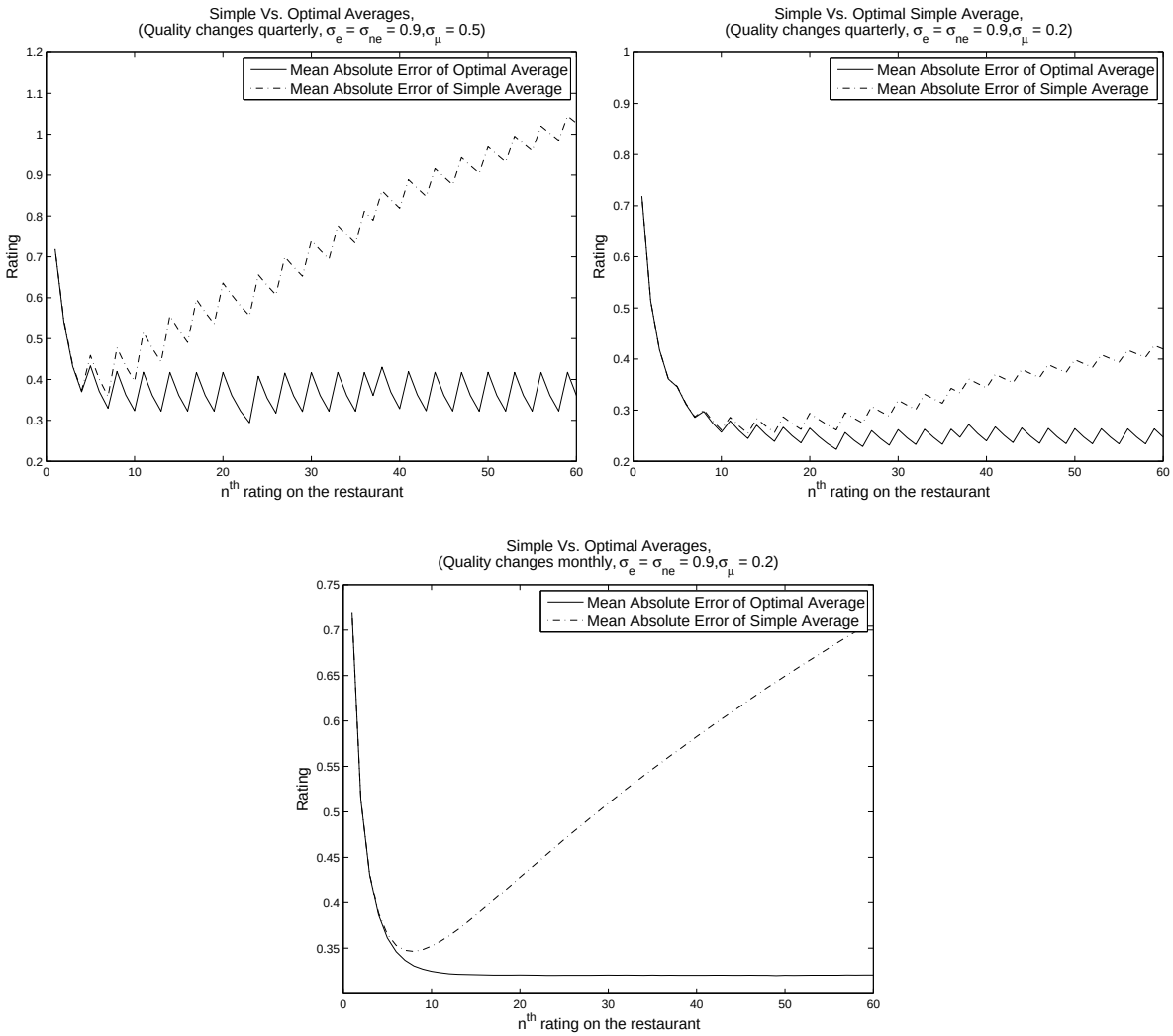
Parameter	ρ	σ	*Restaurant Quality*
(Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.6$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$
(Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = 0.3$,$\sigma_{ne} = 1$	Quality fixed at $\mu = 3$



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 adjusted averages are unbiased estimates for true quality. But, relative to arithmetic average, adjusted 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 E6: **Adjusted and Simple Averages Comparison: Restaurants with Quality Random Walk**

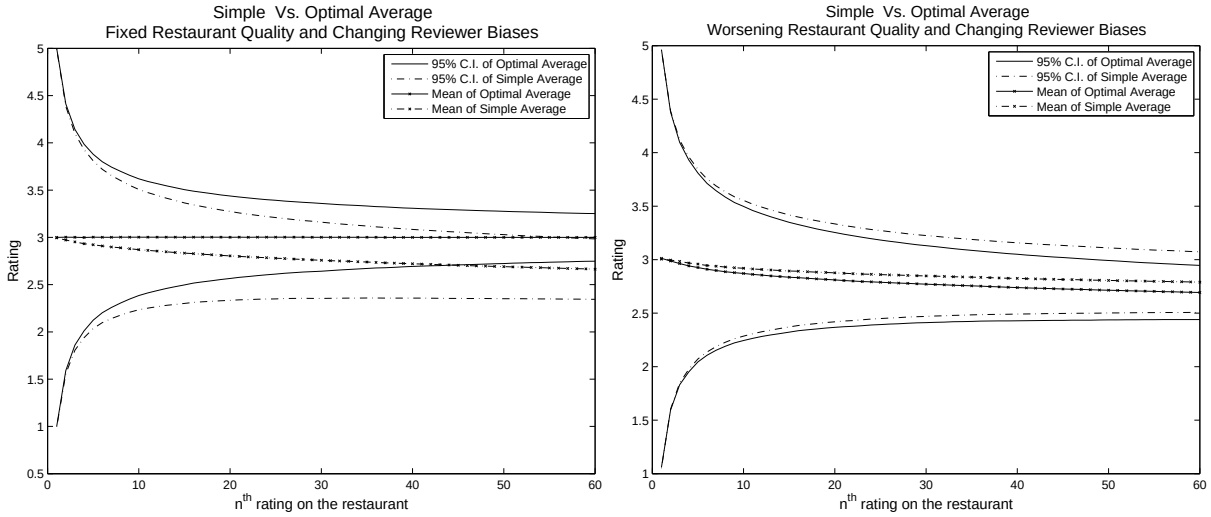
	ρ	σ	*Quality Update Frequency*	*StdDev of* $\Delta_{Quality}$
(Top Left)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.9$	Quarterly	$\sigma_\xi = 0.5$
(Top Right)	$\rho_e = \rho_{ne} = 0$	$\sigma_e = \sigma_{ne} = 0.9$	Quarterly	$\sigma_\xi = 0.2$
(Bottom)	$\rho_e = \rho_{ne} = 0$	$\sigma_e =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



Notes. 1. The above figures plot the mean absolute errors of adjusted 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 adjusted average keeps the same level of mean absolute error. The error of simple average is greater compared with adjusted average if a restaurant has larger variance in quality, and changes quality more frequently.

Figure E7: Simulations when Reviewers are Biased

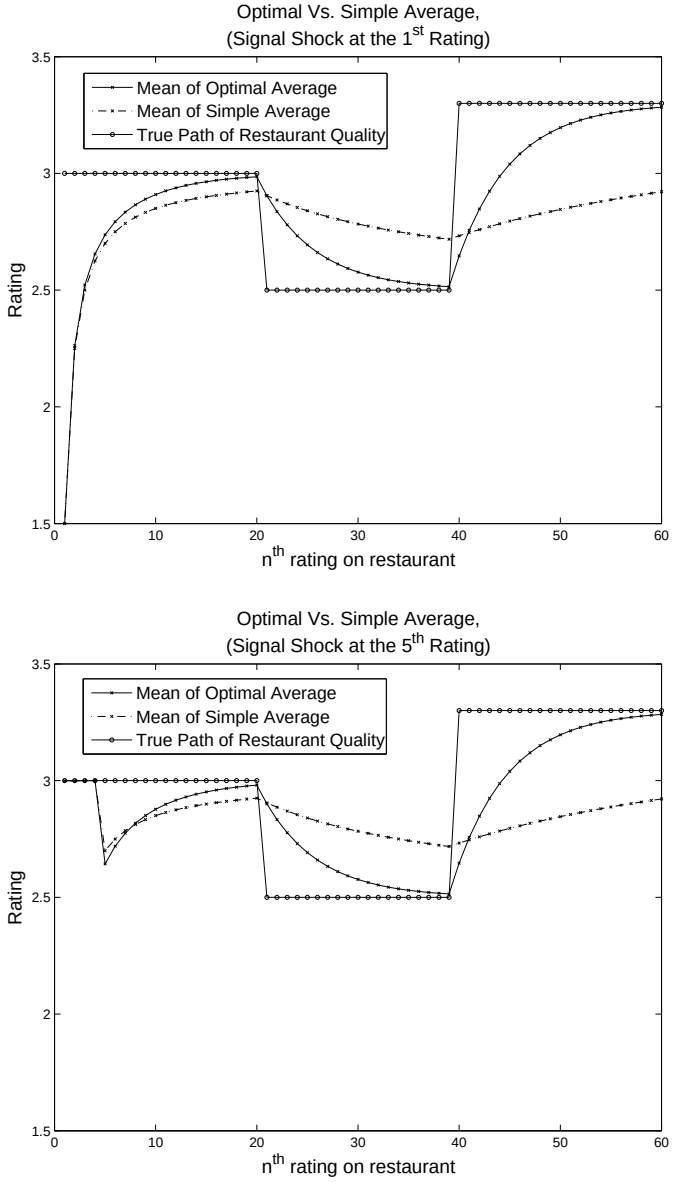
Parameters	Value	Parameters	Value	Parameters	Value
ρ_e, ρ_{ne}	0	$\frac{\partial(\mu+\lambda_e)}{\partial Restaurant\ Age^2}$	4×10^{-6}	$\frac{\partial(\mu+\lambda_e)}{\partial Reviewer\ Frequency}$	0.0256
σ_e, σ_{ne}	1	$\frac{\partial(\mu+\lambda_e)}{\partial Match\ Distance}$	0.0367	$\frac{\partial(\lambda_e-\lambda_{ne})}{\partial Restaurant\ Age^2}$	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, adjusted aggregation is an unbiased estimate for true quality while the simple average is biased without correcting the review bias.

Figure E8: **Adjusted and Simple Averages Comparison: "Fake" Review**

Signal	*Fixed* 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, adjusted aggregation "forgets" about earlier ratings and converges back to the true quality in a faster rate.

Appendix F: "Restaurant Reviews Beliefs Survey" Questionnaire

To test our model against external source of information, we conducted an online survey using Amazon Mturk ("Restaurant Reviews Beliefs Survey," February 1, 2016) in which we asked how respondents used and comprehended restaurant ratings online (we didn't mention Yelp in the survey). In total, 239 Mturk workers responded to our survey. The following shows the screen shot of the questionnaire.