

ABSTRACT

Title of dissertation: ESSAYS ON INFORMATION MANIPULATION
AND OPTIMAL DECISION MAKING

Gustavo Saraiva
Doctor of Philosophy, 2019

Dissertation directed by: Lawrence Ausubel
Department of Economics

This dissertation studies a variety of topics related to information manipulation, such as the manipulation of reviews in online rating platforms, or the act of misreporting one's preferences in matching mechanisms; and how those manipulations affect the overall allocation in the economy.

Chapter 1 analyses the incentives that drive some sellers to fake reviews in online rating platforms, such as Amazon and Yelp. Among other things, I find that sellers' optimal investment in fake reviews is not a monotone function of their reputation. More precisely, sellers that currently possess a very good or very bad history of past reviews have less incentives to solicit fake reviews praising their own products, the intuition being that, for sellers with very bad reputation, it is too costly to pretend that they are high quality sellers; while sellers that have already accumulated a very good reputation do not need to spend much effort in convincing buyers that they are high quality sellers. Moreover, in order to maximize the impact from each fake review, sellers tend to concentrate review manipulation at the initial

stages after they have entered the market.

Chapter 2 develops a theoretical model aimed at explaining the observed polarization on agents' beliefs regarding topics that have objective truths (e.g., such as whether or not global warming is a hoax). The main premises surrounding the model are that rational agents seek to learn the truth about a certain state of the world, but the acquisition of information is costly, and the available information channels are biased and imprecise. The paper vies to understand how the level of bias from those channels affect opinion polarization overall.

Chapter 3 analyses agents' incentives to misreport their preferences or vacancies in large stable matches. I find that, under certain assumptions, those incentives vanish for sufficiently large markets, suggesting that stable matching mechanisms are effectively strategy-proof for sufficiently thick markets.

ESSAYS ON INFORMATION MANIPULATION
AND OPTIMAL DECISION MAKING

by

Gustavo Saraiva

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Lawrence Ausubel, Chair
Professor Daniel Vincent
Professor Emel Filiz Ozbay
Professor Luminita Stevens
Professor Guodong Gao

ProQuest Number: 13895465

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13895465

Published by ProQuest LLC (2019). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

© Copyright by
Gustavo Saraiva
2019

Dedication

To my family

Acknowledgments

I am deeply indebted to my advisor, prof Lawrence Ausubel for his continuous support and guidance. Throughout my PhD I have had the great opportunity of taking some of his classes, which have greatly influenced the type of research that I do today. What I have always found most fascinating about his research field is its strong pragmatism and applicability. Since he became my advisor, Professor Ausubel has given me invaluable feedback on my research and guided me through a tough job market process, to which I am extremely grateful.

Along the different stages of my dissertation, professors Daniel Vincent, Emel Filiz-Ozbay, Luminita Stevens, Andrew Sweeting and Guido Kuersteiner have given me very insightful feedback, which have led to significant improvements on my dissertation. I am very grateful for the time they spent helping me. I also want to thank professor Gordon Gao for serving as the Dean's representative for my dissertation defense.

I have also benefited from comments and discussions with colleagues and Economists from outside the department. I'd like to thank in special to Paulo Saraiva and Ilton Soares for their invaluable feedback.

Last but not least, I want to thank my family for their love and support. This journey would not have been possible without them.

Table of Contents

Dedication	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	vi
List of Figures	vii
1 On Incentives to Manipulate Online Ratings	1
1.1 Introduction	2
1.2 Related Literature	5
1.3 Theoretical Model	11
1.3.1 Firm's profits as a function of its expected quality	12
1.3.2 Allowing the firm to manipulate customers' beliefs through fake reviews	14
1.3.2.1 Customers' Bayesian updates	15
1.3.2.2 Seller's optimal choice of review manipulation	16
1.3.3 Equilibrium	19
1.3.4 Finding the equilibrium numerically	20
1.3.5 When consumers do not anticipate review manipulation	26
1.3.6 Allowing the firm to exit and reenter the market with a new name	28
1.4 Empirical Strategy	34
1.4.1 Database	36
1.4.1.1 Fake review detection	41
1.4.1.2 Measuring reputation	44
1.4.1.3 Covariates	45
1.4.2 Logit model	50
1.4.3 Logit model correcting for classification error	54
1.4.4 Alternative database	56
1.4.4.1 Fake review detection	58
1.4.4.2 Logit Regressions	62
1.4.5 Placebo test	64
1.5 Conclusion	65

2	Opinion Polarization in the presence of noisy and biased channels	69
2.1	Introduction	70
2.2	Related literature	71
2.3	Model	76
2.3.1	Bayesian updates	81
2.4	Existence and uniqueness of a solution to the Bellman equation	82
2.5	Simulations	83
2.6	Allowing agents to fact check the news	84
2.7	Conclusion	86
3	An Improved Bound to Manipulation in Large Stable Matches	87
3.1	Introduction	87
3.2	General framework for the college admission problem	92
3.3	Strategic Manipulation	99
3.4	Large Markets	100
3.5	Equilibrium Analysis	109
3.6	Conclusion	112
A	Appendices for Chapter 1	114
A.0.1	Proofs	114
A.0.2	Jaccard similarity index	116
A.0.3	Variables	117
A.0.3.1	Naïve Bayes estimate of text reliability	117
A.0.3.2	Detecting anomalous peaks on the volume of 5 star reviews	119
B	Appendices for Chapter 3	121
B.1	Dropping strategies are exhaustive	121
B.2	Rejection Chains	125
B.3	Stochastic student-proposing DA algorithm and stochastic rejection chains	131
B.4	Simulated Incentives to misreport preferences or vacancies under different DGP	142
B.5	Equilibrium Analysis	144
	Bibliography	147

List of Tables

1.1	Number of products/reviews collected from each source	41
1.2	Regression Results	51
1.3	Logit regression after correcting for endogenous classification errors. .	56
1.4	Simple Logit regressions using the earphone dataset.	63
1.5	Logit regression after correcting for endogenous classification errors. .	64
1.6	Placebo Tests	66
1.7	Placebo test for the Logit regression, correcting for classification error.	67

List of Figures

1.1	Outline of the model.	12
1.2	Equilibrium as a function of μ , when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$	21
1.3	Average simulated effort of review manipulation chosen throughout the periods, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. η_H corresponds to the average effort of review manipulation chosen by a high type seller (i.e., a seller with $q = 1$), while η_L corresponds to the average effort chosen by a low type seller (i.e., a seller with $q = 0$).	23
1.4	The average simulated evolution of reputation, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. μ_H corresponds to the average simulated reputation from a high type seller (i.e., a seller with $q = 1$), while μ_L corresponds to the average simulated reputation from a low type seller (i.e., a seller with $q = 0$).	25
1.5	The average simulated evolution of reputation, when the seller is allowed and not allowed to fake reviews, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$	26
1.6	Equilibrium as a function of μ , when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$, when consumers know the strategy taken by the seller, and when consumers are naive and believe the seller does not engage in review manipulation (i.e., they believe $\eta(q, \mu) = 0$ for all q, μ).	28
1.7	The average simulated evolution of reputation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. μ_H corresponds to the average reputation from high quality sellers, while μ_L corresponds to the average reputation from low quality sellers.	29
1.8	The average simulated evolution of effort on review manipulation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. μ_H corresponds to the average reputation from high quality sellers, while μ_L corresponds to the average reputation from low quality sellers.	30
1.9	Equilibrium as a function of μ , when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$ and $C = .01$	32

1.10	The evolution of the effort of review manipulation and reputation chosen by a high and a low quality seller, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$	33
1.11	The evolution of the average effort of review manipulation and reputation from high and low quality firms, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$	34
1.12	An example of a seller soliciting fake reviews through RapidWorkers.com.	37
1.13	An example of a seller soliciting fake reviews through a Facebook community.	39
1.14	One of the products from figure 1.13 turns out to be a bestseller on Amazon.	40
1.15	Histogram of number of stars in the sample. Consistent with previous results from the literature, the distribution of number of stars given by reviewers has a J-shape format.	46
1.16	Estimated density of $\tilde{\mu}$ using the Epanechnikov kernel density.	46
1.17	An example of a seller soliciting positive feedback to reviews praising its products.	48
1.18	Average proportion and absolute number of fake reviews chosen by sellers as a function of the time since sellers' first review. Time was discretized into biweekly intervals. The gray area corresponds to bootstrapped 95% confidence intervals for the average proportion of fake reviews.	53
1.19	Histogram of the number of stars per sample. The bars in blue correspond to the sample of wireless earphone products, while the one in orange corresponds to the sample described in section 1.4.1 generated by targeting suspicious products that were either soliciting reviews in online platforms, or were flagged as suspicious on Amazon forums.	58
1.20	Number of 5 star reviews received by a couple of products per day. Product 1 is no longer sold at Amazon, perhaps because Amazon detected suspicious activity surrounding its reviews and thus had the product removed. Regarding product 2, as I write this on May 16, 2019, though it is still sold on Amazon, all its positive reviews (4 and 5 stars) have been removed.	60
1.21	Average number of fake reviews detected using text similarity, compared to the product's grade on Fakespot.com. According to the website, a grade of "A" indicates low level of review manipulation, whereas a grade of "F" indicates a high number of fraudulent reviews. The 95% confidence intervals displayed in the figure were built using 100,000 bootstrap simulations.	61

2.1	Schematics for the timeline of the model, where $f_q^t(\cdot)$ represents the beliefs that the agent has at the beginning of period t regarding the quality of the company; and $(f_p^{n,t}(\cdot q))_{n=1}^N$, the beliefs that the agent has at the beginning of period t on the precision from each channel, conditional on the quality of the company.	79
2.2	Long run polarization of q when $C = .001$ as a function of ρ , where one of the news channel has high precision $(1, \rho)$, while the other has low precision $(\rho, 1)$, and agents are not allowed to fact check the news. The discount factor β equals .5, and all agents start with the same uninformative priors.	84
2.3	Long run polarization of q when $C = .001$ as a function of ρ , where one of the news channel has high precision $(.6 - \rho, .6 + \rho)$, while the other has low precision $(.6 + \rho, .6 - \rho)$, and agents are not allowed to fact check the news. The discount factor β equals .5, and all agents start with the same uninformative priors, and the game is played for 5 periods.	85
3.1	Upper bound $U(m)$ compared with the upper bound derived by Storms (2013), when $\bar{q} = k = 5$	106
3.2	Upper bound $U(m)$ for $q = 1$ and different values of k ranging from 5 to 15, where the upper and lighter lines correspond to higher values of k	107
B.1	Simulated proportion of colleges that had potential incentives to mis-report preferences or vacancies under the SOSM mechanism, for different combinations of γ and β . The simulations were done assuming there were 45 colleges each offering 5 vacancies, and 225 students, so that the number of students equals to the overall number of vacancies. Each student was assumed to have $k = 5$ acceptable choices.	143

Chapter 1: On Incentives to Manipulate Online Ratings

With the proliferation of online rating platforms, there has been an increasing concern over the authenticity of reviews posted online. While much effort has been dedicated to improving fake review detection algorithms, little attention has been spent on understanding the incentives that drives some sellers to solicit fake reviews. To fill this gap, this paper develops a theoretical model in which sellers dynamically choose their effort spent on review manipulation. Among other things, the model predicts that sellers' optimal investment in fake reviews is not a monotone function of their reputation. More precisely, sellers that currently possess a very good or very bad history of past reviews have less incentives to solicit fake reviews praising their own products, the intuition being that, for sellers with very bad reputation, it is too costly to pretend that they are high quality sellers; while sellers that have already accumulated a very good reputation do not need to spend much effort in convincing buyers that they are high quality sellers. Another prediction from the model is that, in order to maximize the impact from each fake review, sellers tend to concentrate review manipulation at the initial stages after they have entered the market. Using data collected from Amazon, I was able to observe those two features from the model at the empirical level by estimating a Logit regression that predicts

the probability of a review being fake as a function of the product’s reputation and the time it took for the review to be posted since the seller entered the market.

1.1 Introduction

With the proliferation of online rating platforms such as Yelp and TripAdvisor, there has been an increasing concern over the authenticity of reviews posted online. In the news one can find several pieces of anecdotal evidence that fake reviews are prolific and have been increasing over the last years (see for instance *Amazon’s Fake Review Problem Is Now Worse Than Ever, Study Suggests (Forbes, Sep 9, 2017)*, *Facebook fake review factories uncovered by Which? investigation, (The Guardian, Oct., 2018)* and *‘The Shed at Dulwich’ was London’s top-rated restaurant. Just one problem: It didn’t exist (Washington Post, Dec 8, 2017)*).

Those practices can add distortions to the market through a series of channels, a few of which include: 1) they can lead customers to make suboptimal decisions;¹ 2) they can also lead customers to perceive reviews as a poor measurement of the product’s quality, and it is a well established result that, when buyers do not know

¹Notice that in some markets a suboptimal decision made by customers can be quite costly, such as in the market for private doctors, where the resulting assignment can impact patients’ health. And in those markets the surge of online platforms as the likes of [vitals.com](#), [ratemds.com](#), [Yelp](#), etc., which display reviews from physicians, has been increasingly playing a bigger role in customers’ decisions. [Shukla, Gao and Agarwal \(2018\)](#), for instance, find that the introduction of reviews in a doctor appointment platform in India has increased the number of online appointments for highly rated doctors by roughly 29.6%, while decreasing the number of appointments from unrated doctors.

the quality of the products sold, adverse selection may occur ([Akerlof \(1970\)](#)); 3) buying fake reviews can be unfair to sellers who refuse to engage in this practice for ethical or legal concerns, etc. These problems have motivated the dissemination of a literature in Computer Science dedicated to detecting and eliminating fake reviews in online platforms (see [section 1.2](#) for details).

In contrast to computer scientists, economists are usually more interested in understanding the causal effects that lead some sellers to fake reviews more than others. One motivation for understanding these causal effects is that such knowledge can be used to perfect existing algorithms used in the detection process. Indeed, by understanding what causes sellers to fake reviews, one can trim down the set of predictors to be used in the detection algorithm to include only the most pertinent variables.

This paper fits into the latter branch of the literature, as it aims to shed light into the incentives that drive some sellers to fake reviews. For that purpose the paper develops a novel theoretical framework in which sellers dynamically choose their optimal amount of effort devoted to faking reviews in online rating platforms such as Yelp, Amazon and TripAdvisor, given that faking reviews is costly to sellers and that customers correctly anticipate some reviews may be fake. The model predicts that, in order to maximize the impact from each fake review, high quality sellers have incentives to concentrate review fraud at the initial stages after they have entered the market. As their reputation improves over time, such sellers gradually fake less reviews. For low quality sellers, however, maintaining a good reputation is unsustainable in the long run, as they systematically receive reviews disparaging

their products from honest consumers. Once their reputation has been squandered, such sellers exit the market and reenter with a new brandname, always concentrating their efforts in review manipulation right after reentering the market, so as to maximize the impact from each fake review. So in markets wherein changing one's name is relatively costless, low quality sellers should be expected to disproportionately fake more reviews than high quality sellers in the long run. Verifying this last prediction empirically may be challenging, since the sellers that change their brandnames usually do so in a concealed fashion that prevents the researcher from accessing their previous identities.

Another qualitative prediction from the model is that the effort spent on review fraud is not a monotone function of the seller's reputation. More precisely, very low or very high reputation levels are usually associated with a low effort on review manipulation, the intuition being that, for sellers with very low reputation, it is too costly to pretend that they are high quality types, which gives them little incentives to fake reviews; similarly, for sellers with very good reputation, the marginal benefit from faking reviews is relatively small since everyone already believes the seller to be of high quality with high probability.

To test some of the predictions from the model, I scraped reviews from different products sold by Amazon that I flagged as suspicious based on the fact that their sellers were (apparently) soliciting fake reviews online (namely, through Facebook and Rapidworkers). I then classified the reviews collected as fake and real based on a series of criteria used in the computer science literature dedicated to detecting fake reviews through supervised learning algorithms. After that I estimated a Logit

model to predict the probability of a review being fake conditional on the seller’s reputation, and on the time it took for the review to be posted since the seller entered the market. Consistent with the predictions from the theoretical model, the results from the regressions suggest that the probability of a review being fake is lower for sellers with very high or very low reputation. Moreover, the probability of a review being fake diminishes with time, which is consistent with the prediction that sellers should focus review manipulation at the initial stages following their entrance (or reentrance with a new brandname) into the market.

1.2 Related Literature

When it comes to related literature done on the theoretical level, one can cite the work from [Mayzlin \(2006\)](#) and [Dellarocas \(2006\)](#). These papers have different premises and predictions as to the types of sellers that have most incentives to fake reviews.

Starting with Mayzlin’s paper, it assumes that consumers randomly observe a single opinion and then, based on that single observation, update their beliefs regarding the quality of sellers, knowing that the opinion that they picked might potentially be fake. Her model leads to the prediction that low quality sellers fake more reviews as compared to high quality sellers. However this conclusion relies on the assumption that each buyer only observes a single review extracted from the overall pool of reviews, when in reality most online rating platforms (e.g., Yelp, TripAdvisor, Amazon, etc.) provide summary statistics of all previous reviews.

Moreover the model implicitly assumes that consumers know exactly how many legitimate reviews have been posted online, even though they do not know the total number of reviews posted (i.e., the number of fake plus real reviews). In addition, her model imposes a series of technical restrictions on exogenous parameters, such as sellers' initial reputation, which prevents one from performing some comparative statics analysis, such as how reputation affects the effort on review manipulation. Finally, her model transpires in a single time window, which prevents one from accessing, for example, whether sellers have incentives to concentrate review fraud at the initial stages after they have entered into the market, or smooth review manipulation throughout the periods.

[Dellarocas \(2006\)](#), on the other hand, uses a different specification that abstracts from some of the technical complications associated with modeling fake review optimization. In particular, it assumes that legitimate reviews generate a stochastic signal (common to all customers) that is correlated with the product's true quality. Because the signal is stochastic, if sellers invest in the distortion of this signal through fake reviews, buyers can not perfectly separate which part of the signal was generated by honest reviews, and which part was explained by fake reviews added to the system.

Different than [Mayzlin \(2006\)](#), [Dellarocas \(2006\)](#) predicts that in some instances high quality sellers may actually fake more reviews than low quality sellers. But those results suffer from a few limitations, including: 1) The paper has a few mistakes when it comes to the derivation of the seller's profit function. 2) It uses an intractable data generating process for product quality. Indeed, it assumes that

the quality parameter is drawn from a normal distribution, which implies that the seller's strategy is a continuous function that maps its type into effort on review manipulation. This implies that the researcher has to rely on guess and verify methods to discover the functional format of the equilibrium strategy. In [Dellarocas \(2006\)](#), the equilibrium they analyze is one in which the seller's optimal investment on fake reviews is an affine function of its quality, which would then imply that a seller with a sufficiently low quality (or sufficiently high quality, depending on whether this affine function is increasing or decreasing) would choose to receive money to have its products disparaged (i.e., they would choose a negative amount of fake reviews), which is a strategy that might be hard to be implementable in practice. Not to mention that the model could potentially have many other equilibria. 3) Finally, like [Mayzlin \(2006\)](#), his model transpires in a single time window, thus preventing researches from accessing whether sellers have incentives to concentrate review fraud at the initial stages following their entrance into the market.

My theoretical framework is very similar to Dellarocas' specification. The main differences and innovations from my model are that: 1) it corrects some issues regarding the derivation of the profit function from the seller; 2) it assumes that there are only two types of quality, high and low, as opposed to a continuum set of types, which makes the model more tractable, and therefore allows one to compute relevant comparative statics; 3) it allows the seller to be forward looking and set its effort on review manipulation dynamically, thus allowing the researcher to derive conclusions regarding dynamic aspects of review manipulation.

A common feature share by these two papers which is also present in my

own model is that they both have the desirable property that consumers correctly anticipate that some reviews are fake. So when looking at signals generated by reviews, consumers curb their expectations by taking into account that some reviews are not perfectly reliable, which is consistent with anecdotal evidence that consumers are aware of the existence of fake reviews.

As to empirical papers that investigate variables that affect review manipulation, one can cite [Luca and Zervas \(2016\)](#) and [Mayzlin, Dover and Chevalier \(2014\)](#). Among other things, these two papers find, using different databases, that chain restaurants (in [Luca and Zervas \(2016\)](#)) and chain hotels (in [Mayzlin, Dover and Chevalier \(2014\)](#)) are less likely to fake positive reviews praising their products, since they offer a standard service that already has a solidified reputation.

[Luca and Zervas \(2016\)](#) also run regressions that seem to support their conjecture that sellers with lower reputation have more incentives to fake reviews, and as their reputation improve, they gradually fake less reviews (hence, the creative title from their paper: “fake it till you make it”). To test this conjecture, they use positive reviews (4 or 5 stars) filtered by Yelp as a proxy to measure the effort of review manipulation spent by restaurant owners, then regress the number of filtered reviews per time interval as a function of the total number of 1, 2, 3, 4 and 5 stars accumulated by the restaurant in previous periods. A limitation from this specification is that, by grouping the number of filtered reviews into time intervals the researchers lose pertinent information regarding each individual review that could be used to correct for classification errors (i.e., to correct for real reviews that were wrongly filtered by Yelp, as well as fake reviews that Yelp failed to filter). Moreover,

the number of reviews filtered by Yelp may not be a very good measure as to the effort spent on review manipulation, since, according to Yelp’s website, they not only filter reviews that have high chances of being fake, but also reviews that are likely to be less relevant to consumers (say, because the reviews lack useful content, or they are too old, etc.). Finally, even though the authors find that having more previous 4 and 5 stars are usually associated with less fake reviews in the current period, they haven’t actually created a univariate measure of reputation to test their conjecture. One contribution from my paper is that it uses a different database that targets the detection of fake reviews exclusively. Moreover I use a logit specification that corrects for endogenous classification errors by using data at the individual level. And finally, in order to measure the the impact that reputation has on the incentives to fake reviews, I create a univariate measure of sellers’ reputation, as opposed to using a vector of the number of previous 1, 2, 3, 4 and 5 stars received.

There is another branch of the literature that focuses in analyzing the impact that reviews have on sales. To cite a few papers, ?, ? and [Shukla, Gao and Agarwal \(2018\)](#) find that positive reviews have a positive and significant impact on sales (? use data from book reviews at Amazon and bn.com, while ? uses data from restaurant reviews on Yelp, and [Shukla, Gao and Agarwal \(2018\)](#) use review data from a doctor appointment platform in India). My paper, on the other hand, takes those results as given, and focuses instead in identifying the types of sellers that have most incentives to fake reviews.

This paper uses a combination of methodologies employed in the computer science literature to create training databases for the purpose of fake review detection.

Essentially, to detect fake reviews through supervised machine learning techniques, researchers need to feed the machine with some examples of reviews that they know to be fake, and another subsample of reviews that they know to be real, so that the machine can learn to distinguish the patterns from each group. The challenge is that in practice researchers can not tell for sure whether a review is fake or not, after all, fake reviews are supposed to be convincing. As a matter of fact, some experimental evidence suggest that humans are in general poor judges when it comes to detecting fake reviews (see for instance [Ott et al. \(n.d.\)](#)). So in order to build a training sample, researches usually spot reviews that are “clearly fake” based on some some baseline criteria that rely on computational methods. That is, while it is virtually impossible to determine from the naked eye whether a review is fake or not, by using automated methods that process big amounts of data, one can find reviews that are almost certainly fake.

[Kaghazgaran, Caverlee and Alfifi \(n.d.\)](#), for instance, looked at Amazon products that were soliciting fake reviews on the crowdsourcing platform RapidWorkers, and then classified a review as fake if the reviewer in question had posted reviews to two or more different products from the list. The premise behind this criterion lies in the fact that fake reviews are usually mass produced, and that the probability that a customer happens to review two or more products that were crowdsourcing fake reviews on the same platform by pure chance is very small, given that Amazon sells millions of different products. [Jindal and Liu \(n.d.\)](#), on the other hand, classify a review as fake if its review text is a near duplicate to some other review from the sample. The intuition behind this criterion lies in the fact that, since fake reviews

are usually mass produced, fake reviewers have a tendency to copy and paste the same text to describe different products. In my paper I combine these two criteria to classify reviews as fake and real, while also adding a new criterion, which, to the best of my knowledge, has not been exploited in previous literature (though the method draws some resemblance to the one used by Mukherjee et al. (2012) [Mukherjee, Liu and Glance \(n.d.\)](#) which regards groups of reviewers that provide feedback to the same products as suspicious).

1.3 Theoretical Model

The outline of the theoretical model can be described as follows: nature initially determines the type of the seller as being high or low quality with an exogenous probability μ_0 . After learning its type, the seller chooses how much to invest on review manipulation. Then a random signal v_1 is generated that is observable to consumers. The signal v_1 is positively correlated with the firm's quality and its investment on review manipulation. After observing the signal, potential buyers compute μ_1 , their updated beliefs regarding the probability that the seller is high type. After consumers update their beliefs, the seller chooses the optimum price p_1 from its product and then the heterogeneous consumers decide whether or not to purchase the product. After the firm's profits are realized for that period, the firm goes to the next period with a new reputation μ_1 and is matched with a new set of customers, wherein the same process is repeated iteratively.

Figure [1.1](#) depicts the outline of the model, where $\eta(q, \mu)$ corresponds to the

effort of review manipulation chosen by the firm as a function of its type q , and its current reputation μ .

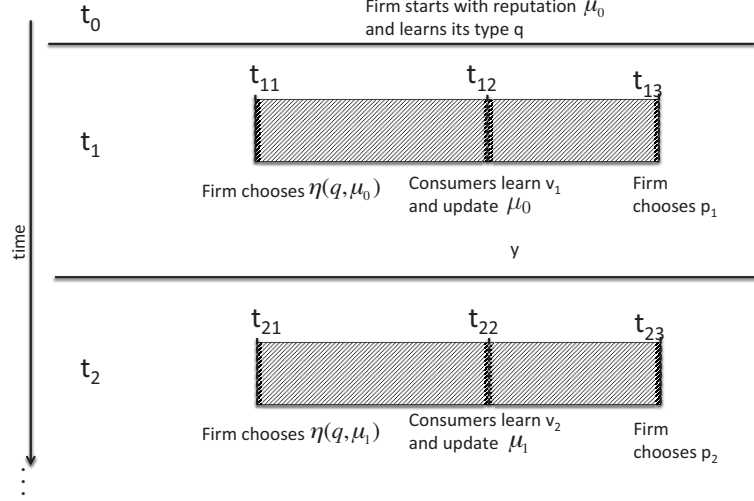


Figure 1.1: Outline of the model.

Later on, section 1.3.6 adds the possibility that at any point a seller can pay a fixed cost to exit and reenter the market with a new name. Sellers will of course only choose to do so once their reputation (i.e., their μ 's) have reached a sufficiently low point. Unsurprisingly, our simulations predict that the reputation of high quality sellers tend to improve over time, so that they usually find no need to resort to this tactic. Low quality sellers, on the other hand, tend to constantly exit and reenter the market with a new name, since their reputation tends to deteriorate over time as a result of honest reviews.

1.3.1 Firm's profits as a function of its expected quality

A monopolist wishes to sell a product with quality q that is unknown to potential customers. Time is discrete and finite, and indexed by $t \in \{0, 1, 2, \dots\}$. At

each period the firm is matched with a continuum of potential consumers uniformly distributed between $[0, 1]$ indexed by i . The utility that a consumer located at $i \in [0, 1]$ gets from purchasing a product with quality q and price p_t is given by:

$$u_i = q - p_t + i.$$

So implicitly this specification assumes that the firm is located at point 1 from the unit interval, and that customers located closer to the firm place a higher valuation for the product. As highlighted by [Tirole \(1988\)](#), the connotation of location does not have to be geographical: it can represent differences in tastes which causes consumers to have heterogenous willingness to pay for a certain product.

Letting $\mu_t \equiv \mathbb{E}_t(q)$ denote the expected quality from the firm given customers' beliefs at time t , we have that a consumer i will purchase the product if and only if

$$\begin{aligned} \mu_t - p_t + i &\geq 0 \\ \iff i &\geq p_t - \mu_t. \end{aligned}$$

This implies that if $p_t > 1 + \mu_t$ the demand is zero; if $p_t < \mu_t$ the demand is the entire unit interval; and finally, if $\mu_t \leq p_t \leq 1 + \mu_t$, the demand is given by $1 - (p_t - \mu_t)$. So in the end the demand faced by the firm is given by

$$D(\mu_t, p_t) = \begin{cases} 1, & \text{if } p_t < \mu_t \\ 1 - p_t + \mu_t, & \text{if } \mu_t \leq p_t \leq 1 + \mu_t \\ 0, & \text{if } p_t > 1 + \mu_t \end{cases}$$

Now assume that, after observing μ_t , the firm chooses the price p_t that maximizes its revenues $D(\mu_t, p_t)p_t$ (i.e., the firm is assumed to face zero marginal cost

of production, so that its profits equals to its total revenue). We also make the high level assumption that customers do not update their beliefs regarding q after observing the price p_t . This is more a result than an assumption, since it can be shown that, in the event prices can be used as a signal, there exists a “pooling” Bayesian equilibrium in the sense that all types with the same reputation charge the exact same price, given by the price that maximizes their revenue. Intuitively, for games of incomplete information in which the costs from sending a signal is the same for all types (in the current situation, the signal being the price), one should expect all types to send the same signal.² So if all firm types are to choose the same price, they might as well choose the price that maximizes their expected revenue.

If $0 \leq \mu_t \leq 1$, then the optimal price chosen by the firm is given by $p_t = (1 + \mu_t)/2$, which yields the firm an expected revenue of

$$\omega(\mu_t) = \frac{(1 + \mu_t)^2}{4}.$$

1.3.2 Allowing the firm to manipulate customers’ beliefs through fake reviews

Now assume that at each period the firm has the ability to influence μ_t by exerting some effort $\eta_t \geq 0$ in the fabrication of reviews that praise its own products.³

²As an example, in Akerlof’s market of lemons ([Akerlof \(1970\)](#)), all sellers are assumed to charge the same price, irrespective of the quality of the cars being sold.

³Fake reviews disparaging the firm’s rivals should have a similar effect as to fake reviews that praise the firm’s products. The main reason as to why I restrict attention on fake positive reviews is because empirically it is hard to detect the culprits from negative fake reviews (it could be

We assume the cost from choosing $\eta_t \geq 0$ is given by $\lambda\eta_t^2$. At each period t consumers get a noisy signal about the quality from the firm, given by

$$v_t = q + \eta_t + \varepsilon_t,$$

where q and ε_t are independent random variables that are not observable by customers. The term $q + \varepsilon_t$ from this expression can be interpreted as the part from the signal generated from honest reviews, while η_t is the fraction from the signal attributed to review manipulation financed by the seller.⁴ After observing v_t , customers update their beliefs regarding the distribution of q to form their expectation μ_t of q , and then decide whether or not to purchase the product.

We now closely examine how customers update their beliefs for a specific data generating process (DGP) for q and ε_t .

1.3.2.1 Customers' Bayesian updates

Assume that $q \in \{0, 1\}$, and let $\mu_0 = \mathbb{E}(q) = \text{Prob}(q = 1)$. Also assume that $(\varepsilon_t)_{t=1}^\infty$ is iid with $\varepsilon_t \sim N(0, \sigma^2)$. Then, if in period 1 the firm was to choose $\eta = \eta_H$ when $q = 1$, and $\eta = \eta_L$ when $q = 0$, we would have from Bayes' rule that consumers' updated beliefs that the firm is of quality $q = 1$ after observing v_1 should

anyone of the firm's rivals), whereas fake positives are usually orchestrated by the firm that is having its products praised.

⁴For tractability, this framework does not model customers' incentives to leave reviews; instead, it just assumes that the signal generated from sellers' honest reviews are stochastic and positively correlated with their quality. For a theoretical framework that examines buyers' incentives to post reviews, see [Campbell, Mayzlin and Shin \(2017\)](#).

be given by:

$$\mu_1 = \frac{\mu_0 e^{-\frac{(v_1 - 1 - \eta_H)^2}{2\sigma^2}}}{\mu_0 e^{-\frac{(v_1 - 1 - \eta_H)^2}{2\sigma^2}} + (1 - \mu_0) e^{-\frac{(v_1 - \eta_L)^2}{2\sigma^2}}},$$

In general, denoting $\boldsymbol{\eta} : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}^+$ as the amount of effort dedicated in faking reviews chosen by the seller as a function of its type $q \in \{0, 1\}$ and customers' beliefs $\mu_{t-1} \in [0, 1]$, we have that, starting at initial beliefs μ_0 , customers' beliefs and the seller's choices obey the following Markov process: for all $t = 0, 1, 2, \dots$,

$$\mu_t = \frac{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}} + (1 - \mu_{t-1}) e^{-\frac{(v_t - \boldsymbol{\eta}(0, \mu_{t-1}))^2}{2\sigma^2}}},$$

where

$$v_t = q + \boldsymbol{\eta}(q, \mu_{t-1}) + \varepsilon_t,$$

$(\varepsilon_t)_{t=1}^\infty$ is iid with $\varepsilon_t \sim N(0, \sigma^2)$, and q is the quality of the firm that is defined initially in period 0, and it is equal to 1 with probability μ_0 , and 0 with probability $1 - \mu_0$.

Implicitly we have made the high level assumption that consumers expect the strategy chosen by the seller, $\boldsymbol{\eta} : \{0, 1\} \times [0, 1] \rightarrow \mathbb{R}^+$, to only depend on the seller's type and on the previous beliefs μ_t held by customers. As shown in the next section, given such beliefs pertaining the strategy chosen by the seller, q and μ_t will indeed be sufficient statistics for the seller's optimal policy in period t .

1.3.2.2 Seller's optimal choice of review manipulation

Once consumers update their beliefs μ_{t+1} in period $t + 1$, the firm's expected quality is given by μ_{t+1} , which yields the firm a profit of $\omega(\mu_{t+1}) = (1 + \mu_{t+1})^2/4$ (see section 1.3.1).

So if customers expect the firm to adopt strategy $\boldsymbol{\eta}(q, \mu)$, we have that, starting at the initial μ_0 , a firm with quality $q \in \{0, 1\}$ chooses a sequence of $(\tilde{\boldsymbol{\eta}}(q, \mu^t))_{t=1}^{\infty}$ that solves:

$$\max_{\tilde{\boldsymbol{\eta}}(q, \mu^{t-1})} \sum_{t=1}^{\infty} \delta^{t-1} [\mathbb{E}_{t-1}[\omega(\mu_t)] - \lambda \tilde{\boldsymbol{\eta}}(q, \mu^{t-1})^2] \quad (1.1)$$

$$s.t. \quad \mu_t = \frac{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1} e^{-\frac{(v_t - 1 - \boldsymbol{\eta}(1, \mu_{t-1}))^2}{2\sigma^2}} + (1 - \mu_{t-1}) e^{-\frac{(v_t - \boldsymbol{\eta}(0, \mu_{t-1}))^2}{2\sigma^2}}}, \quad (1.2)$$

$$v_t = q + \tilde{\boldsymbol{\eta}}(q, \mu^{t-1}) + \varepsilon_t, \quad (1.3)$$

where $(\varepsilon_t)_{t=0}^{\infty}$ is iid with $\varepsilon_t \sim N(0, \sigma^2)$, and $\mu^{t-1} = (\mu_1, \mu_2, \dots, \mu_{t-1})$ is the entire history of beliefs up to time $t-1$, and $\delta \in [0, 1)$ is the firm's discount factor. At this point it is important to emphasize the distinction between $\tilde{\boldsymbol{\eta}}(q, \mu^{t-1})$ and $\boldsymbol{\eta}(q, \mu_{t-1})$. $\tilde{\boldsymbol{\eta}}(q, \mu^{t-1})$ is the strategy adopted by the firm, while $\boldsymbol{\eta}(q, \mu_{t-1})$ is what customers think what the strategy from the firm will be, which the firm takes as given. This is actually one of the main distinctions between this model and standard advertising models: in a standard advertising model, the amount of advertising is observed by customers, so the firm takes into account the direct impact that advertising has on customers' beliefs pertaining the strategy adopted by the firm; but in the present model customers do not update their beliefs regarding the strategy taken by the firm once they observe a signal, since customers can not observe the effort undertaken by the firm in review fraud ([Mayzlin \(2006\)](#)).

Because the expected payoff from the firm at period t only depends on the choice of $\tilde{\boldsymbol{\eta}}$ made by the firm at that period and on the variables q and μ_t , we can write the above sequential problem as a functional equation, where the state

variables are μ_t and q (notice that q is determined in period 0 and does not change over time). Therefore, from the principle of optimality, one can find the solution to the sequential problem 1.1 by solving the following Bellman equation:

$$V(q, \mu) = \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv \quad (1.4)$$

$$s.t. \quad \mu' = \frac{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}}}{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}} + (1-\mu)e^{-\frac{(v-\eta(0,\mu))^2}{2\sigma^2}}}. \quad (1.5)$$

Proposition 1.3.1 *Given $\eta(q, \mu)$, and imposing the constraint that the amount of fake reviews chosen by the firm, $\tilde{\eta}$, can not exceed a certain upper limit $\bar{\eta} > 0$, i.e, $\tilde{\eta} \in [0, \bar{\eta}]$, we have that the Bellman equation 1.4 has a unique solution.*

So given that customers believe that the seller adopts strategy $\eta(q, \mu)$, there is a unique solution to the seller's problem of choosing the optimal expenditure on review manipulation.

One can also easily show that, for the extreme points in which $\mu = 0$ or $\mu = 1$, the seller's optimal strategy consists on choosing $\tilde{\eta} = 0$, regardless of consumers' guess regarding the strategy taken by the seller, $\eta(q, \mu)$. Indeed, at those points the signal generated from reviews can not affect customers' beliefs, so that the seller has no incentives to try to influence the signal. This extreme result can be relaxed by allowing the seller's type to change over time according to a certain Markovian process. But qualitatively, adding that additional friction does not affect the main predictions of the model. So we now proceed to describe the equilibrium from this economy.

1.3.3 Equilibrium

Informally, the perfect Bayesian equilibrium (PBE) equilibrium from this economy is given by a policy function from the seller, and a policy function for customers such that: 1) Customers' maximize their expected utility when deciding whether or not to purchase a product, given their beliefs regarding the seller's type; 2) the seller maximizes its expected profits given customers' beliefs and customers' strategy; 3) Customers' beliefs regarding the seller's type are correctly updated through Bayes' rule.

Definition 1.3.1 (*Equilibrium*) *Given the initial probability of a firm being of high type, μ_0 , and the firm's quality q , a PBE from this economy is characterized by a strategy $\boldsymbol{\eta} : \{0, 1\} \times [0, 1]$ dictating the effort chosen by the firm at the beginning of each period as a function of its quality $q \in \{0, 1\}$ and its current reputation $\mu \in [0, 1]$, and consumers' beliefs, such that, for every $(q, \mu) \in \{0, 1\} \times [0, 1]$,*

$$\boldsymbol{\eta}(q, \mu) \in \arg \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv \quad (1.6)$$

$$s.t. \quad \mu' = \frac{\mu e^{-\frac{(v-1-\boldsymbol{\eta}(1,\mu))^2}{2\sigma^2}}}{\mu e^{-\frac{(v-1-\boldsymbol{\eta}(1,\mu))^2}{2\sigma^2}} + (1-\mu) e^{-\frac{(v-\boldsymbol{\eta}(0,\mu))^2}{2\sigma^2}}}, \quad (1.7)$$

and customers update their beliefs through Bayes' rule, so that $(\mu_t)_{t=1}^{\infty}$ obey the following stochastic process:

$$\mu_t = \frac{\mu_{t-1} e^{-\frac{(v_t-1-\boldsymbol{\eta}(1,\mu_{t-1}))^2}{2\sigma^2}}}{\mu_{t-1} e^{-\frac{(v_t-1-\boldsymbol{\eta}(1,\mu_{t-1}))^2}{2\sigma^2}} + (1-\mu_{t-1}) e^{-\frac{(v_t-\boldsymbol{\eta}(0,\mu_{t-1}))^2}{2\sigma^2}}}, \quad \forall t \geq 1,$$

where $v_t \sim N(q - \boldsymbol{\eta}(q, \mu_{t-1}), \sigma^2)$.

1.3.4 Finding the equilibrium numerically

An alternative way of interpreting the PBE concept is to think of the seller as choosing a policy function $\tilde{\eta}(q, \mu)$, and then having customers guessing a policy function $\eta(q, \mu)$ chosen by the seller, and then requiring that:

- I) Given $\eta(q, \mu)$, the strategy adopted by the seller, $\tilde{\eta}(q, \mu)$, is a solution to the Bellman equation 1.4 (i.e., the seller chooses the optimal amount of fake reviews given consumers' expectations regarding the strategy chosen by the seller);
- II) $\tilde{\eta}(q, \mu) = \eta(q, \mu)$ (i.e., customers correctly guess the strategy adopted by the seller).

This way of thinking about the PBE motivates the usage of the following algorithm for finding the equilibrium:

Algorithm 1 (Finding the PBE numerically)

1. *Guess a strategy $\eta(\cdot, \cdot)$.*
2. *Given this strategy, solve the Bellman equation 1.4 (say, by iterating the value function) to obtain a new guess $\tilde{\eta}(\cdot, \cdot)$ for the policy function, then go to the next step.*
3. *Compare $\tilde{\eta}(\cdot, \cdot)$ obtained in the previous step with $\eta(\cdot, \cdot)$. If these two policy functions are sufficiently close to each other (i.e., if $\sup_{(q, \mu)} |\tilde{\eta}(q, \mu) - \eta(q, \mu)|$ is sufficiently small), an approximation to the fixed point representing the*

seller's equilibrium strategy has been found, so stop the algorithm; else redefine $\eta(\cdot, \cdot) = \tilde{\eta}(\cdot, \cdot)$ and repeat step 2.

Applying this algorithm, we obtain the equilibrium strategy from the seller as depicted in figure 1.2. As it is clear from this figure, regardless of its type, a seller optimally chooses to exert more effort on review manipulation for intermediate values of μ , the intuition being that, for very low levels of reputation the seller finds it too costly to signal that it is of high quality, whereas a seller that has already accumulated a very good reputation does not need to prove that it sells a high quality product. In mathematical terms, reputation levels of $\mu = 0$ or $\mu = 1$ are absorbing states: once a seller achieves those reputation levels, they can not be altered.

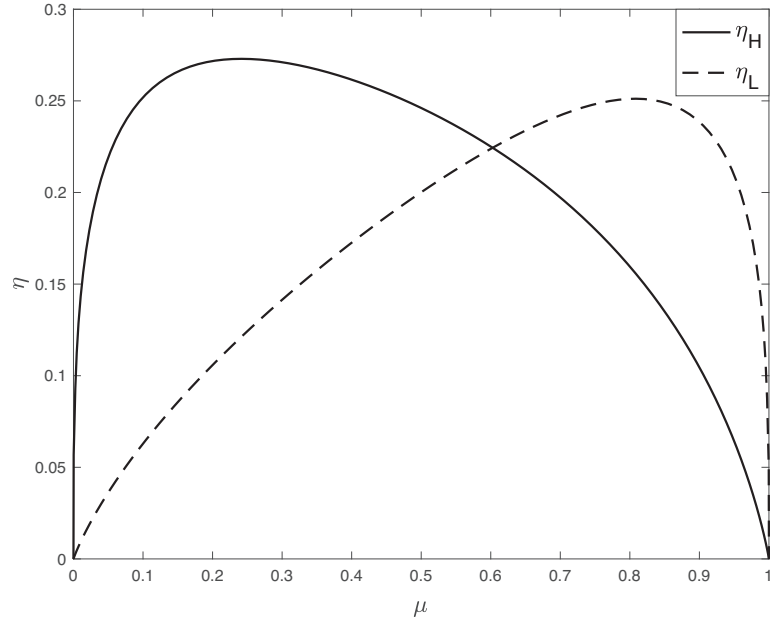


Figure 1.2: Equilibrium as a function of μ , when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$.

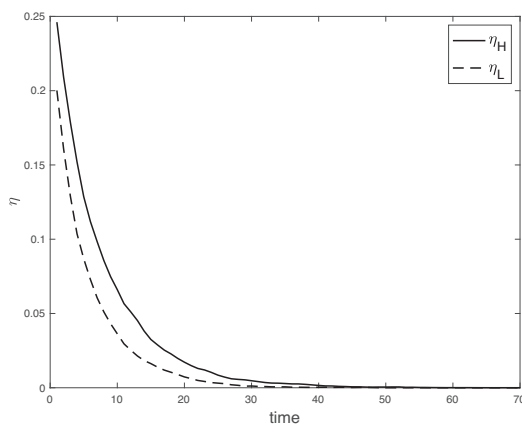
Another interesting feature from this equilibrium is that low quality sellers do not necessarily exert more effort on review manipulation. Which type spends most

effort on review manipulation depends on the current level of reputation held by the seller. Indeed, given a very low level of reputation, a high quality seller should spend more effort on review manipulation, and the opposite should hold when the seller has a very high reputation.

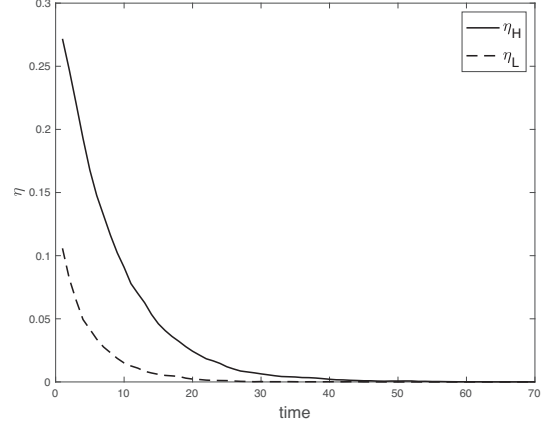
The intuition being this result can be explained as follows. Imagine that a seller currently possesses a very low reputation. In that case, the disutility from getting a bad signal is not so great, since the seller is already close to “rock bottom”. The gain in utility, however, from getting a good signal is more promising, as it can help the seller to separate itself from low types. In that case, because it is too costly for low types to pretend that they are high types, a high type should spend proportionally more on review manipulation. Analogously, if the seller has already accumulated a good reputation, then the marginal benefit from getting a good signal is relatively small, as compared to the disutility from getting a very bad signal. Because high quality sellers are very unlikely to get a very bad signal, it should be the low types, on that case, that will put more effort on the fabrication of fake reviews.

This result has implications on the dynamic choices made by the seller. Indeed, if the initial μ_0 is very low (i.e., if a firm in the market is most likely to be of low quality), then high quality types should be expected to be the ones spending most effort on review manipulation throughout the periods, as depicted in figure 1.20(a). If, however, μ_0 is large, then it is the low quality sellers that should be expected to spend most effort on review manipulation over the periods, as depicted in figure 1.3(c).

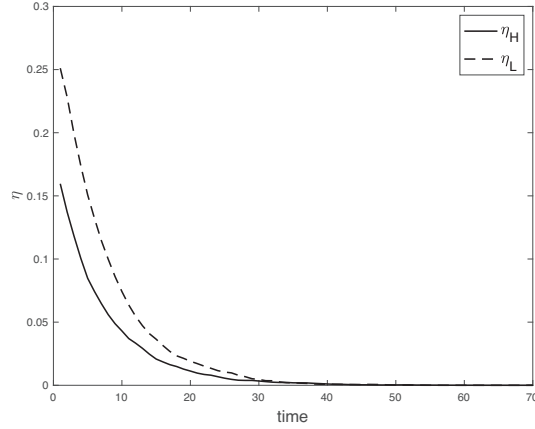
From figure 1.3 one can also see that both types spend most effort on review



(a) $\mu_0 = .2$



(b) $\mu_0 = .5$



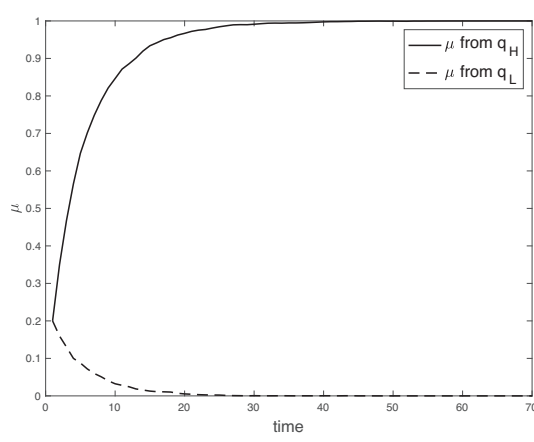
(c) $\mu_0 = .5$

Figure 1.3: Average simulated effort of review manipulation chosen throughout the periods, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. η_H corresponds to the average effort of review manipulation chosen by a high type seller (i.e., a seller with $q = 1$), while η_L corresponds to the average effort chosen by a low type seller (i.e., a seller with $q = 0$).

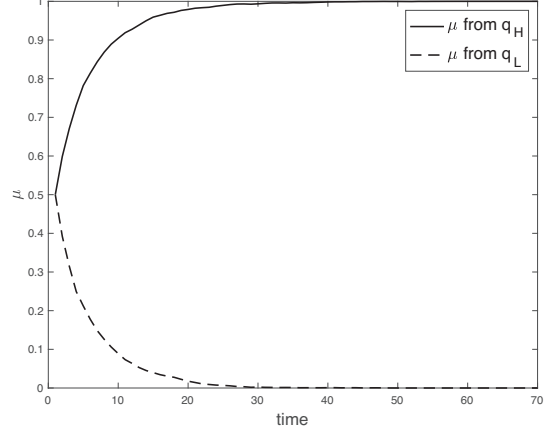
manipulation at the initial stages after they have they have opened their businesses, so as to maximize the impact from each fake review.

Now turning the attention to the evolution of reputation held by the seller, one can see from figure 1.4 that the reputation from a high quality seller tends to improve over time, while the reputation from a low quality seller gradually deteriorates. This happens because, as high quality sellers systematically receive positive reviews praising their products, their reputation tends to improve over time. For low quality sellers, however, it is too costly to maintain a high reputation in the long run due to the fact that they systematically receive negative reviews from honest consumers. So the model essentially predicts that in the long run customers learn the truth regarding the quality from sellers.

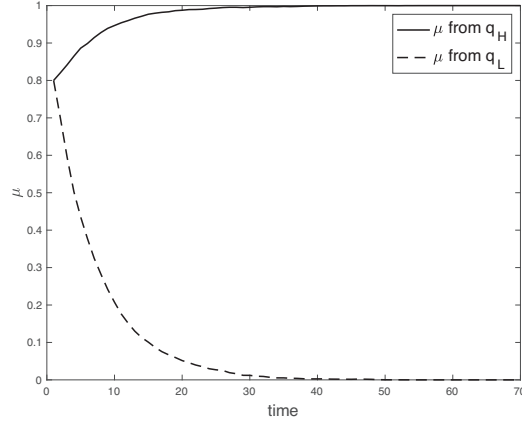
One can also compare the rate at which buyers learn the truth about the seller in this model with the case in which sellers are not allowed to fake reviews, say, because the monitoring of fake reviews is very intense or because the punishment applied to those caught faking reviews is very harsh, so that sellers have no incentives to fake reviews. Figure 1.5 compares the evolution of reputation for a high and low quality seller, when fake reviews is allowed to take place, and when it is not. As it is clear from these plots, the rate at which buyers learn the truth about the seller depends on the initial probability that the seller is of high type. Indeed, as discussed previously, high types will have incentives to fake more reviews than low types for low levels of reputation (see figure 1.2). In that case reviews are actually more informative when sellers are allowed to fake reviews, as the gap between the signal generated by high quality sellers and low quality sellers is greater. So in that



(a) $\mu_0 = .2$



(b) $\mu_0 = .5$



(c) $\mu_0 = .5$

Figure 1.4: The average simulated evolution of reputation, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$. μ_H corresponds to the average simulated reputation from a high type seller (i.e., a seller with $q = 1$), while μ_L corresponds to the average simulated reputation from a low type seller (i.e., a seller with $q = 0$).

case, buyers learn the truth about the type from the seller at a faster rate when fake reviews are allowed to take place, as depicted in figure 1.5(a). On the contrary, for high levels of reputation, it is the low type seller that has more incentives to fake reviews. So in that case allowing sellers to fake reviews closes the gap between the signal generated from high and low types, thus making reviews less informative, which in turn decreases the rate at which buyers learn the truth about the sellers' types as depicted in figure 1.5(b).

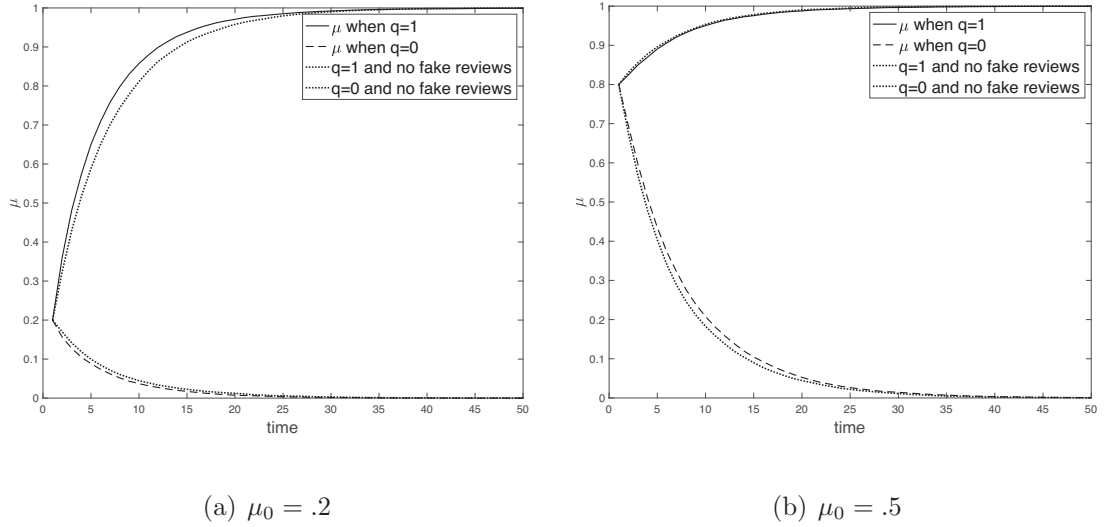


Figure 1.5: The average simulated evolution of reputation, when the seller is allowed and not allowed to fake reviews, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$.

1.3.5 When consumers do not anticipate review manipulation

The results from the previous section were built with the assumption that consumers know the strategy taken by the seller in equilibrium, i.e., that consumers correctly anticipate that some reviews may be fake, and that high and low quality

sellers manipulate reviews in different proportions. But one can also imagine scenarios in which customers are unaware of the existence of fraudulent reviews, or at least greatly underestimate how prevalent they are. So this section presents the results from the model in a scenario in which consumers incorrectly believe that the effort on review manipulation is zero for both high and low quality sellers.

Figure 1.6 displays the equilibrium strategy from both high and low quality sellers in this new environment with naive consumers, together with the equilibrium strategy from the standard version of the model presented in the previous section. At least for the set of parameters under consideration ($\delta = .8$ and $\lambda = \sigma^2 = 1$), when consumers are unaware of the existence of fake reviews, high quality sellers tend to engage in more review manipulation, while low quality sellers end up faking less reviews. As this diminishes the gap between the signals from high and low quality sellers, consumers take longer to learn the true type from the seller, as displayed in figure 1.7.

But qualitatively speaking, the results from either version of the model are very similar. Figure 1.8 displays the seller's average simulated effort on review manipulation through time for both scenarios. In both scenarios the seller tends to fake more reviews at the beginning, and as its reputation gradually improves (for a high type seller) or deteriorate (for a low type seller) it gradually reduces its effort on review manipulation as time goes by.

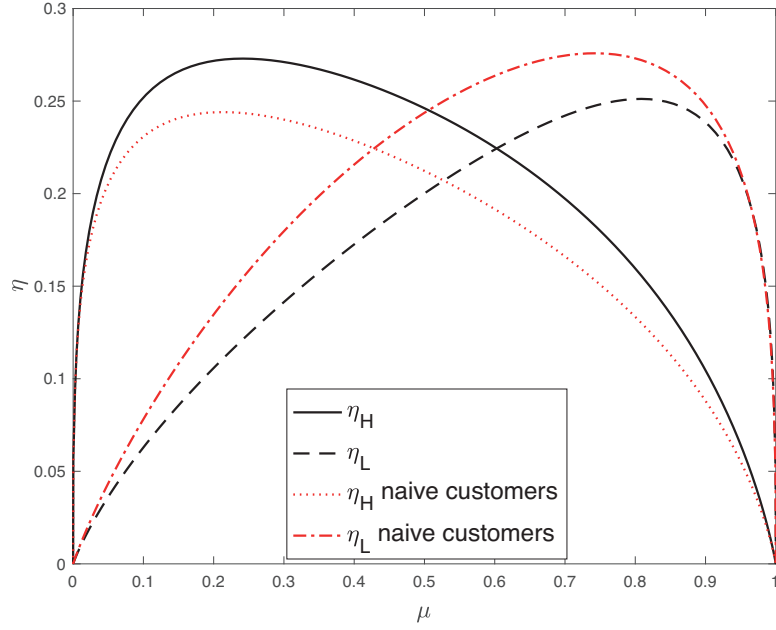


Figure 1.6: Equilibrium as a function of μ , when $\delta = .8$, $\lambda = 1$ and $\sigma^2 = 1$, when consumers know the strategy taken by the seller, and when consumers are naive and believe the seller does not engage in review manipulation (i.e., they believe $\eta(q, \mu) = 0$ for all q, μ).

1.3.6 Allowing the firm to exit and reenter the market with a new name

Now assume that the market starts with N sellers. As before, at each period a seller acts as a monopolist on its own market and they are each matched with a continuum of consumers with mass 1, and a seller's optimal profits given that customers believe that its expected quality is μ is given by $\omega(\mu) = (1 + \mu)^2/4$ (see section 1.3.1). At the beginning of each period, each seller retires with an exogenous probability p_e . At every period, $[p_e N]$ new sellers enter the market, where they are each high type ($q = 1$) with probability μ_0 and low type ($q = 0$) with probability

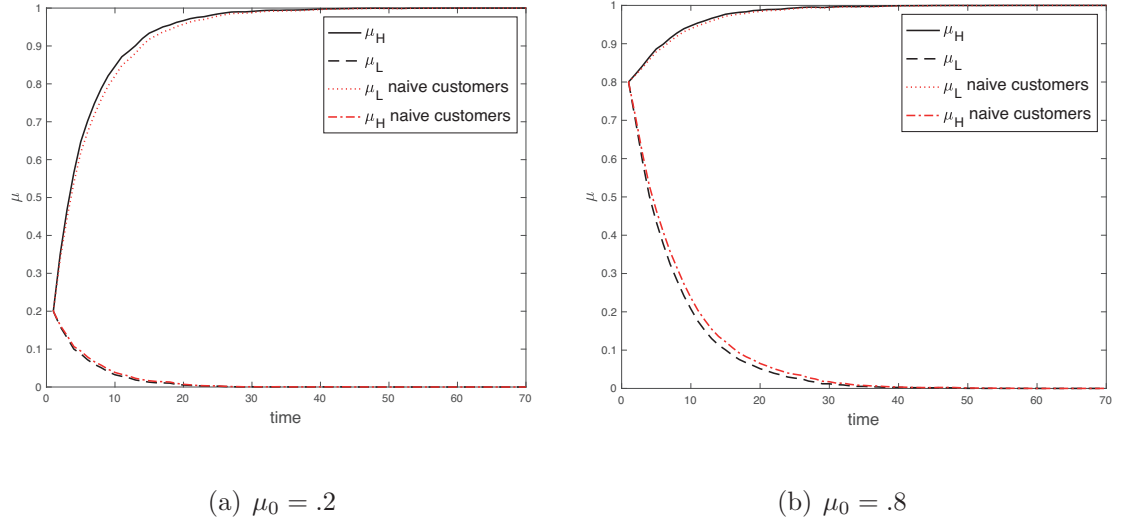


Figure 1.7: The average simulated evolution of reputation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. μ_H corresponds to the average reputation from high quality sellers, while μ_L corresponds to the average reputation from low quality sellers.

$1 - \mu_0$. But now sellers can “pretend” to retire and reenter the market with a new name, after paying a fixed cost $C > 0$. If consumers were oblivious of this scheme, they would believe a newcomer to be of high quality with ability μ_0 . But it is assumed that customers correctly anticipate that some firms may try to exit and reenter the market with a new name in order to hide a potential bad reputation obtained from previous reviews.

For technical reasons, it is assumed that at each period there is a small probability ρ_s that the seller is not allowed to exit and reenter the market. Without this assumption buyers would be allowed to have any beliefs whatsoever regarding the effort spent on review manipulation from a seller with a sufficiently low level of rep-

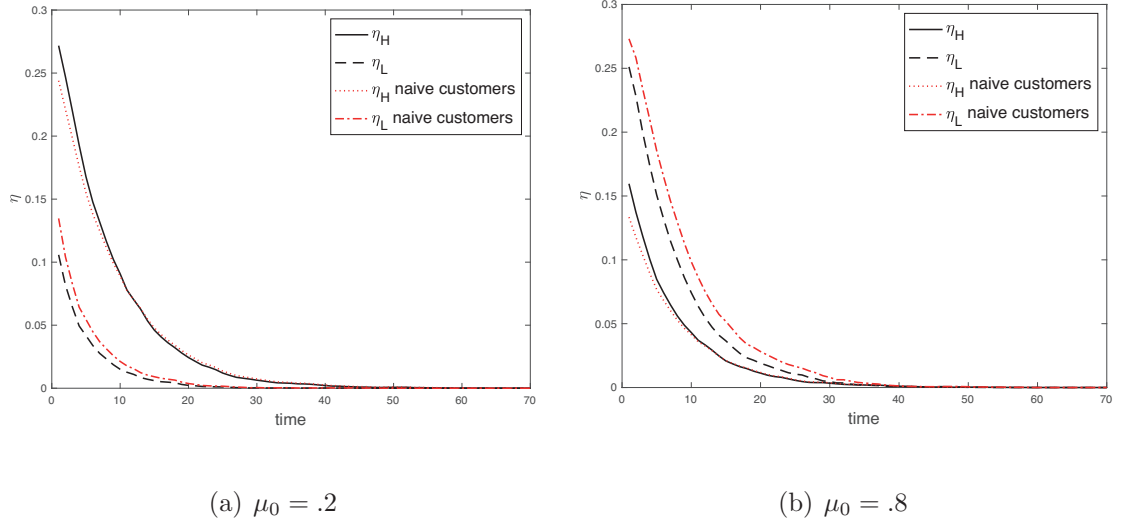


Figure 1.8: The average simulated evolution of effort on review manipulation when the seller is faced with sophisticated (black lines) or naive (red lines) customers, starting at different μ_0 's, when $\delta = .8$, $\lambda = 1$, $\sigma^2 = 1$. μ_H corresponds to the average reputation from high quality sellers, while μ_L corresponds to the average reputation from low quality sellers.

utation. Indeed, once a seller's reputation goes below a certain threshold, the seller optimally chooses to exit and reenter the market with a new name. So if sellers were always allowed to do that, in equilibrium one would never observe a seller with a very low level of reputation choosing some effort of review manipulation, which would then allow buyers to have any arbitrary beliefs regarding how much a seller would spend on fake reviews on those hypothetical scenarios. Because arbitrary beliefs pertaining decisions taken outside the equilibrium path can potentially affect the decisions made on the equilibrium path, it is important to impose some discipline on customers' beliefs on contingencies that are never reached in equilibrium. One way to accomplish that is by using equilibrium refinements. For example, instead

of assuming that agents play a PBE, one could assume that they play a sequential equilibrium. But for the current model, imposing discipline on customers' beliefs can be more easily achieved by simply assuming that there is a small probability ρ_s that the seller is not allowed to exit and reenter the market. This way, all feasible contingencies can be reached with positive probability, which implies that customers' beliefs regarding the actions taken by the firm in each contingency have to coincide with the firm's actual actions taken on those contingencies.

Henceforth a seller is defined as an *apparent newcomer* for period t if the seller has either started selling its product in period t , or if the seller was already selling its product before period t but changed its name in period t , so as to erase his past reputation. In that case, deriving the reputation from apparent newcomers can be difficult, since the reputation from those sellers should depend on the strategies chosen by the firms in equilibrium. So in order to compute the reputation from those firms I rely on Monte Carlo simulations. More precisely, I add an outer loop to algorithm 1 and iteratively compute the realized expected quality from newcomers in the long run to then use that statistic as a new guess for the reputation from newcomers, and keep repeating this process until a convergence criterion is reached.

Algorithm 2 (When firms can exit and reenter the market)

1. *Guess $\tilde{\mu}_0$, the probability that a an apparent newcomer is of high type in the long run (notice that because the term “an apparent newcomer” also encompasses old sellers that pretend to be new ones, $\tilde{\mu}_0 \neq \mu_0$).*
2. *Given $\tilde{\mu}_0$, compute the strategy taken by the firm by implementing a procedure*

similar to algorithm 1, then go to the next step.

3. Given the sellers' strategy conduct Monte Carlo simulations to compute $\tilde{\mu}_0'$, the long run expected probability that an apparent newcomer is of high type. If $|\tilde{\mu}_0' - \tilde{\mu}_0|$ is sufficiently small, stop the algorithm, else redefine $\tilde{\mu}_0 = \tilde{\mu}_0'$ and repeat step 2.

Applying this algorithm to a given set of parameters, we find a similar optimal strategy for the seller, as the one obtained previously, as depicted in figure 1.9.

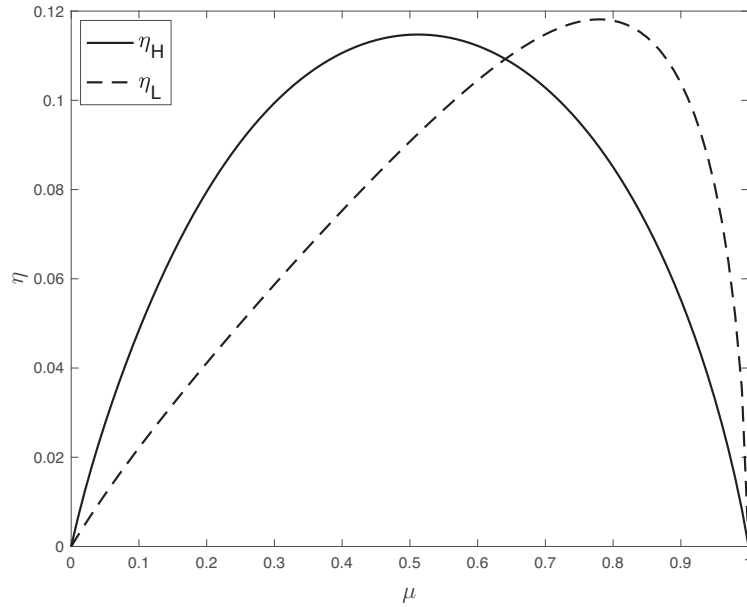
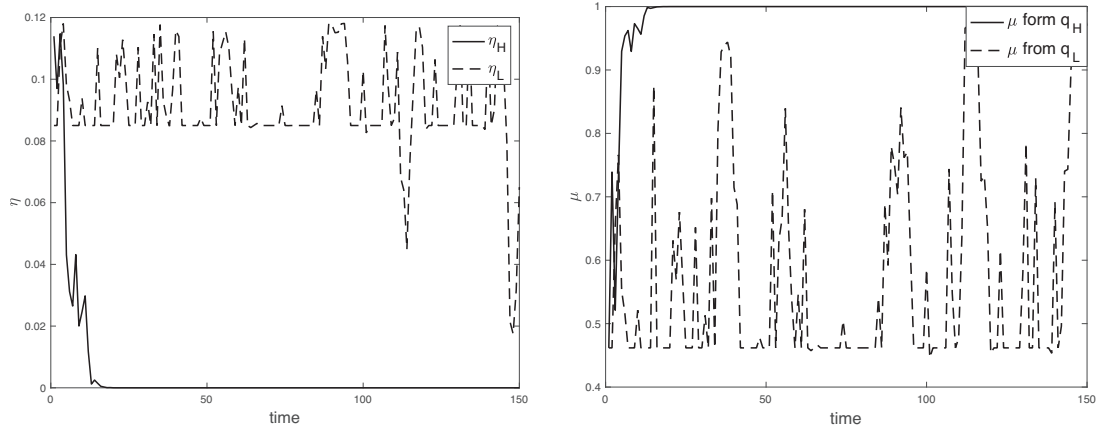


Figure 1.9: Equilibrium as a function of μ , when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$ and $C = .01$.

But because now sellers can exit and reenter the market with a new name once their reputation has been squandered, the dynamics from their effort on review manipulation changes. In particular, the reputation from low quality sellers tend to reach low levels constantly, which leads them to periodically exit and reenter the market with a new name, always concentrating review manipulation at the time

they reenter the market, so as to maximize the impact from each fake review. This phenomenon can be visualized in figure 1.10(a), which depicts the simulated effort of review manipulation from a single high and low type seller, while figure 1.11(b) depicts their corresponding level of reputation, assuming they are never selected to retire (recall that there is an exogenous probability p_e that at each given period the seller retires).



(a) $\mu_0 = .2$

(b) $\mu_0 = .5$

Figure 1.10: The evolution of the effort of review manipulation and reputation chosen by a high and a low quality seller, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$.

Figure 1.11 averages the strategy from sellers and their respective reputation over several simulations. As it is clear from those plots, consumers on average do not learn the type from low quality sellers in the long run, as they systematically exit and reenter the market with a new name once their reputation has been squandered. This result suggests that building obstacles that prevent sellers from anonymously selling their products under different account names should be pursued by online

platforms such as Amazon, so as to guarantee that reviews are informative.

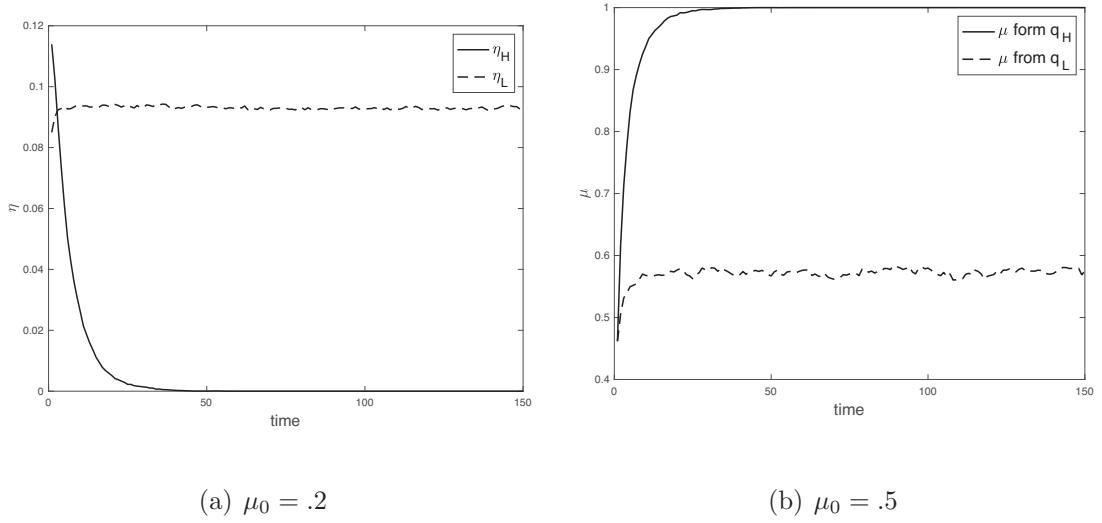


Figure 1.11: The evolution of the average effort of review manipulation and reputation from high and low quality firms, when $\delta = .95$, $\lambda = 1$, $\sigma^2 = 1$, $\mu_0 = .5$, $p_e = .5$, $C = .01$, $N = 100$.

1.4 Empirical Strategy

The theoretical model from the previous section provides a series of predictions regarding the pattern form review manipulation observed in online rating platforms. One of them is that the process of review manipulation is usually more concentrated during the initial periods following a seller's entrance (or reentrance with a new brandname) into the market, as sellers wish to maximize the impact from each fake review (see figures 1.3 and 1.11(a)). Another prediction is that the amount of effort dedicated in review manipulation does not vary monotonically with the seller's reputation. Indeed, from figures 1.2 and 1.9, sellers with very high or very low reputation levels will usually spend less effort manipulating reviews. For sellers

with low reputation, that happens because they find it too costly to pretend that they are high quality types, in which case they optimally choose to either give up trying to convince buyers that they are high quality sellers, or to exit and reenter the market with a new brand name in order to reenter the game with a fresh reputation. Analogously, for sellers with a very good reputation, the marginal benefit from faking reviews is small, since at that point buyers are already pretty much convinced that sellers are of high quality.

In order to verify to which extent those two predictions are observed empirically, I first scraped data from reviews posted on Amazon from a set of products that I classified as suspicious on the basis that they were soliciting positive fake reviews for their products on the internet (namely, on Facebook and Rapidworkers). Using that data I then apply a set of criteria to identify fake reviews from those sellers. Once fake reviews are identified, I then run a Logit regression to estimate the probability that a review is fake as a function of the product's reputation and the time since the product was first introduced in the market.

Because this dataset is comprised exclusively of suspicious products, it allows the researcher to detect fake reviews more easily. Indeed, if the researcher knows that two different products were involved in review solicitation, and he also observes that the same customer posted reviews on these two suspicious products, then the researcher can safely assume that those reviews are almost certainly fake, given that Amazon has a million of other products from which that customer could have chosen from, thus making the observed event very unlikely to have happened merely by chance. Clearly this detection criterion wouldn't be effective at all if applied to

products for which one has no prior knowledge about their involvement in review solicitation.

Therefore, an advantage about using this dataset to study fake reviews is that it allows the researcher to detect fake reviews more easily, which in turn diminishes the occurrences of classification errors in the sample (i.e., it diminishes the number of instances in which the researcher incorrectly classifies a fake review as real and vice versa).

But using this dataset also has a few caveats, one of them being that it may be susceptible to selection bias. Indeed, by focusing the analysis on sellers who are known to solicit fake reviews on the internet, the resulting sample may end up with an overrepresentation of fake reviews. For that and other reasons, on section [1.4.4](#) I then repeat a similar estimation analysis using a different set of products from Amazon, one in which sellers were not targeted in the sampling process.

Consistent with the theoretical model, the results from both regressions indicate that the probability of a review being fake decreases with time, and it varies non-monotonically with the seller's reputation, where very high or very low reputation levels are associated with a lower probability of the review being fake.

1.4.1 Database

This study uses a brand new dataset of reviews scraped from Amazon. The dataset contains information from 16,935 reviews from 206 different products. Information from each review includes the review text, the review title, the date the

review was posted, the number of stars given by the reviewer, whether the review came from a verified purchase, whether the review contained pictures or videos, whether the review received positive feedback, etc.

As to the 206 products for which reviews were collected, they were individually selected based on the fact that their sellers were soliciting fake review on online platforms such as RapidWorkers and Facebook. Knowing that these sellers were soliciting fake reviews online allows one to more easily detect which of the reviews posted were fake (see section 1.4.1.1), which then allows the researcher to identify the characteristics from fake reviews.

Figure 1.12 provides an example of an Amazon seller soliciting fake reviews through RapidWorkers.

The screenshot shows a task page on RapidWorkers. The title is "New Amazon Review - 2 min task - \$0.30" in red, followed by "fast and easy". On the right, there is a "Please select" button with a green icon. Below it, a link says "Not interested in this job". A "PROOF BOX" section asks the user to "Enter the proof in the box below" and includes a note: "* If a printscreen is asked, use a free service like this one: [http://prtsc.ca](\"http://prtsc.ca\")". A large empty box is provided for the proof. Below the title, task details are listed: "Work done: 5/10", "You will earn: \$0.30", "This task takes less than 2 minutes to finish", "Campaign ID : 59d351c8-4d18-4215-899d-45713257911a", and "Campaign Name : New Amazon Review - 2 min task - \$0.30 fast and easy". A section titled "You can accept this job if you are from THESE COUNTRIES ONLY:" lists "USA". A "Campaign isn't working?" section asks users to report issues and includes a "Click to Report" link. A "What is expected from workers?" section provides instructions: "Visit [https://goo.gl/zCdE5m](\"https://goo.gl/zCdE5m\")", "Read the listing and understand the product.", "Leave a 5 star review with 30 words minimum.", and "Duplicate reviews will be marked as unsatisfied - so please only complete this once."

Figure 1.12: An example of a seller soliciting fake reviews through RapidWorkers.com.

Figure 1.12 helps to illustrate two aspects about fake reviews solicited through

RapidWorkers: 1) when sellers use this website to solicit fake reviews, they provide a link specifying the url from the product in question, which can then be used by the researcher to identify the seller responsible for the fake review solicitation. 2) Moreover, notice that the amount paid per fake review is usually very small, given that this website focus on the solicitation of non-verified purchase reviews, which are virtually costless to fabricate. This small amount paid for unverified purchase reviews is consistent with anecdotal evidence provided in the News (see for instance *A Rave, a Pan, or Just a Fake?* (*New York Times*, 2011)) and other scientific research that also relies on RapidWorkers to spot fake reviews, such as Kaghazgaran et al. (2017) [Kaghazgaran, Caverlee and Alfifi \(n.d.\)](#).

As to verified purchase reviews, since they are perceived as being more reliable, and since they are more costly to fabricate, sellers are usually willing to pay a higher price for those type of fraudulent reviews, oftentimes by completely reimbursing reviewers after they have purchased the product and left their positive reviews. Figure 1.13 shows an example of a seller soliciting verified purchase fake reviews through a community on Facebook dedicated solely to facilitate the transaction of Amazon fake reviews.

Unlike Rapidworkers, on Facebook sellers tend to be less brazen when it comes to fake review solicitation. For starters, on Facebook sellers usually talk using code language. For the example in figure 1.13, “PP” stands for PayPal, which means that the buyer will be reimbursed through PayPal after purchasing the product through Amazon and leaving a positive review. “US” means that the purchase must be purchased in the US. And “PM” stands for private messaging, meaning

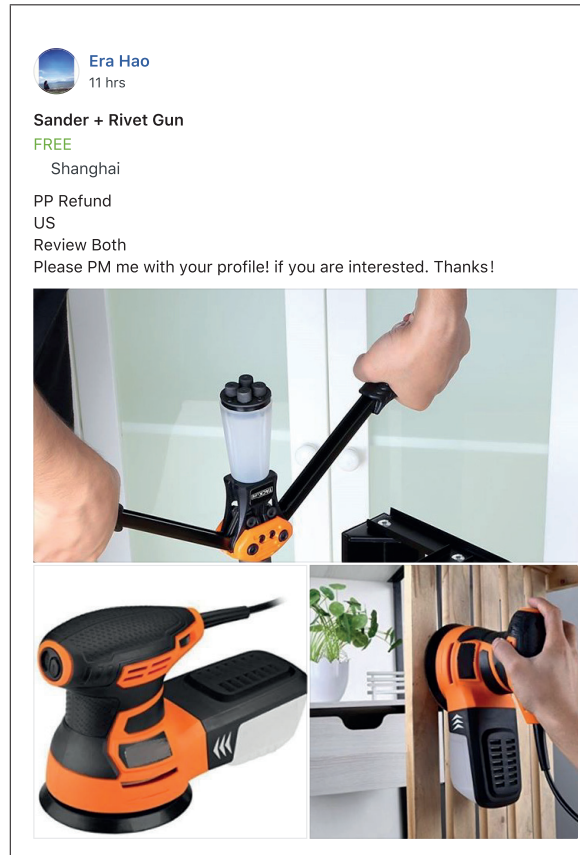


Figure 1.13: An example of a seller soliciting fake reviews through a Facebook community.

that whoever is interested in the gig should contact the seller through Facebook's private message system.

On top of that, sellers that solicit fake reviews through Facebook usually do not provide the url from their product. Instead they only display a picture and sometimes a small product description. This makes it harder for researchers and Amazon staff to detect the sellers responsible for the reviews solicitation. However, by doing a Google image search, one can identify some of these products. Indeed, for the product in question, a Google image search leads to the page depicted in figure 1.14. From the figure, one can see that, surprisingly, the product in question

has an Amazon best-seller stamp.

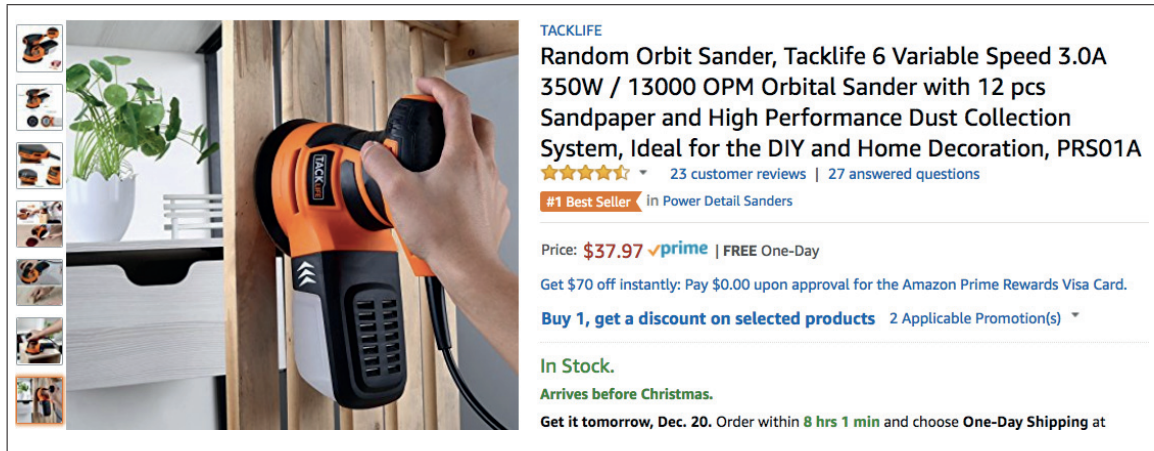


Figure 1.14: One of the products from figure 1.13 turns out to be a bestseller on Amazon.

So by doing a google image search on some of the products listed in Facebook communities, I was able to identify sellers that were probably soliciting fake reviews for their products.

Finally, another set of suspicious sellers were targeted using information posted from Amazon’s seller forum, where users would complain about the credibility of some of the reviews from certain products. As an example, a post from the forum would read:

“...This item gets over 5 reviews a day verified reviews and they are all 5 stars how is that even possible. Link below:

<https://www.amazon.com/Ashwagandha-EnhancerArtichokeEnhancedSupplement/dp/B06X...>

Is this legit or fake reviews. Am I missing out on some method to have this many reviews or are they all paid for...”

Edited by: Wholesale promo on Apr 30, 2017 1:03 PM

So I searched for similar posts and targeted the products being flagged as suspicious by concerned users as well as some other products that were being reviewed by the same set of customers.

In the end I ended up with a database of reviews 26,971 from 247 different products. Some of those products were targeted based on the fact that they were either soliciting fake reviews through RapidWorkers or Facebook, or they were being flagged as suspicious by concerned users on Amazon’s seller forum. Table 1.1 provides some summary statistics about the dataset.

	RapidWorkers	Facebook	Amazon Seller Forum	Total
Number of products	61	165	21	247
Number of reviews	4,724	20,685	1,562	26,971

Table 1.1: Number of products/reviews collected from each source

1.4.1.1 Fake review detection

As mentioned earlier, this study focuses in analyzing the patterns from positive fake reviews only. While fraudulent negative reviews aimed at a seller’s rivals should have a similar effect on sales as compared to fabricating positive reviews praising the seller’s own products; for the first tactic it is usually difficult to determine the agent(s) responsible for the review fabrication: they could have originated from any of the seller’s rivals, or even some disgruntled consumer. For positive fake reviews, on the other hand, it is invariably the seller receiving the positive fake reviews that

is behind their solicitation.

Now, given Amazon 1 to 5 star rating metric, classifying a review as positive or negative can be subjective: if the average rating given on Amazon was around 2 stars, then a 3 star would actually be considered a good review. But given that the great majority of reviews posted on Amazon are 5 star reviews (see figures 1.15 and 1.19), I classify a review as positive if and only if it has more than 4 stars.⁵ So reviews with 4 or 5 stars will be our candidates for positive fake reviews.

Depending on the source that led us to include a certain product in our list, a different set of criteria was used to determine whether a positive review from that product was fake or not. The combination of the two criteria listed below was used to categorize reviews from products in which fake review solicitation was happened on RapidWorkers or Facebook:

- I) If two different reviews were sufficiently similar to one another in terms of their text Jaccard similarity index, and the reviews in question had more than 5 words, and they were both from products in which fake review solicitation happened in the same online platform, and they both gave the seller at least 3 stars, then those reviews were classified as fake.⁶

⁵Jabr and Zheng [Jabr and Zheng \(2014\)](#) adopt the same criteria for classifying an Amazon book review as positive.

⁶To compute the Jaccard similarity index I used shingles containing 4 consecutive words, and I employed a hashing algorithm that addresses the computational burden of computing the actual Jaccard Similarity index by computing an unbiased estimator of the index. For details, check section A.0.2 from the appendix.

II) If a reviewer id was linked to two or more reviews from two or more different products that were soliciting fake reviews on the same online platform, and if the corresponding reviews gave the sellers a grade of at least 3 stars, then they were classified as fake.

All other reviews were classified as real. Of course, this process inevitably lead us to incorrectly classify some actual fake reviews as real. But one should keep in mind that, when it comes to fake review detection, no classification is perfect. Certain methods, however, can be implemented to correct for potential misclassification errors, as discussed later in section 1.4.3.

As to reviews from products that were being discussed on Amazon seller forums, a more conservative approach had to be implemented in order to classify them as fake or real. Indeed, from the way in which products were targeted using Amazon seller forums (see section 1.4.1), intersections among the products reviewed by customers are to be expected even when the reviews in question are real. So if one finds a buyer reviewing two different products from the list of products targeted using Amazon seller forum, that does not constitute a strong indication that the review is fake, so that criterion II) listed above does not effectively detect fake reviews for that sample. So for products targeted using Amazon seller forum, while still preserving criterion I), I replaced criterion II) for the more conservative classification rule:

II*) Consider an undirected graph where each reviewer is linked to the products that they review. If a cycle containing 2 or 3 reviewers is formed, then the corresponding reviews responsible for that cycle are classified as fake.

Applying the criteria described above to our dataset, 3,834 of the total of 26,971 reviews are classified as fake, while the remaining 23,137 reviews are classified as real. That is, approximately 14% of the reviews from the sample were classified as fake. Though that is a lot of fake reviews, one should keep in mind that, when building our sample we deliberately targeted suspicious products, so as to simplify the task of detecting fake reviews. Therefore one should not interpret this percentage as an accurate depiction as to how prevalent fake reviews are in online platforms such as Amazon.

1.4.1.2 Measuring reputation

The theoretical model from section 1.3 defined a seller's reputation as customer's beliefs that the seller is of high type. Of course, such probability is not observable in practice, which motivates the usage of some statistic that captures this reputation. One possibility would be to use the average number of stars received by a seller at any given time. The problem with this statistic, however, is that it would imply that if a product had a single review, and that review happened to give the seller 5 stars, then the seller would have the highest reputation score that one could possibly get.

So instead I use the following statistic as a proxy for reputation: 1) each seller starts with a score of zero. 2) For each review received by the seller, the seller's score is added or subtracted by a certain number depending on whether the review gave him 1, 2, 3, 4 or 5 stars. A 1-star review reduces the score from the seller by

2 points, a 2-star review reduces the raw score by 1 point, a 3-star review does not affect the score, a 4-star review adds 1 point to the score, and a 5-star review adds 2 points to the score. 3) After computing the raw score from seller s at time t , $r_{s,t}$, I then normalize it to a 0 to 1 scale by computing

$$\tilde{\mu}_{s,t} \equiv \frac{1}{1 + \exp(-r_{s,t}/\sigma_r)}, \quad (1.8)$$

the final statistic used to measure the seller's reputation, where σ_r is the standard deviation of the raw score $r_{s,t}$.

Because most reviews on Amazon tend to have 5 stars, as depicted in histogram 1.15, the lower ψ is, the more the distribution of reputation $\tilde{\mu}$ will be concentrated around 1, the maximum reputation level possible. To prevent such distribution from having positive mass at 1, the smoothing parameter was set at $\psi = 30$, for which the density estimate is depicted in figure 1.16. As is visible from this figure, very low levels of reputation are rarely observed, which again is associated with the fact that the great majority of reviews posted on Amazon tend to be 5 star reviews.

As a final observation, notice that this measure of reputation does not take into account that some reviews may be fake. So on that regard it is more aligned with the version of the model in which consumers are naive and do not expect some reviews to be fake (see section 1.3.5).

1.4.1.3 Covariates

Some covariates were added in our regressions in order to control for endogenous shocks. Two of these covariates use text analysis to estimate the probability

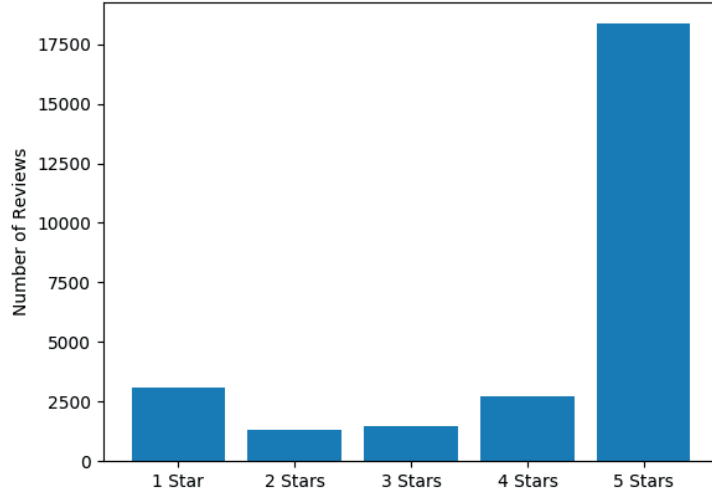


Figure 1.15: Histogram of number of stars in the sample. Consistent with previous results from the literature, the distribution of number of stars given by reviewers has a J-shape format.

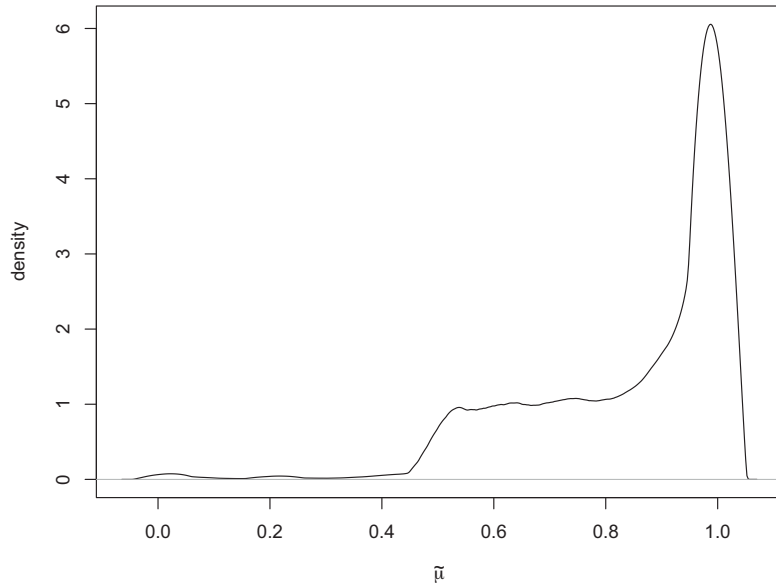


Figure 1.16: Estimated density of $\tilde{\mu}$ using the Epanechnikov kernel density.

that a review is real conditional on its content. One of these two variables uses content from the review title, while the other uses content from the review text. I call these variables “reliability from review title” and “reliability from review text”.

Each of these variables are dummies that assume value 1 if the contents of the review text (review title, resp.) are more likely to have been generated by legitimate reviews. For both of these measures I also imposed the restriction that a review with 1 to 3 stars is real with probability 1 (recall that this paper focuses on the detection of positive fake reviews, since for those cases it is clear who is responsible for the solicitation of fake reviews: the seller who is getting its product praised by dishonest reviewers). The method used for computing these dummies is the naïve Bayes' estimate. In spite of its name, this statistic has proven to perform surprisingly well as compared to more sophisticated methods. This, added to its simplicity, has lead this statistic to be commonly employed in the computer science literature.

Other covariates include whether or not a review contained a picture or a photograph. Controlling for these variables is important since, from several examples that I found on Facebook and RapidWorkers, buyers who were soliciting fake reviews would occasionally pay extra if a photograph or video was added to the review, supposedly to make the reviews more convincing.

Another important variable consists on the amount of feedback received by a review.⁷ The impact that this variable should have on the probability of a review being fake is ambiguous. Indeed, on the one hand, one would expect that more positive feedback would imply that the others found the review to be useful, and therefore less likely to be fake. But on the other hand, one can see cases in which sellers solicit positive feedback on positive reviews, as depicted in figure 1.17. Because of this, it could actually be the case that more positive feedback are associated

⁷As I write this in late 2018, Amazon only displays positive feedback, not negative ones.

with a higher probability of a review being fake. To their surprise, Jindal and Liu (2008) Jindal and Liu (n.d.) find that the latter occurs in their Amazon dataset, i.e., they find that more positive feedback actually increases the probability of a given review being fake, probably due to the fact that some sellers solicit *fake feedback on fake reviews*.

Amazon HELPFUL VOTES Needed


Work done: 3/⁵⁰
You will earn: \$0.11
This task takes less than 5 minutes to finish
Campaign ID : 59e38c46-b098-4511-85f7-4eda3257911a
Campaign Name : Amazon HELPFUL VOTES Needed

You can accept this job if you are from THESE COUNTRIES ONLY:

International

Campaign isn't working?

Campaign isn't working? If a Campaign does not work, please report that immediately. Include Campaign name and Campaign ID [Click to Report](#)



What is expected from workers?

THIS IS AN AMAZON.COM HELPFUL VOTE. YOU CAN DO THIS USING ANY AMAZON ACCOUNT THAT CAN POST REVIEWS.

YOU MUST VOTE YES

Use the link below

Please VOTE YES ON ALL REVIEWS LISTED BELOW

https://www.amazon.com/gp/customer-reviews/R3SLQR3BGP3JTZ/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B075LP9BSY https://www.amazon.com/gp/customer-reviews/R12I5TIT2QHFAG/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=B075JRHLK
https://www.amazon.com/review/R1U7S7YH6LQ5FL/ref=pe_1098610_137716200_cm_rv_eml_rv0_rv
https://www.amazon.com/review/R1A01I2I65VF7N/ref=cm_cr_srp_d_rdn_perm?ie=UTF8

Figure 1.17: An example of a seller soliciting positive feedback to reviews praising its products.

An additional variable consists on the number of words from a review text. Again, the effect of this variable on the probability of a review being fake is ambigu-

ous. On the one hand one could argue that, since fake reviewers are usually mass produced, the length from fake reviews are expected to be shorter. Some of the examples discussed earlier, however, suggest that the opposite can occur. Indeed, given that sellers occasionally ask for pictures or videos to be added to the reviews, or ask for positive feedback to be given to favorable reviews, it seems that sellers are not only interested in acquiring a high volume of positive fake reviews, but they also want those reviews to be convincing. So it may actually be the case that the text from fake reviews are on average longer than real reviews, due to the extra effort put by fake reviewers so as to make their reviews look more convincing.

Last, but not least, a dummy that determines whether or not a review came from a verified purchase is added to the vector of covariates. This variable is potentially a good predictor as to whether or not a review is fake. Indeed, since unverified purchase fake reviews are usually a lot cheaper to acquire as compared to verified purchase ones, one would expect unverified purchases to be more likely to be fake. One force, however, that can potentially change the effect from this variable is the selection bias present in our sample. Indeed, most of our sample consists on products for which fake reviews were being solicited on Facebook. Given that the reviews solicited on Facebook were mostly (if not exclusively) verified purchase ones, this can cause our data to have an overrepresentation of verified purchase fake reviews.

1.4.2 Logit model

Let $y_{i,s,t}$ be a binary variable that indicates whether review i from product s posted at time t is fake or not, i.e., $y_{i,s,t} = 1$ if the review is fake, and $y_{i,s,t} = 0$ if the review is real. Next, define the latent variable

$$y_{i,s,t}^* = \tau_{i,s,t}\beta_0 + \beta_1\tilde{\mu}_{s,t} + \beta_2\tilde{\mu}_{s,t}^2 + \mathbf{z}_{i,s,t}\gamma + v_{i,s,t}, \quad (1.9)$$

where $\tau_{i,s,t}$ is the time it took for review i from product s to be posted since the first review received by seller s , $\tilde{\mu}_{s,t}$ is the statistic derived from expression 1.8, which measures the reputation from seller s up to time t ; $\mathbf{z}_{i,s,t}$ is the vector of covariates described in section 1.4.1.3, and $v_{i,s,t}$ is an idiosyncratic error term with cdf $F(\cdot)$.

Assume that

$$y_{i,s,t} = \begin{cases} 1, & \text{if } y_{i,s,t}^* \geq 0 \\ 0, & \text{if } y_{i,s,t}^* < 0 \end{cases}.$$

Given the predictions from the theoretical model, one would want to test the following hypothesis: 1) $\beta_0 < 0$, so that older reviews are more likely to be fake, and 2) that β_1 is positive, while β_2 is negative, in such a way that sellers with very low or very high reputation have less incentives to fake reviews. Therefore, denoting $\mathbf{x}_{i,s}$ as the vector containing the variables of interest, i.e.,

$$\mathbf{x}_{i,s,t} = [\tau_{i,s,t}, \mu_{i,s,t}, \mu_{i,s,t}^2],$$

the loglikelihood function from this model can then be written as:

$$l(\beta, \gamma) = \sum_{i,s,t} [y_{i,s,t} \log(F(\mathbf{x}_{i,s,t}\beta + \mathbf{z}_{i,s,t}\gamma)) + (1 - y_{i,s,t}) \log(1 - F(\mathbf{x}_{i,s,t}\beta + \mathbf{z}_{i,s,t}\gamma))].$$

	<i>Dependent variable:</i>	
	$y = \mathbb{1}(\text{review is fake})$	
	(1)	(2)
Constant	−0.795 (0.607)	−5.715*** (0.531)
μ	9.591*** (1.612)	16.095*** (1.414)
μ^2	−7.199*** (1.043)	−11.888*** (0.910)
time	−0.008*** (0.001)	−0.015*** (0.001)
Dummy for text reliability	−2.607*** (0.059)	
Dummy for title reliability	−1.488*** (0.065)	
Numb. helpful feedback	−0.019*** (0.006)	
Verified purchase	−0.209*** (0.065)	
Has images or videos	0.607*** (0.086)	
Observations	18,440	18,440
Log Likelihood	−5,241.875	−7,552.646
Akaike Inf. Crit.	10,501.750	15,113.290
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01	

Table 1.2: Regression Results

Table 1.2 displays the results from Logit regressions (i.e., assuming that $F(\cdot)$ is the cdf from a logit distribution). Since the estimates from all specifications are very similar we will analyze the results from a single specification, namely column (1) from the table. Probit regressions also yield similar results and are therefore omitted.

Consistent with the predictions from the theoretical model, the time coefficient is negative and statistically significant, so that the longer it takes for a review to be posted since the seller entered the market, the less likely the review is to be classified as fraudulent. This result is also consistent with the plot depicted in figure 1.18, which displays the average proportion of fake reviews chosen by sellers as a function of time. As it is clear from the graph, the bulk of fake reviews is mostly concentrated around the few weeks following a seller's entrance (or reentrance) into the market. The small increase on fake reviews at the right tail of the graph can be attributed to the fact that this measure becomes increasingly less precise as time advances, since the number of observations used to compute this statistic decreases as time increases (i.e., the longer the time span, the less sellers there are in the sample that have lived long enough to be included in the average).

Turning back to the results from the regression, the coefficients for reputation and reputation squared suggest an inverted U-shape relationship between reputation and effort on review manipulation. This is consistent with the model's prediction that sellers that have either accumulated a very high or very low reputation have less incentives to fake reviews.

As to coefficients from covariates, they mostly have the sign that one would

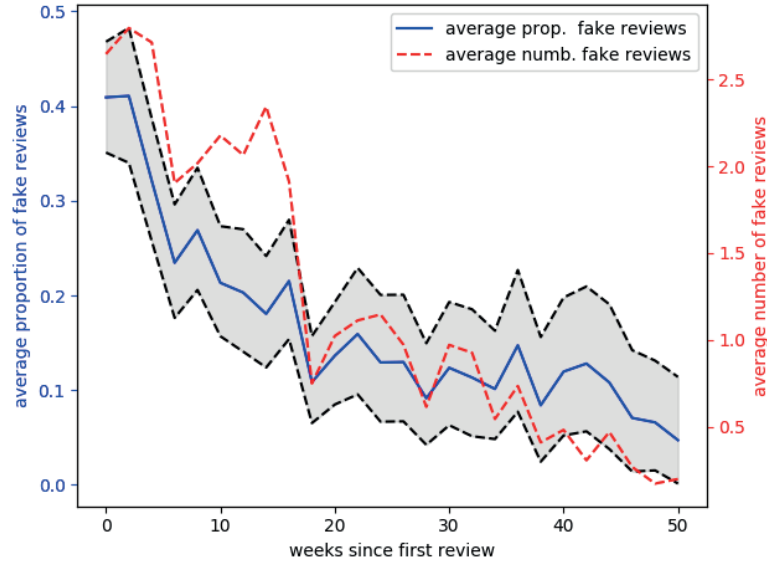


Figure 1.18: Average proportion and absolute number of fake reviews chosen by sellers as a function of the time since sellers’ first review. Time was discretized into biweekly intervals. The gray area corresponds to bootstrapped 95% confidence intervals for the average proportion of fake reviews.

normally expect. For instance, the coefficients for the reliability index dummies, which capture the probability that a review is real based on their text content, have negative signs, which means that reviews with more reliable review texts and review titles are unsurprisingly less likely to be fake (see section [A.0.3.1](#) for a detailed explanation as to how these dummies were constructed).

A review coming from a verified purchase decreases the probability of a review being fake, most likely due to the fact that verified purchase reviews are more costly to fake. And finally, a review having images or videos increases the probability of a review being fake. That is probably due to the fact that sellers occasionally solicit pictures or videos to be added to fake reviews, so as to make them more convincing.

1.4.3 Logit model correcting for classification error

The logit model presented in section 1.4.2 assumed that the variable $y_{i,s,t}$ used to classify reviews as fake or real was flawless, i.e., that there were no instances in which some fake reviews were incorrectly classified as real, and vice versa. But in practice the researcher can not determine with absolute certainty whether a review is fake or real, so that one should expect a certain degree of misclassification to be present in the dataset. In our case, even though reviews were only classified as fake when very strong evidence supported that those reviews were in fact fake (see section 1.4.1.1), it is very likely that some of the reviews from our sample were incorrectly classified as real. So in essence our variable of interest $y_{i,s,t}$ is not observable. What is observable instead is $y_{i,s,t}^o$, an indicator variable that equals 1 if the researcher classified review i from product s posted at time t as fake, and zero otherwise, where occasionally we may have $y_{i,s,t}^o \neq y_{i,s}$.

Because the presence of misclassifications of the dependent binary variable causes the probit and Logit estimates to be biased and inconsistent, I use an estimation approach proposed by [Tennekoon and Rosenman \(2016\)](#) that corrects for endogenous misclassifications. Formally, let $\mathbf{z}_{i,t,s}$ be a vector of covariates that can predict whether or not a review is fake, such as the length from the review, whether or not the review was from a verified purchase, whether or not the review contained a picture or a video, etc. Then we assume that the probability that a review is classified as fake when the review is in indeed fake conditional on the vector of

covariates $\mathbf{z}_{i,s,t}$ is given by:

$$Prob(y_{i,s,t}^o = 1 | y_{i,s,t} = 1, \mathbf{z}_{i,s,t}) = F_o(\mathbf{z}_{i,s,t}\gamma),$$

where $F_o(\cdot)$ is a cdf. Because reviews from our sample were only classified as fake when strong evidence supported that those reviews were indeed fraudulent (see section 1.4.1.1), I assume that a real review from our sample is never incorrectly classified as fake, i.e.,

$$Prob(y_{i,s,t}^o = 0 | y_{i,s,t} = 0, \mathbf{z}_{i,s,t}) = 1.$$

Therefore, letting $\mathbf{x}_{i,s,t}$ denote the vector of explanatory variables of interest (namely, the time it took for the review was posted, and the categorical dummies indicating the cohort of reputation from the product at the time the review was posted), we have that the overall probability of observing $y_{i,s,t}^o = 1$ given the covariates from the model is given by

$$\begin{aligned} Prob(y_{i,s,t}^o = 1 | \mathbf{x}_{i,s,t}, \mathbf{z}_{i,s,t}) &= Prob(y_{i,s,t} = 1 | \mathbf{x}_{i,s,t}) Prob(y_{i,s,t}^o = 1 | y_{i,s,t} = 1, \mathbf{z}_{i,s,t}) \\ &= F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma) \end{aligned}$$

With these probabilities, we can then build the loglikelihood function

$$l(\beta, \gamma) = \sum_{i,s,t} [y_{i,s,t}^o \log(F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma)) + (1 - y_{i,s,t}^o) \log(1 - F(\mathbf{x}_{i,s,t}\beta) F_o(\mathbf{z}_{i,s,t}\gamma))],$$

and maximize it to obtain estimates of β and γ .

The results from this regression are depicted in table 1.3. Again, the results from the regression are very similar to the ones obtained earlier in section 1.4.2, table 1.2. Looking at the variables of interest, they exhibit the same patterns as the

ones derived earlier: older reviews are more likely to be fake, and the probability of a review being fake is smaller for very low and very high levels of reputation $\mu_{i,s,t}$.

	variable	estimates	p-values	std errors
x	Constant	-164.6***	3.58e-14	2.17e+01
	$\mu_{i,s,t}$	828.9***	2.25e-12	1.18e+02
	$\mu_{i,s,t}^2$	-663.9***	3.37e-12	9.54e+01
	time	-0.016***	3.506e-18	1.82e-03
z	Constant	2.07***	0.0000	7.02e-02
	Dummy for text reliability	-2.605***	0.0000	6.29e-02
	Dummy for title reliability	-1.53***	0.0000	6.86e-02
	Numb. helpful feedback	-0.015	1.00014e-02	5.804e-03
	Verified Purchase	-0.285***	1.019621e-05	6.468e-02
	Has images or videos	0.613**	7.7998e-12	8.95e-02
Observations: 18,440		pseudo R^2 : 0.3953554		

Table 1.3: Logit regression after correcting for endogenous classification errors.

1.4.4 Alternative database

As mentioned at the beginning of section 1.4, the database collected from sellers who were either caught soliciting fake reviews or were flagged by users for being involved in suspicious activity may suffer from selection bias. Indeed, by focusing the analysis on those sellers, the resulting sample may end up with an

overrepresentation of fake reviews, which could then affect the resulting estimates from the Logit regressions. Moreover, restricting the analysis to those sellers may limit the overall sample size, as manually finding suspicious sellers is a tedious and time consuming process. And finally, the resulting dataset is highly heterogenous, as it includes several different types of products, ranging from cheap electronic devices to children’s toys, which can lead our model to be misspecified.

To address these issues, I collected a separate dataset comprised exclusively of wireless headsets sold at Amazon, not targeting any seller from such category in the sampling process. The reason I chose wireless headsets is because one can find evidence in the news that fake reviews for these products are very prolific on Amazon, thus making the analysis for this market economically relevant (see for instance *How merchants use Facebook to flood Amazon with fake reviews (April 23, 2018)*). The dataset was then used to estimate a model similar to the one presented in section [1.4.3](#).

The dataset is comprised of 278,829 reviews from 1,134 different headphone products. So sellers on average received approximately 246 reviews, which is significantly higher than the average number of reviews from the previous sample. But regarding the distribution of stars, they are very similar for both samples as depicted by figure [1.19](#).

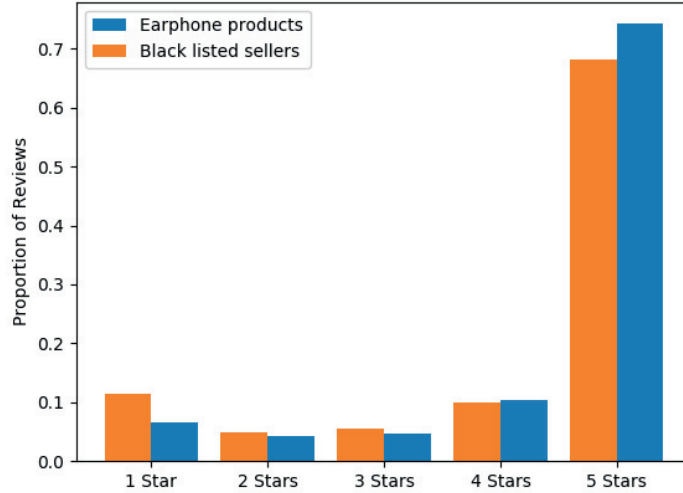


Figure 1.19: Histogram of the number of stars per sample. The bars in blue correspond to the sample of wireless earphone products, while the one in orange corresponds to the sample described in section 1.4.1 generated by targeting suspicious products that were either soliciting reviews in online platforms, or were flagged as suspicious on Amazon forums.

1.4.4.1 Fake review detection

Because there is no prior evidence to suggest that a particular seller from this new dataset solicited fake reviews, I no longer employ criterion number 1.4.1.1 presented in section 1.4.1.1 in the fake review detection process. Instead, I rely solely criterion 1.4.1.1, i.e., I classify a review as fake if its text content is sufficiently similar to some other review from the sample.⁸ Applying this unique classification rule to the sample results in 37,921 of the total of 278,829 reviews being classified as fake,

⁸Moreover, the computational burden from criterion II*) increases, while its efficacy decreases as the sample size increases, which is another reason not to use this criterion on the new dataset.

which amounts to approximately 19% of reviews.

While this percentage may already seem alarmingly high, it greatly underestimates the actual proportion of fake reviews for this market. Indeed, just as an illustration, consider a couple of products in the sample for which the Jaccard similarity index accused less than 15% of their reviews from being fake. By inspecting those two products more closely, one can find that: 1) more than 99% of their reviews were 5 star and unverified purchase reviews, and 2) they were mostly concentrated around a few days during the time period reviews were posted for these products, as depicted in figure 1.20. So it is safe to assume that for these two products the Jaccard similarity index alone was not capable to capture all suspicious activity. For this reason we add in our Logit regression variables aimed at detecting potential classification errors, such as adding dummies that assume value 1 when the review was posted during a spike of positive reviews. The criteria for detecting those spikes is explained in more detail on the appendix session [A.0.3.2](#).

Now to see how the detection of fake reviews using Jaccard similarity compares with other detection methods, I compute the average number of fake reviews obtained through this method for each different grade category received on *Fakespot.com*, a platform dedicated to detecting fraudulent reviews on online rating platforms such as Amazon and Yelp. Figure 1.21 displays the average number of fake reviews per product for each grade category, where “A” corresponds to the best possible grade (i.e., to the lowest level of review manipulation) that a product can get, and “F” corresponds to the worst grade possible. The plot depicts a negative relationship between grades and the number of reviews detected as fake using the

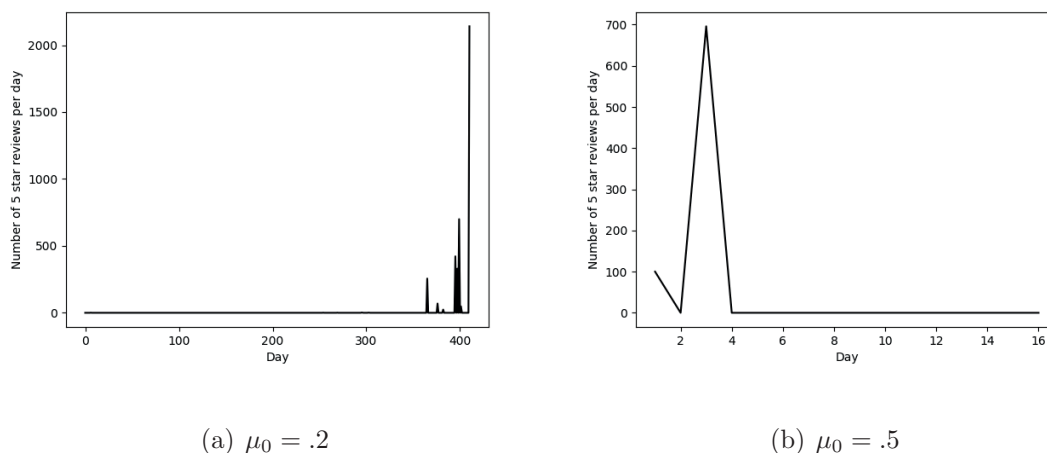


Figure 1.20: Number of 5 star reviews received by a couple of products per day. Product 1 is no longer sold at Amazon, perhaps because Amazon detected suspicious activity surrounding its reviews and thus had the product removed. Regarding product 2, as I write this on May 16, 2019, though it is still sold on Amazon, all its positive reviews (4 and 5 stars) have been removed.

Jaccard similarity index. If the relationship is not perfectly decreasing, that is probably due to the combination of two factors: 1) We are using a single criterion to detect fake reviews, namely, the level of text similarity among the reviews, whereas Fakespot seems to use a machine learning algorithm that computes the probability of a review being fake using a combination of several different criteria; 2) Moreover, Amazon has its own fake review detector, and it excludes fraudulent reviews from its platform on a regular basis. This implies that on several occasions we would encounter a product that engaged in a lot of suspicious activity, and yet had a high score on Fakespot, solely because their fraudulent reviews were removed by Amazon by the time we checked its grade. To mitigate the selection bias caused by hav-

ing Amazon removing some of the suspicious reviews from the platform, for many products in our sample we scraped their corresponding reviews at several different days, so that our database contains many reviews that were filtered by Amazon, in addition to reviews that were posted after Amazon’s filtering.

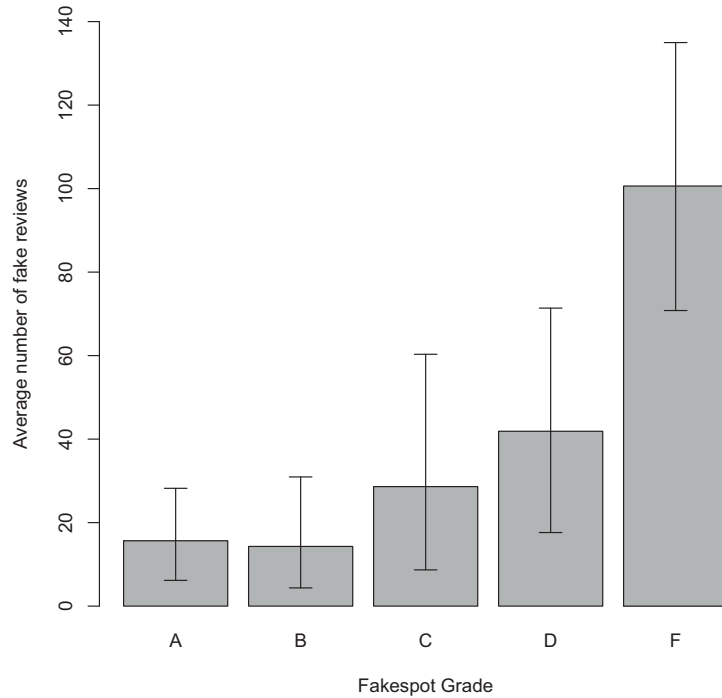


Figure 1.21: Average number of fake reviews detected using text similarity, compared to the product’s grade on Fakespot.com. According to the website, a grade of “A” indicates low level of review manipulation, whereas a grade of “F” indicates a high number of fraudulent reviews. The 95% confidence intervals displayed in the figure were built using 100,000 bootstrap simulations.

1.4.4.2 Logit Regressions

Table 1.4 reports the results from standard Logit regressions. The results are mostly similar to the ones obtained before: for the fully specified model, the coefficients of interest follow the right direction predicted by the theoretical model, namely, that as times goes by, reviews are less likely to be fake; and that the relationship between reputation and the probability of a review has a downward parabola shape.

Also, once we add the dummy coefficient that assumes value 1 when the review was posted during an abnormal peak of 5 stars, we see that reviews are more likely to be classified as fake during those periods. That is not surprising given that in our sample 45% of the reviews posted during abnormal peaks were classified as fake, while only 17% of the reviews outside those peaks were classified as fake.

Moving to the model specification that addresses classification error, we get coefficients similar to the ones obtained using the previous dataset, as displayed in table 1.5. The only main difference between the two regressions are the signs from the coefficients for reputation, which now display signs not consistent with the inverted U shape relationship between reputation and effort on review manipulation. But other than that, all other coefficients exhibit the same signs as in the previous regressions.

	<i>Dependent variable:</i>	
	$y = \mathbb{1}(\text{review is fake})$	
	(1)	(2)
Constant	−0.457*** (0.156)	0.156 (0.129)
μ	6.006*** (0.459)	−4.615*** (0.378)
μ^2	−3.851*** (0.320)	4.218*** (0.263)
time	−0.001*** (0.0001)	−0.006*** (0.0001)
Peak dummy	0.512*** (0.018)	
Dummy for text reliability	−2.147*** (0.013)	
Dummy for title reliability	−1.175*** (0.017)	
Numb. helpful feedback	0.002*** (0.0003)	
Verified Purchase	−1.313*** (0.019)	
Has images or videos	0.166*** (0.036)	
Observations	232,176	232,176
Log Likelihood	−88,368.930	−122,434.000
Akaike Inf. Crit.	176,757.900	244,876.100

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1.4: Simple Logit regressions using the earphone dataset.

	variable	estimates	p-values	std errors
x	constant	70.9***	4.36e-20	7.72
	$\tilde{\mu}_{i,s,t}$	-2.49e+02***	6.33e-19	28
	$\tilde{\mu}_{i,s,t}^2$	2.26e+02***	5.52e-19	25.352
	time	-2.87e-03***	2.40e-14	3.76e-04
z	constant	1.86***	0.000	2.765e-02
	peak dummy	0.523***	0.000	1.867e-02
	Dummy for text reliability	-2.22***	0.000	1.41e-02
	Dummy for title reliability	-1.24***	0.000	1.824e-02
	Numb. helpful feedback	2.29e-03***	1.80e-10	3.596e-04
	Verified Purchase	-1.35***	0.000	1.954e-02
	has images or videos	0.16***	1.29e-05	3.6724e-02
	Observations: 232,176	pseudo R^2 : 0.2933641		

Table 1.5: Logit regression after correcting for endogenous classification errors.

1.4.5 Placebo test

A placebo test was conducted in order to certify that the correlations obtained in the previous sections were not spurious. To perform the test, we first randomly classify reviews as fake or real according to the empirical distribution from the sample. Since roughly 19% of the reviews in the sample were classified as fake, we randomly choose a review to be assigned as fake with probability .19. Then we run

the same regressions as before with this new random assignment.⁹ The results from those regressions are displayed in tables 1.6 and 1.7, from which one can see that, with the exception of intercepts, all coefficients are statistically insignificant.

1.5 Conclusion

This paper develops a theoretical model in which sellers dynamically choose the effort spent on review manipulation. One of the predictions from the model is that the effort spent on review manipulation tends to be concentrated during the initial periods following a seller’s entrance (or reentrance with a new brandname) into the market, since sellers wish to maximize the impact from each fake review. Another prediction from the model is that the amount of effort dedicated in review fraud does not vary monotonically with the firm’s reputation. Indeed, sellers that currently possess a very good or very bad reputation will usually spend less effort in review manipulation, the intuition being that, for very low levels of reputation, the seller finds it too costly to signal that it is of high quality, whereas a seller that has already accumulated a very good reputation does not need to prove that it is a high quality type. In mathematical terms, very high or very low reputation levels are absorbing states: the closer a seller is to those states, the harder it is to depart from them, which in turn defeats the purpose of trying to influence signals generated from reviews.

Another interesting prediction from the theoretical model is that low quality

⁹The variables “Reliability index from review title” and “Reliability index from review text” were also recomputed based on the new assignment.

	<i>Dependent variable:</i>	
	$y = \mathbb{1}(\text{review is fake})$	
	(1)	(2)
Intercept	1.142*** (0.131)	1.141*** (0.129)
Time	−0.0001 (0.0001)	−0.0001 (0.0001)
$\tilde{\mu}$	0.160 (0.386)	0.221 (0.380)
$\tilde{\mu}^2$	−0.127 (0.269)	−0.174 (0.265)
peak dummy	−0.010 (0.017)	
Reliability index from review title	−0.001** (0.0003)	
Reliability index from review text	−0.00002 (0.00002)	
Numb. of words	−0.0001 (0.0003)	
Numb. helpful feedback	0.004 (0.016)	
Has Images	−0.020 (0.027)	
Has videos	0.049 (0.124)	
Observations	232,176	232,176
Log Likelihood	−125,382.700	−125,387.500
Akaike Inf. Crit.	250,787.400	250,782.900

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1.6: Placebo Tests
66

	variable	estimates	p-values	std errors
x	intercept	0.882	0.919	8.655
	$\tilde{\mu}_{i,s,t}$	14.539	0.555	24.639
	$\tilde{\mu}_{i,s,t}^2$	-11.553	0.494	16.908
	time	0.093	0.422	0.116
z	intercept	1.204***	0.000	0.020
	peak dummy	-0.013	0.449	0.018
	Reliability index from review title	-0.063	0.029	0.029
	Reliability index from review text	-0.002	0.218	0.002
	Numb. helpful feedback	-0.000	0.870	0.000
	Verified Purchase	-0.012	0.512	0.019
	has images	-0.016	0.560	0.028
	has videos	0.052	0.674	0.125
	Observations: 204,219	pseudo R^2 : 4.95e-05		

Table 1.7: Placebo test for the Logit regression, correcting for classification error.

sellers do not necessarily exert more effort on the fabrication of fake reviews as compared to high quality sellers. Indeed, which type spends most effort on review manipulation depends on the current level of reputation held by the seller. For very low levels of reputation, it is the high quality seller that spends most effort in review manipulation, while the opposite holds for low levels of reputation.

While the benchmark version from the model predicts that in the long run

buyers eventually learn the true type from the seller; by allowing sellers to exit and reenter the market with a new name, we observe that low quality sellers tend to resort to this tactic very frequently, thus preventing customers from learning the true type from low quality sellers in the long run. So this result suggests that one way of making reviews more informative in online platforms such as Amazon and TripAdvisor, is by building obstacles that prevent sellers from anonymously selling their products under different account names.

In order to verify some of the model's predictions empirically, I collected data from products for which sellers were soliciting fake reviews on Amazon. After classifying the reviews posted on those products as fake or real, I then ran a Logit regression to estimate the probability of a review being fake as a function of the seller's outstanding level of reputation and the time it took for the review to be posted. Consistent with the predictions from the theoretical model, I find that the probability of a review being fake decreases with time, and it varies non-monotonically with the seller's reputation, where very high or very low reputation levels are associated with a lower probability of the review being fake.

These results have potential practical applications when it comes to fake review detection. Indeed, the performance from machine learning algorithms can potentially be improved with the inclusion of a measure of sellers' reputation as well as the time it took for the review to be posted since the seller entered into the market, as predictor variables. In research currently underway, I plan to compare the prediction power from neural network algorithms that include and exclude these variables as predictors.

Chapter 2: Opinion Polarization in the presence of noisy and biased channels

Polarized opinions are often observed on topics that have objective truths. Such polarization should not occur if the information available to the public is perfectly reliable, or if everyone obtains information through the same channels. I develop a dynamic model in which rational Bayesian updaters select among different news channels in hopes of learning the truth about a certain state of the world. The signals generated from news channels are not perfectly accurate (i.e., the news can be imprecise or slanted), and those who read the news take that into account when updating their beliefs regarding the topic in question. A novel feature from my model is that it allows agents to develop different opinions regarding the reliability of each news channel. Although agents initially start with the same priors regarding the relevant state of the world and the quality of the news channels, their posterior beliefs may diverge, as they gather information from different outlets. Preliminary evidence seems to suggest that an increase in the level of bias from news channels can actually help customers distinguish the direction of bias from each channel, thus improving their ex ante utility. However, results on polarization still require further tests.

2.1 Introduction

In the last years, many countries have observed a surge in polarized political opinions. Many attribute that to an increase in the dissemination of false information perpetrated by malicious agents, which are intensified by social media platforms such as Facebook and Twitter, which, some would argue, create an echo chamber environment in which agents cherry pick posts that are likely to confirm their prior beliefs.

But if agents are utility maximizers and they learn from their mistakes, then why should they systematically pick information from unreliable channels?¹ This paper aims to model the process that leads to such behavior. The model assumes agents to be rational Bayesian updaters who wish to seek the truth about a certain state of the world in order to take some action that directly affect their utility. For example, one may want to know the true financial health of a certain company in order to decide whether or not to invest in it; or to know whether or not vaccinations

¹Following the notation from the seminal paper [Shannon \(1948\)](#) which initiated the field of Information Theory, I denote a news channel as any intermediate agent responsible for the transmission of some original piece of information. So that includes not only TV news channels such as Fox News and CNN, but also newspapers, blog channels or word of mouth. Perhaps because the term “channel” is usually associated with TV channels, some economists prefer using the term “news source”, which I avoid using here as the term “source” has a different connotation in Information Theory.

against measles disease cause autism, so as to decide whether or not to have their kids vaccinated.

Our model is comprised of a discrete and finite set of news channels, with accuracy that is unknown to customers. So upon reading news from a certain channel, the agent not only updates his beliefs pertaining the fact reported, but also his beliefs regarding the quality of the news channel. Different news channels are assumed to be equally costly, so that the ultimate reason someone prefers to watch, say, Fox News, over CNN, has nothing to do with the former being a cheaper source of information than the latter, but rather, it is due to the viewer’s past experience with each of those news channels.

Though the results from the model are quite preliminary, simulations seem to suggest that, by keeping the overall level of precision from channels constant while increasing their bias has little to no impact on the observed level of opinion polarization. Surprisingly, increases in the level of bias from news channels can actually improve customers ex ante utility.

2.2 Related literature

This paper is closely related to [Nimark and Sudaresan \(2018\)](#), which also devise a theoretical framework in which agents dynamically acquire information through noisy and biased channels in order to learn the truth about a certain state of the world. However, in their framework agents are assumed to know ex ante the exact level of bias and precision from each available news channel, so that the

only reason they would pick poor and unreliable channels is because they cost less. However in our modern age of information and technology, it is relatively easy to find examples in which reliable information can be obtained at virtually zero monetary cost. Indeed, on the internet one can find for free several credible news articles and scientific papers that are only one click away from unreliable conspiracy theory websites. So the assumption that the monetary costs of different news channels is what drives polarization of beliefs may not apply to certain contexts.²

In fact, in many instances agents seem eager to get information from low quality channels because they overestimate the reliability from those channels, not just because they are cheaper. As an example, when it comes to gathering information as to whether vaccinations can cause autism or other diseases, publicly available information (or misinformation) can be found supporting either state of the world. So one of the main motivations of my model is to relax the assumption that more reliable channels cost more, and instead, assume that agents try to pick the best channels at their disposal, while at the same time allowing them to dynamically learn the quality from those channels.

²Nimark and Sudaresan (2018) justify the assumption that information from more precise channels are more costly because they are more technical, and therefore harder to understand. They even mention as an example that reading and understanding a scientific article from Nature concerning global warming is arguably more costly than watching a heated and sensationalistic debate on the television addressing the topic. However, in their model the signal provided by each news channel is binary (e.g., high or low). And the cost from processing a binary signal should not depend on its precision. So implicitly, they are assuming that more precise information are sold at a higher price (i.e., the cost is generated from acquiring, not processing information).

In a different framework, [Gentzkow and Shapiro \(2006\)](#) develop a static model in which Bayesian updaters seek to learn the truth about a certain state of the world through noisy channels; but now customers do not know the reliability of each channel, leading them to update not only their beliefs regarding the state of the world, but also on the reliability of the news channel that they picked. The model also endogenizes the amount of slant coming from each channel: sellers may want to bias the information they provide towards readers' priors, so as to improve their perception among the public. But due to its static nature and its focus on analyzing the news providers' incentives to slant information, it is not an ideal framework for explaining how polarization on beliefs are generated.

Another set of related papers are [Mullainathan and Shleifer \(2005\)](#), [Xiang and Sarvary \(2007\)](#) and [Yildirim, Gal-Or and Geylani \(2013\)](#), which, like [Gentzkow and Shapiro \(2006\)](#), mostly focus on the incentives from channels to slant the news in order to attract a bigger audience. But different from [Gentzkow and Shapiro \(2006\)](#), these papers make reduced form assumptions regarding agents' utilities. [Mullainathan and Shleifer \(2005\)](#), for instance, make the reduced form assumption that slanted news reduces one's utility quadratically. Moreover, it assumes that, for biased readers, the more distant one's posterior belief is from her prior, the lower is her ex post utility from having gathered information from that source. My paper distances itself from this approach by assuming instead that all agents in the economy are rational expected utility maximizers who wish to learn the true state of the world in order to make an optimal binary decision, such as whether to invest in a certain company, or whether to vaccinate their kids against measles disease. This

way, tendencies to flock toward information sources that confirms one’s priors will arise naturally from rational Bayesian update, and not due to some ad hoc utility function that penalizes posteriors that drift away from one’s prior.

Also, as mentioned above, these papers mostly focus their analysis on the news providers’ incentives to slant information. While endogenizing the overall amount of slanting in the system has the potential of enriching the model, it reduces its tractability, and in some instances may even add undesirable distortions. [Xiang and Sarvary \(2007\)](#) for instance recognize that one concern surrounding their specification is that they make the requirement that news channels that have best access to information are the ones that disseminate most misinformation. That is because they assume that, since those channels have gathered more information, they are in a position of displaying only a small fraction of such information. So implicitly they assume that news channels are not capable of lying: they can only omit certain truths and replace them with “alternative facts”, so that the more information they gather, the more “truths” can be omitted in their broadcasts. As a result, it becomes actually more costly to sell slanted news, as opposed to selling unbiased information. While understanding the incentives that lead channels to become more or less biased can have several useful policy applications, in my model I abstract from those issues by treating news bias completely exogenous, and focus entirely on the impact that those biases have on agents’ beliefs.

In a somewhat different branch of the literature, papers like [Papanastasiou \(2017\)](#) and [Bikhchandani, Hirshleifer and Welch \(1992\)](#) focus on informational cascades models, which address the process through which misinformation can be dis-

seminated perpetually in a chain of information sharing. The main intuition from such result lies in the fact that, if an information has been shared by a sufficiently high number of individuals before reaching a new agent, then the agent will think that the information was most likely already fact checked by at least someone else that preceded the agent in the chain. If the information was false, then whoever fact checked the news would not have passed it on; but since it was passed on, the agent concludes that the information is most likely true. Because the agent assumes the information to be true with high probability, he doesn't bother to fact check it before sharing it with the next person.

Building on this principle, [Azzimonti and Fernandes \(2018\)](#) write a model in which agents dynamically share their beliefs in a social network. The network allows one to add bots or other malicious agents to check the overall effect that they have on opinion polarization. One drawback from their specification, however, is that the connections on the network are completely exogenous. Moreover the model relies on heuristic assumptions that forces agents to listen to information from biased and potentially malicious channels, when unbiased channels are freely available. So an interesting research project would be to adapt [Azzimonti and Fernandes \(2018\)](#) framework by making network connections and agents' attention to each connection endogenous.

2.3 Model

An economy is comprised of N news channels and M agents, indexed by n and m , respectively. News channels convey information about a certain company c . The company has quality $q \in \{q_H, q_L\}$, where $1 \geq q_H > q_L \geq 0$. Time is discrete and infinite, and at each period $t \in \{1, 2, \dots\}$ the company generates a random signal $y_t \in \{-1, 1\}$, where the probability $y_t = 1$ is given by $q \in \{q_H, q_L\}$ (because $q_H > q_L$, this implies that a high quality company is more likely to generate a positive signal at each period).

At each period, news channels can either report the realized signal accurately or inaccurately, depending on their precision, which we shall define next. Following [Nimark and Sudaresan \(2018\)](#), the precision from a channel n consists on a pair of probabilities $(p_H^n, p_L^n) \in [0, 1]^2$, where p_H^n is the probability that the news channel reports a high signal when the actual signal is high, and p_L^n is the probability that the news channel reports a low signal when the actual signal is low. As an example, $(p_H^n, p_L^n) = (1, 1)$ would be a perfectly precise channel (it always reports the true signal), whereas any news channel such that $p_H^n + p_L^n = 1$ would be completely uninformative/imprecise, as it would report either signal with probability .5 regardless the true quality of the company. The precision can also capture the level of bias from a certain news channel. For instance, a channel with precision $(p_H^n, p_L^n) = (1, 0)$ would always report favorable news pertaining the company, so one could interpret such channel as being biased towards praising the company (as it turns out, such news channel is also completely uninformative, since it always praises the company

no matter what the actual fact is).

We denote $p = (p_H^n, p_L^n)_{n=1}^N$ as the vector of precision of the N channels, and assume that the quality from the company, q , and vector of precision p from the news channels are randomly drawn from the joint distribution $f_{q,p}(\cdot)$. We denote $f_q(\cdot)$ and $f_p(\cdot)$ as the marginal density of q and p , respectively. The quality from the company, q , and precision from the news channels, p , remain constant throughout the game, and they are both unknown to customers. Customers wish to learn the true type of the company in order to make an optimal investment decision in every period. If the company is of high quality and the consumer invests in the company in period t , he gets a payoff of 1 in that period, while his payoff from not investing is zero; but if the company is of low quality, then the agent gets a payoff of 0 from investing, and 1 from not investing. So the agent essentially gets a payoff of 1 when he makes the right decision, which is to invest when the company is of high quality, and not invest, when the company is of low quality; and a payoff of 0 when he makes the wrong decision, which is to invest when the company is of low quality, and not invest, when the company is of high quality. So if at a certain period an agent believes the firm to be of high quality with probability $Prob(q = q_H) \geq .5$, then he should invest in the company in that period, thus resulting in an expected payoff of $Prob(q = q_H)$ for that period, whereas if $Prob(q = q_H) < .5$, the agent should not invest, resulting in an expected payoff of $1 - Prob(q = q_H)$.³

³Though we framed the model in terms of investment decisions, it might also have applications in other areas, such as understanding the opinion formation of political candidates, or topics that have objective truths, such as whether or not global warming is manmade.

At each period t an agent can either pick a news from one of the n channels at a cost $C > 0$, or pick no news at all. We assume that the cost C is the same for every news channel, regardless of their precision. This assumption diverges from the method utilized by [Nimark and Sudaresan \(2018\)](#), where they assume that the cost of processing information from a channel is proportional to the channel capacity. We depart from that assumption because, though [Shannon \(1948\)](#) has proven that the maximum rate at which one can learn an unobserved signal through a noisy channel is bounded above by the channel's capacity, in our model agents can only pick a single observation from a channel in each period, and the cost from processing a binary datapoint (1 bit) should not depend on its precision. The only reason one would use channel capacity to measure the cost of processing a binary datapoint is if one believes that the market equilibrates to the point that more reliable information are *sold* at a higher price. But one can find anecdotal evidence in which that is not always the case, in fact, many channels with different degrees of reliability are freely available on the internet.

In the event the agent picks a news from channel n , he updates his beliefs regarding q and the distribution of the precision (p_H^n, p_L^n) from that news channel conditional on q , through Bayes' rule. In the event the agent does not read any news at time t , he does not make any updates.

After updating their beliefs, agents make their investment decision. As described earlier, if their updated beliefs on the firm's quality are such that $Prob(q = q_H) \geq .5$, then they should invest in that period, thus resulting in an expected payoff of $Prob(q = q_H)$ for that period, whereas if $Prob(q = q_H) < .5$, then they should

not invest, resulting in an expected payoff of $1 - \text{Prob}(q = q_H)$.

So to summarize, the timeline of the model goes as follows: 1) at the beginning of each period agents pick a signal from one of the news channels (or no signal); 2) after they choose a news channel and observe the signal, they update their beliefs pertaining the quality of the company, and the precision from the news channel that they picked; 3) after that, they make an investment decision (invest, or not invest); 4) and finally, they go to the next period and repeat the same process iteratively (see figure 2.1).

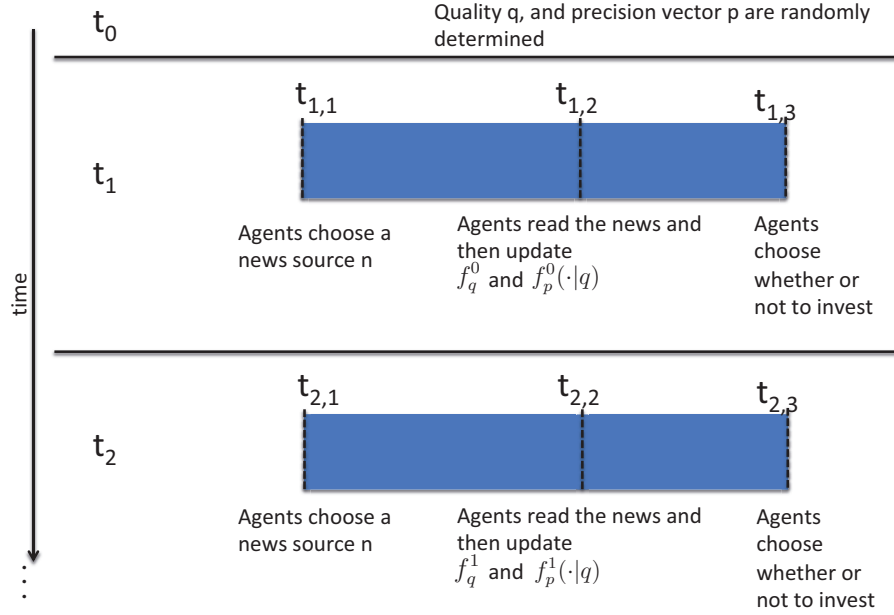


Figure 2.1: Schematics for the timeline of the model, where $f_q^t(\cdot)$ represents the beliefs that the agent has at the beginning of period t regarding the quality of the company; and $(f_p^{n,t}(\cdot|q))_{n=1}^N$, the beliefs that the agent has at the beginning of period t on the precision from each channel, conditional on the quality of the company.

Assuming agents have a discount factor $\beta \in [0, 1)$, we have that their dynamic problem is a solution to the following Bellman equation

$$V(f_q^t, f_p^t(\cdot|q)) = \max_{n \in \{0, 1, 2, \dots, N\}} \mathbb{E}_t(U_n^{t+1}) - C\mathbb{I}_{n \neq 0}(n) + \beta \mathbb{E}_t[V(f_q^{t+1}, f_p(\cdot|q))], \quad (2.1)$$

where the state variables are $f_q^t(\cdot)$, the beliefs the agent has at the beginning of period t regarding the quality of the company; and $f_p^t(\cdot|q)$, the beliefs the agent has at the beginning of period t on the precision from each news channel, conditional on the quality of the company. The choices available to each player in every period consists on the set of channels the agent can choose, $n \in \{1, 2, \dots, N\}$, plus the option of $n = 0$, which corresponds to the action of not picking any channel. The term

$$U_n^{t+1} = \max\{(f_q^{t+1}(q_H), 1 - f_q^{t+1}(q_H))\} \quad (2.2)$$

is the expected utility of making the optimal investment decision in period t , after the agent has learned the signal from channel n and updated $f_q^t(\cdot)$ through Bayes' rule. Notice that we take an expectation on U_n^{t+1} , since the agent's beliefs regarding the quality of the company, $f_q^{t+1}(\cdot)$, might change depending on the news the agent receives in that period. Moreover, the cost C is only paid in the event the person picks one of the available news channels, which is why we multiply it by the indicator function

$$\mathbb{I}_{n \neq 0}(n) = \begin{cases} 1, & \text{if } n \neq 0 \\ 0, & \text{if } n = 0 \end{cases}$$

We now proceed to explain in detail how agents update their beliefs upon receiving a news, i.e., how to update f_q^t and $f_p^{n,t}(\cdot|q)$ in the event the agent picks channel n .

2.3.1 Bayesian updates

Recall from the previous section that we are assuming the company's quality q to be a binary random variable ($q \in \{q_H, q_L\}$). We will also assume that the channels' precision p are discrete random variables as well. For the case in which these random variables are continuous, one only needs to replace the summations from the formulas below by integrals.

So suppose an agent starts a period with belief $f_q^t(\cdot)$ regarding the quality from the company in question, and conditional belief $f_p^t(\cdot|q)$ on news channels' precision. Then the prior joint distribution of the company's quality q and channels' precision p is given by

$$f_{q,p}^t(q, p) = f_q^t(q) f_p^t(p|q).$$

So upon reading a news $\hat{y} \in \{-1, 1\}$ from channel n in period t , where $\hat{y} = 1$ represents a news that portrays the company favorably, and $\hat{y} = -1$ a news that portrays the company unfavorably, we have that the updated joint distribution of q and p is given by

$$f_{q,p}^{t+1}(q, p|\hat{y}) = \frac{\text{Prob}(q \cap p \cap \hat{y} | f_{q,p}^t(q, p))}{\text{Prob}(\hat{y} | f_{q,p}^t(q, p))}$$

So if $\hat{y} = 1$, the updated joint distribution is given by

$$f_{q,p}^{t+1}(q, p) = \frac{f_{q,p}^t(q, p)(qp_H^n + (1-q)(1-p_L^n))}{\sum_{\tilde{q}} \sum_{\tilde{p}} f_{q,p}^t(\tilde{q}, \tilde{p})(\tilde{q}\tilde{p}_H^n + (1-\tilde{q})(1-\tilde{p}_L^n))},$$

whereas if $\hat{y} = -1$, the the updated joint distribution is given by

$$f_{q,p}^{t+1}(q, p) = \frac{f_{q,p}^t(q, p)(q(1-p_H^n) + (1-q)p_L^n)}{\sum_{\tilde{q}} \sum_{\tilde{p}} f_{q,p}^t(\tilde{q}, \tilde{p})(\tilde{q}(1-\tilde{p}_H^n) + (1-\tilde{q})\tilde{p}_L^n)},$$

Once the updated joint distribution is computed, one can recover the marginal distribution of q :

$$f_c^{t+1}(q) = \sum_{\tilde{p}} f_{q,\tilde{p}}^{t+1}(q, \tilde{p}),$$

which is then used to compute $Prob(q = q_H)$ from expression 2.2.

We can also recover the distribution of p conditional on q by computing:

$$f_p^{t+1}(p|q) = \frac{f_{q,p}^{t+1}(q, p)}{f_c^{t+1}(q)}.$$

After observing an event that happens with zero probability, the agent's posterior can take any form. So for example, if $Prob(\hat{y}|f_{q,p}^t(q, p)) = 0$, then we can assume without loss of generality that $f_{q,p}^{t+1}(q, p|\hat{y}) = .5$.

2.4 Existence and uniqueness of a solution to the Bellman equation

Proposition 2.4.1 *There exists a unique $V(\cdot, \cdot)$ that is a solution to the Bellman equation 2.1.*

Proof: Let $C(X)$ be the set of real bounded continuous functions with the sup norm defined over the set of possible probability mass functions for q and p conditional on q . If $V \in C(X)$, then clearly the function

$$T(V) = \max_{n \in \{0,1,2,\dots,N\}} \mathbb{E}_t(U_n^{t+1}) - C\mathbb{I}_{n \neq 0}(n) + \beta \mathbb{E}_t[V(f_q^{t+1}, f_p(\cdot|q))] \quad (2.3)$$

also belongs to $C(X)$. Indeed,

$$\mathbb{E}_t(U_n^{t+1}) = \sum_{\hat{y}} Prob(\hat{y}|f_{q,p}^t(q, p)) \max\{(f_q^{t+1}(q_H|\hat{y}), 1 - f_q^{t+1}(q_H|\hat{y}))\}$$

is clearly continuous and bounded (above by 1 and below by 0). So for every $n \in \{0, 1, 2, \dots, N\}$, the expression

$$\mathbb{E}_t(U_n^{t+1}) - C\mathbb{I}_{n \neq 0}(n) + \beta \mathbb{E}_t[V(f_q^{t+1}, f_p(\cdot|q))]$$

is bounded and continuous since it is a convex combination of bounded and continuous functions. Because the maximum of a finite set of bounded and continuous function is also bounded and continuous, we have that $T(V)$ is bounded and continuous.

Now the operator $T : C(X) \rightarrow C(X)$ clearly satisfies the Blackwell sufficient conditions for a β -contraction. Because $C(X)$ is a Banach space, this implies that the operator $T(\cdot)$ has a unique fixed point in $C(X)$, by the contraction mapping theorem. ■

2.5 Simulations

Figure 2.2 provides [Esteban and Ray \(1994\)](#) measure of polarization for different levels of news bias, when there are only 2 news channels available, and where q and p can each assume only 2 possible distinct values (if q and p had a bigger support, the number of state variables from the functional equation 2.1 would be larger, thus triggering curse of dimensionality issues).

Figure 2.3 displays the ex ante utility that agents get as a function of media bias, when the game lasts for only 5 periods (computations for an infinite version of the model exhibits the same pattern) which is essentially the value function from the first period evaluate at agents' prior beliefs. As displayed in the figure, agents'

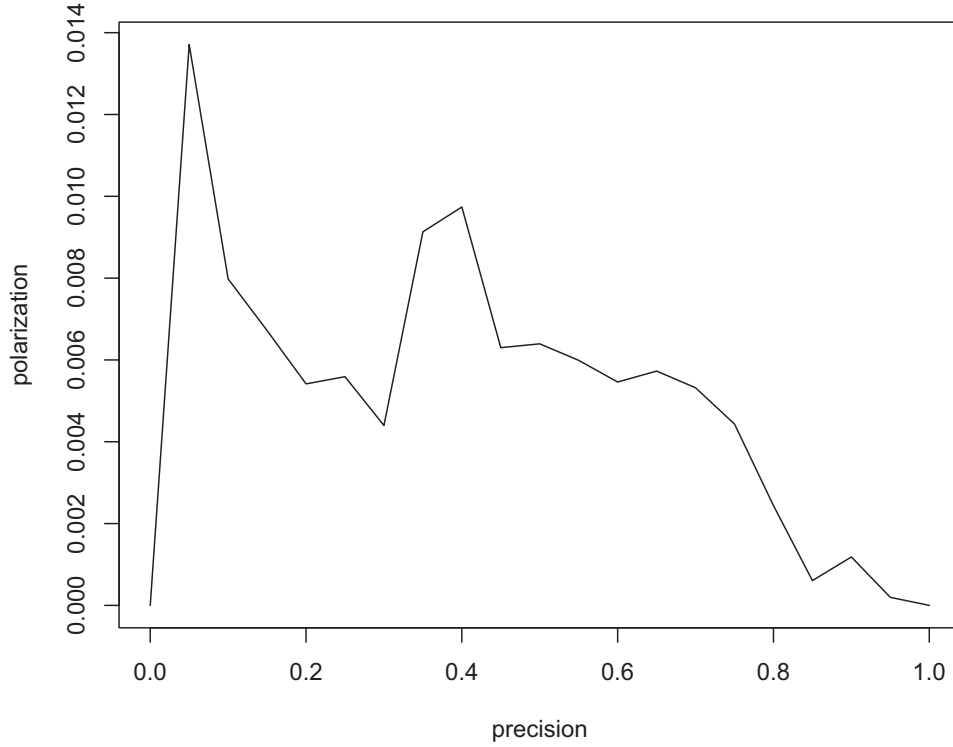


Figure 2.2: Long run polarization of q when $C = .001$ as a function of ρ , where one of the news channel has high precision $(1, \rho)$, while the other has low precision $(\rho, 1)$, and agents are not allowed to fact check the news. The discount factor β equals .5, and all agents start with the same uninformative priors.

expected utility actually increase with media bias, probably due to the fact that, with an increase in bias, agents can more easily discern the direction of channel bias, thus simplifying their decision making process.

2.6 Allowing agents to fact check the news

In the model described in section 2.3, agents had no control over the quality of information they received from each channel. But in practice one might expect

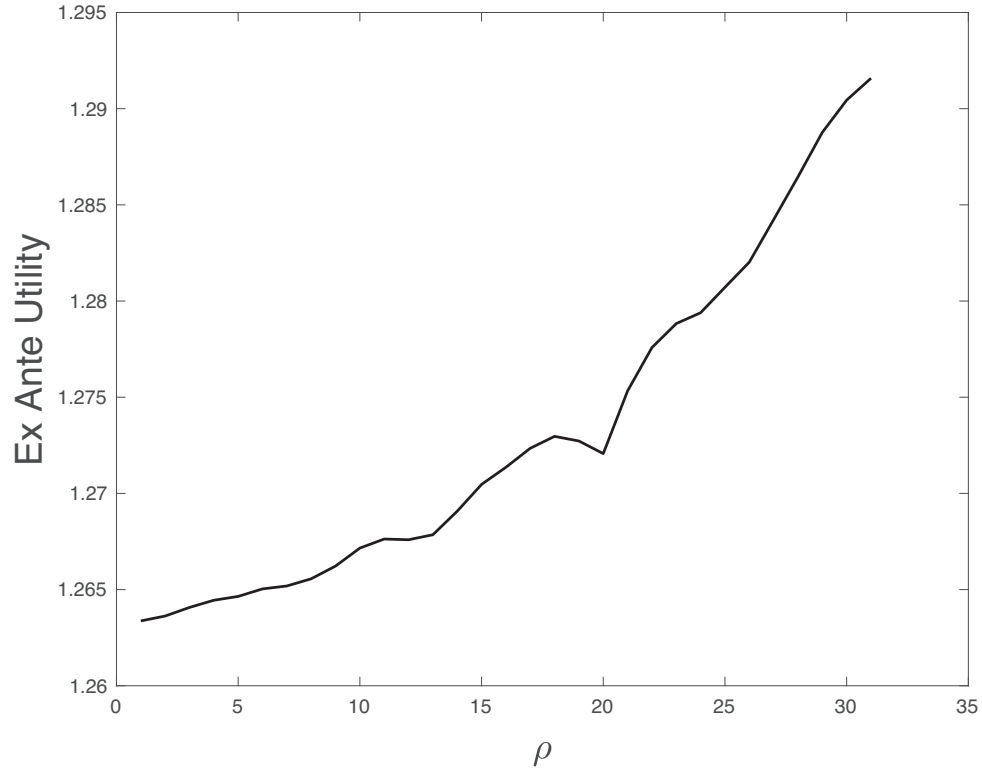


Figure 2.3: Long run polarization of q when $C = .001$ as a function of ρ , where one of the news channel has high precision $(.6 - \rho, .6 + \rho)$, while the other has low precision $(.6 + \rho, .6 - \rho)$, and agents are not allowed to fact check the news. The discount factor β equals .5, and all agents start with the same uninformative priors, and the game is played for 5 periods.

agents to have the ability to collect more reliable information by exerting a higher effort on their search for knowledge. Indeed, upon reading a certain news article that discusses the financial health of a corporation, the reader might be able to check the reliability of the information provided in the article by comparing it with other sources of information, including raw data on the firm's stock value, or data collected from the firm's annual report.

So now we will assume that, upon reading a news, agents can pay an additional

cost to fact check the signal provided. More precisely, after receiving a signal \hat{y} , readers can pay an additional cost K to learn the true signal y generated by the company at that time period. Paying that additional cost would help agents to learn the precision of the channels and quality of the company at a faster rate.

Adding this fact checking mechanism still does not contribute to generating a pattern of the effect of bias on polarization. So this is a topic that deserves future study.

2.7 Conclusion

Preliminary evidence seems to suggest that an increase in the level of bias from news channels can actually help customers distinguish the direction of bias from each channel, thus improving their ex ante utility. The seemingly no effect that bias has on polarization is concerning, and probably requires some model modifications. I believe changing the fact checking mechanism by making it more similar to [Gentzkow and Shapiro \(2006\)](#), would probably solve the issue. More precisely, instead of assuming that agents pay a fixed cost to fact check the news they just received, at the end of each period some agents randomly receive the information as to whether the news they read is fake or real.

Chapter 3: An Improved Bound to Manipulation in Large Stable Matches

This paper derives an upper bound to the expected proportion of agents that have incentives to misreport their true preferences or vacancies in many-to-one stable matching mechanisms. The paper shows that the upper bound converges to zero as the number of participants in the market goes to infinity at a faster rate as compared to previous results in the literature. Moreover, the paper relaxes some of the assumptions that are typically made in order to get this type of result. In particular, it relaxes a requirement that would cause the market to end up with many unfilled positions, thus causing the convergence result to be applicable to more competitive environments. So this paper adds evidence to the fact that, though stable matching mechanisms are not entirely strategy-proof, in practice, when the number of participants in the market is sufficiently large, they can be treated as being effectively strategy-proof.

3.1 Introduction

Centralized two sided matching mechanisms have been widely used in the US and abroad to allocate scarce resources in situations where market based solutions

(such as auctions) are not ideal or can be perceived as unfair. Examples include the college admission problem and matching programs for medical residency ([Roth and Stomayor \(1992\)](#)). Ideally, a matching mechanism should satisfy some set of desirable properties. One such property is *stability*. A matching mechanism is said to be stable if it is individually rational (i.e., agents weakly prefer the match they get under the mechanism over staying unmatched) and if every pair of agents prefers the match they get under the mechanism over being matched to one another. One of the benefits of a matching mechanism being stable is that it gives agents incentives to follow through the recommendation made by the matchmaker, not to mention that it causes agents to perceive the resulting allocation as fair.¹

It is a well known result, however, that a stable matching mechanism can not be strategy-proof for both sides of the market. In other words, in a stable matching mechanism, there always exist a set of preferences such that revealing one's true preferences and vacancies is not a dominant strategy for at least one side of the market. The lack of strategy-proofness can affect a mechanism's efficacy. Indeed, in the context of the college admission problem, if truthful reporting is not a dominant strategy, students or colleges can end up wasting too much effort acquiring information on other agents' preferences and vacancies before making their

¹Making sure that a mechanism is fair is particularly important when it comes to matching design given that one of the motivations for implementing a matching mechanism in the first place as opposed to market based solutions such as an auction, is because market based solutions tend to be perceived as unfair in certain contexts (e.g., it would be unfair if only the rich kids got allocated to the top schools).

optimal decisions. Moreover, if agents misreport their preferences or vacancies, then the resulting allocation may not be stable with respect to the true preferences and vacancies, which can then lead some agents to either sue the matchmaker under the basis that the resulting allocation is *unfair*, or ignore its recommendation. And finally, colleges may wish to underreport their number of vacancies in order to attract better students, which can then lead to an overall shortage of vacancies offered by colleges.

Though this impossibility result may cause concerns surrounding the implementation of stable matching mechanisms, simulation exercises have suggested that in practice the proportion of agents that can successfully manipulate stable matching mechanisms is relatively small, and the differences in allocations obtained by implementing different stable matching mechanisms also tends to be negligible (see [Roth and Peranson \(1999\)](#)). Those simulations have led authors to conjecture that, as long as a market has a sufficiently high number of participants (i.e., as long as it is sufficiently thick),² and as long as the number of acceptable choices from each agent on at least one side of the market is bounded, then the proportion of agents in the market that can benefit by misreporting their preferences or vacancies should be small.

Some authors have proven this conjecture analytically: [Immorlica and Mahdian \(2005\)](#) for one-to-one matchings (i.e., for the marriage market) and then [Kojima](#)

²In this paper, a market being thick loosely means that it has a large number of participants. It should not be confused with the connotation that the term has in [Kojima and Pathak \(2009\)](#), where it is used to designate a formal technical assumption regarding the distribution of preferences.

and Pathak (2009) who extended Immorlica and Mahdian (2005) results to many-to-one matchings (i.e., to the college admission problem), though focusing on a very specific type of stable mechanism: the student optimal stable match mechanism (SOSM). However, as Kadam (n.d.) observed, in some very conservative environments, Kojima and Pathak (2009) results require an absurdly large number of participants in the market to guarantee that the mechanism is effectively strategy-proof, sometimes more than 10^{600} participants.³ In a survey, Pathak (2011) recognizes that it is still an open question as to whether a tighter upper bound to these convergence results can be found.

More recently, Storms (2013) has made a considerable improvement on Kojima and Pathak (2009) upper bound, by showing that one can focus on a very specific type of strategy in order to determine whether a college has incentives to misreport its preferences or vacancies. He also extended Kojima and Pathak (2009) results to all other stable matching mechanisms. Though Storm’s improvement on the rate of convergence is significant, it is still arguably slow, which can lead one to question whether those results have practical applications to existing markets.⁴

A common feature shared by all of the aforementioned papers is that they derive their main convergence results by applying the same logic: first they show

³As I write this in 2018, the current estimated number of atoms in the universe ranges between 10^{78} and 10^{82} .

⁴Apart from the papers cited above, more recently Azevedo and Budish (2019) provided a different approach to show that certain mechanisms, including deferred acceptance mechanisms, are effectively strategy-proof for large markets. Their convergence rate to the equilibrium, however, is even slower, as their result is proven for a more generic set of mechanisms, including auctions.

that the only reason why a college would want to misreport its preferences in the SOSM mechanism is because, by rejecting an acceptable student, that student can potentially make a proposal to her next best college, which in turn can accept the offer, and as a result, displace a less preferred student for which it was already tentatively matched with. The displaced student can then make an offer to her next best choice, thus perpetuating a chain of new offers and rejections that can potentially lead to a new proposal being made to the college that started this chain reaction. Those papers then proceed to show that for large markets such chain is likely to be disrupted due to a new offer being made to a college with open vacancies before a new offer is ever made to the college that triggered the chain of rejections. So their results rely on assumptions that guarantee that the market ends up with many unfilled seats after implementing the SOSM mechanism (one way to achieve this is by assuming that the market has more vacancies than students).

In this paper I modify this approach by showing instead that those chains of rejections are more likely to come to a halt as the result of a student in the chain having exhausted all of his acceptable choices. For this reason, the results presented in this paper do not rely on the market having disproportionately more vacancies than students. This is particularly important given that many matching markets are so competitive that virtually all vacancies end up being filled, such as the Brazilian centralized college admission market.⁵ The rate of convergence for truth telling that I get is also significantly faster than the ones obtained in the aforementioned

⁵Because public Brazilian universities are completely tuition free, there is a fierce competition for getting into those universities, resulting in virtually all vacancies being filled.

literature.

3.2 General framework for the college admission problem

This section describes the main environment and introduces some basic notation. Much of the notation and definitions (such as stability and SOSM) are standard in the Matching literature, and therefore can be skipped by someone already familiar with those concepts.

A market is comprised of two disjoint sets of participants: a set $C = \{c_1, c_2, \dots, c_m\}$ of m colleges and a set $S = \{s_1, s_2, \dots, s_n\}$ of n students. Each college c is assumed to have a limited number of vacancies given by $q_c \in \mathbb{N}$. Each student can either be matched with a single college or stay unmatched, and each college c can either be matched with a subset of students not exceeding its capacity q_c , or stay unmatched. Each student s has a list of strict rational preferences over C and \emptyset , denoted by \succ_s , and each college c has a list of strict rational preferences over the $2^{|S|}$ set of subsets of students plus the empty set \emptyset , denoted by \succ_c . If $\emptyset \succ_s c$, then student s is said to prefer being unmatched over being matched to college c , or that college c is unacceptable to student s . If $\emptyset \succ_c \tilde{S}$ then college c is said to prefer being unmatched over being matched with the subset of students \tilde{S} , or that the subset of students \tilde{S} is unacceptable to college c . Because a college c will only be allowed to be matched with at most q_c students, it is assumed without loss of generality that a college prefers to remain unmatched over being matched with any group of students exceeding its capacity, i.e., $\emptyset \succ_c \tilde{S}$ for all \tilde{S} such that $|\tilde{S}| > q_c$. A market is then a

tuple $(S, C, \succ_S, \succ_C, q)$, where $\succ_S = (\succ_s)_{s \in S}$, $\succ_C = (\succ_c)_{c \in C}$ and $q = (q_c)_{c \in C}$.

For a given preference relation \succ_c from college c , define R_c as the preference relation over individual students derived from \succ_c , i.e., R_c is a preference relation defined on $S \cup \emptyset$ such that $s R_c s'$ if and only if $\{s\} \succ_c \{s'\}$ for all $s, s' \in S \cup \emptyset$.

Definition 3.2.1 *College c 's preferences \succ_c are responsive if $s R_c s'$ implies that for any $\tilde{S} \subseteq S$ with $s \in \tilde{S}$ and $s' \notin \tilde{S}$ implies that $\tilde{S} \succ_c \tilde{S} \cup \{s'\} \setminus \{s\}$.*

Informally, colleges having responsive preferences simply means that they rank students regardless of the set of other students that they are already matched with. Throughout this paper all colleges are assumed to have responsive preferences. This assumption is commonly made in the matching literature for tractability reasons. In particular, when preferences exhibit complementarities, a stable matching allocation may not even exist (Roth and Stomayor (1992)).⁶

Throughout the analysis, only preferences over individual students, R_c , will be required, not \succ_c . For this reason, sometimes \succ_C will be replaced by R_C in a market's description, likewise (S, C, \succ_S, R_C, q) , where $R_C = (R_c)_{c \in C}$ corresponds to colleges' preferences over individual students.

Assume that each student only has $k \leq m$ acceptable choices, and that all students are acceptable to all colleges. While the hypothesis that all students are acceptable to all colleges is without loss of generality, the convergence results derived in this paper require that at least one side of the market has a limited number of acceptable choices. This assumption was also required in previous literature (namely,

⁶Though Storms (2013) is able to prove his results for a class of preferences that is a superset of responsive preferences, I abstract from such generalization in this paper.

in [Immorlica and Mahdian \(2005\)](#), [Kojima and Pathak \(2009\)](#) and [Storms \(2013\)](#)), the reason being that for stable matching mechanisms in which the number of agents on each side of the market (in the current context, the number of students vs the number of vacancies) are approximately equal to one another, and in which every agent prefers any match over staying unmatched (i.e., all matches are acceptable) then it can be shown that the market has a high proportion of individuals that can potentially benefit by misreporting their preferences, even in large markets ([Ashlagi, Kanoria and Leshno \(2015\)](#)).

Though the assumption that agents have a limited set of acceptable choices may seem restrictive, it can be justified by the empirical evidence that in most real world centralized matching markets agents not allowed to elicit more than a fixed number of acceptable choices. So the results derived in this paper provide a good approximation to matching mechanisms in which this constraint is not believed to be binding for most participants.

Definition 3.2.2 *In market $(S, C, \succ_S, \succ_C, q)$, a matching μ is a correspondence mapping $C \cup S$ into itself, possibly assuming empty values, such that:*

1. $\forall c \in C, \mu(c) \subseteq S$ (possibly, with $\mu(c) = \emptyset$), and $|\mu(c)| \leq q_c$.
2. $\forall s \in S, \mu(s) \subseteq C$ (possibly, with $\mu(s) = \emptyset$), and $|\mu(s)| \leq 1$.
3. $\forall c \in C$ and $s \in S, s \in \mu(c)$ if and only if $\mu(s) = c$.

In words, a matching is correspondence that maps students to colleges. Condition 1 from definition 3.2.2 means that a college c is either matched to a subset of

students which does not exceed its capacity q_c , or is unmatched, in which case it is said that college c is matched with itself (when $\mu(c) = \emptyset$). Condition 2 means that a student is either matched to a single college or is unmatched (when $\mu(s) = \emptyset$). And condition 3 means that a college c is matched with a subset of students that includes student s if and only if student s is matched with college c .

Let \succeq_j be the weak preference relation associated with \succ_j . An individual $j \in C \cup S$ is said to weakly prefer the matching μ over the matching μ' if and only if $\mu(j) \succeq_j \mu'(j)$, and with abuse of notation that is denoted by $\mu \succeq_j \mu'$.

For a given market $(S, C, \succ_S, \succ_C, q)$, a matching μ is said to be **individually rational** to agent $j \in C \cup S$ if $\mu(j) \succeq_j \emptyset$. For this same market, a college-student pair (c, s) is said to **block the matching allocation** μ if $c \succ_s \mu(s)$ and $s R_c s'$ for some $s' \in \mu(c)$.

Definition 3.2.3 *For a given market $(S, C, \succ_S, \succ_C, q)$, a matching μ is **stable** if:*

1. μ is individually rational to all agents in the market,
2. There is no college-student pair that blocks μ .

It can be shown that when preferences from colleges are responsive, μ is stable if and only if it is a core allocation, i.e., if and only if there is no coalition involving any number of participants that can block μ (Roth and Stomayor (1992)).

In the current environment with a finite number of agents, one can always find a stable matching by adopting the well known Gale and Shapley deferred acceptance algorithm with students proposing. The algorithm works as follows:

Algorithm 3 (*Student-proposing deferred acceptance algorithm*)

Round 1) In the first round, each student applies to her most preferred college. Then each college c accepts its preferred students among the ones making a proposal, and rejects all other students in excess of its capacity q_c . Students who are accepted by a college are tentatively matched with that college.

Round k) At every subsequent step, students who are not currently tentatively matched propose to their most preferred college among the ones they consider acceptable and have not rejected them yet, and each college c accepts its preferred students among the ones currently making a proposal and the ones they were tentatively matched with in the previous round. Students in excess of the college's capacity are rejected.

The algorithm continues until each student is either tentatively matched with an acceptable college or has already been rejected by all of their acceptable colleges. The algorithm must eventually stop, as no student proposes to the same college more than once, and because there is a finite number of students and colleges in the market. As stated earlier, the resulting allocation obtained through this algorithm is stable ([Roth and Stomayor \(1992\)](#)).

In a centralized matching mechanism, students and colleges report their preferences and number of vacancies to a matchmaker, and based on those reports the matchmaker determines the final matching outcome μ . Formally, in market (S, C, \succ_S, R_C, q) , a matching mechanism is defined as a function ψ that maps students' reported preferences over colleges \succ'_S , colleges' reported preferences over in-

individual students R'_C , and colleges' reported number of vacancies q' into a matching allocation $\psi(\succ'_S, R'_C, q')$. Notice, that according to this definition, colleges report their preferences only over individual students as opposed to reporting their complete preferences over all the possible subsets of students that they can get matched with. This is done because when colleges' preferences are responsive their preferences over individual students are enough to characterize whether a matching mechanism is stable or not (Roth and Stomayor (1992)).

Denote $\psi(\succ'_S, R'_C, q')(i)$ as the matching allocation obtained by agent $i \in C \cup S$ in mechanism ψ , when reported preferences and vacancies are $\{\succ'_S, R'_C, q'\}$. A matching mechanism is said to be stable if the resulting matching allocation is stable with respect to agents' reports.

Definition 3.2.4 *In market (S, C, \succ_S, R_C, q) , a matching mechanism ψ is stable if, given any report (\succ'_S, P'_C, q') , the resulting matching allocation $\psi(\succ'_S, R'_C, q')$ is stable in market $(S, C, \succ'_S, R'_C, q')$.*

One particular stable matching mechanism that has become increasingly popular in the school choice problem is the one that implements the Gale and Shapley algorithm with students proposing (i.e., the one in which agents elicit their preferences and vacancies to the matchmaker, who then implements the Gale and Shapley algorithm over those reports to determine the final allocation).

A desirable property of this mechanism is that it is stable with respect to agents' reported preferences and number of vacancies. As mentioned earlier, one of the reasons stability is desirable is because it implies that agents have incentives to

follow through the recommendation made by the matchmaker, assuming of course that they have reported their preferences and vacancies truthfully. Associated with stability is also the notion of fairness, as in a stable match a student indexed by s will not envy the allocation from another student who ended up in a college for which student s had a higher priority. For this reason, in the context of school choice, the stability property is sometimes referred to as *justified envy free* property (Abdulkadiroglu and Snmez (2003)).

Another desirable property of the Gale and Shapley mechanism is that, among all the stable matchings, this is the one that provides the highest utility to students, i.e., from the students' perspective, this matching mechanism weakly Pareto dominates all other stable matches (Roth and Stomayor (1992)). This is also the reason why the allocation obtained through this mechanism is referred in the literature as the student optimal stable match, or SOSM for short.

Finally, as mentioned in the introduction, it can be shown that this mechanism is *strategy-proof* for students, that is, students have a weakly dominant strategy of reporting their true preferences to the matchmaker (Roth and Stomayor (1992)). Therefore, in this mechanism students do not need to think strategically when eliciting their preferences, thus greatly simplifying their decision making process.

In this mechanism, however, some colleges may potentially have incentives to misreport their true preferences or to underreport their true number of vacancies. One way of eliminating colleges' incentives to misreport their preferences would be by adopting the college proposing version of the Gale and Shapley algorithm, which yields the most desirable stable match for colleges. However, in such an environment

it would then be the students the ones willing to misreport their preferences, and some colleges would still have incentives to underreport their vacancies (Sönmez (1997)). As it turns out, it is impossible to find a stable mechanism in which those incentives completely disappear, as explained in the next section.

3.3 Strategic Manipulation

An agent is said to be able to manipulate a matching mechanism ψ if, by reporting preferences or vacancies other than their true preferences or vacancies, the agent gets a better match, assuming that everyone else reports their preferences and vacancies truthfully. Formally, incentives to manipulate a matching mechanism can be stated as follows:

Definition 3.3.1 *For a given market $(S, C, \succ_S, \succ_C, q)$, a college c is said to be able to manipulate a matching mechanism ψ if there exists a pair of preference list and vacancy (R'_c, q'_c) such that*

$$\psi(\succ_S, (R'_c, R_{-c}), (q'_c, q_{-c}))(c) \succ_c \psi(\succ_S, R_C, q)(c).$$

Analogously, a student s is said to be able to manipulate a matching mechanism ψ if there exists a preference list \succ'_s such that

$$\psi((\succ'_s, \succ_{-s}), R_C, q)(s) \succ_s \psi(\succ_S, R_C, q)(s).$$

Even though students can not manipulate the SOSM mechanism, colleges may potentially have incentives to either misreport their true preferences or to underre-

port their number of vacancies.⁷ And for any stable match, there always exists a set of preferences such that one or more agents in one side of the market can manipulate the mechanism (Roth and Stomayor (1992)).

Notice that the form of strategic manipulation defined above does not include *prearranged matches*, the practice of forming a match before the centralized matching mechanism takes place. Section 3.4 provides a brief comment regarding this specific type of manipulation. Unfortunately, unless agents have common knowledge regarding the distribution of other agents' preferences and vacancies, market thickness may not be enough to eliminate the formation of those early matches.

The next section describes the main result from this paper, namely that though strategic manipulation in stable matches is a theoretical possibility, under some regularity conditions, a very small fraction of those agents can actually manipulate a stable matching mechanism, provided that the market is sufficiently large.

3.4 Large Markets

The evaluation of the frequency with which agents can manipulate a matching mechanism requires preferences to be random. So this section starts by describing

⁷Under the current environment where students' preferences remain unaltered throughout the entire matching process, it can be easily shown that colleges would never have strict incentives to overreport their capacities, i.e., of undertaking a strategy resembling *overbooking*. In practice, however, preferences may change throughout the matching process as a result of changes in home addresses, changes in school location, inadequate provision of entitled services at assigned schools, etc. But those frictions are not analyzed in this paper.

the data generating process (DGP) governing agents' preferences, and then it proceeds to show that, as the market becomes sufficiently large, the probability that a generic agent can manipulate a stable matching mechanism converges to zero as the number of participants in the market goes to infinity.

So starting with the description of the DGP governing agents' preferences, it is assumed that, associated with each college c_i there is a number $p_{c_i} \in (0, 1)$, which from now on shall be referred to as the popularity of college c_i , and associated with each student s_i there is a number $p_{s_i} \in (0, 1)$, the popularity of student s_i . Those popularities are assumed to add up to one, i.e., $\sum_{i=1}^m p_{c_i} = 1$ and $\sum_{i=1}^n p_{s_i} = 1$. Without loss of generality agents are ordered from highest to lowest popularity, i.e.,

$$p_{c_1} \geq p_{c_2} \geq \cdots \geq p_{c_m},$$

and

$$p_{s_1} \geq p_{s_2} \geq \cdots \geq p_{s_n}.$$

Given these popularities, the probability that the preference list from an arbitrary college c is $(s_{i_1}, s_{i_2}, \cdots s_{i_n})$ is given by

$$\prod_{j=1}^n \frac{p_{s_{i_j}}}{1 - \sum_{l=0}^{j-1} p_{s_{i_l}}},$$

and analogously, the probability that the preference list from an arbitrary student s is $(c_{i_1}, c_{i_2}, \cdots c_{i_k})$ is given by

$$\prod_{j=1}^k \frac{p_{c_{i_j}}}{1 - \sum_{l=0}^{j-1} p_{c_{i_l}}},$$

where $p_{s_{i_0}} = p_{c_{i_0}} = 0$. That is, ordered sampling without replacement is used to determine agents' preferences.

Notice that the greater the discrepancy between popularities, the more correlated the realized preferences become. For instance, if $p_{c_i} = (1 - \varepsilon)\varepsilon^{i-1}$ and ε is very small, then most students are likely to end up with the same preferences, whereas $p_{c_1} = p_{c_2} = \dots = p_{c_m} = 1/n$ would cause students' preferences to become completely uncorrelated with one another. Throughout this paper, no restrictions whatsoever are imposed on colleges' popularity, and therefore on the correlation level of students' preferences. But in terms of colleges' preferences, it is assumed that they are the least correlated as possible, that is:

$$p_{s_1} = p_{s_2} = \dots = p_{s_n} = \frac{1}{n}.$$

Though this assumption may seem restrictive, it has been noticed that in general more correlated preferences are associated with a lower number of agents in the market being able to successfully manipulate a stable matching mechanism. In particular, if all agents in one side of the market have the exact same preferences, then no agent in the market will have incentives to misreport their preferences or vacancies in any stable matching mechanism, which would make the convergence result derived in this paper trivial. Though [Ashlagi, Kanoria and Leshno \(2015\)](#) provide a counterexample in which more correlated preferences can actually increase the level of manipulation in the market, cases like those are relatively rare, as shown by his simulations and confirmed by our own simulations presented in section [B.4](#) in the appendix. Therefore, though a very specific DGP governing colleges' preferences is used, that is done entirely for tractability reasons (see the last paragraph from section [B.4](#) in the appendix), and the convergence of truth telling is expected to

hold at an even faster rate for different DGP's.

With that in mind, a stochastic market is defined as a tuple (S, C, P_C, q) , where $P_C = \{p_{c_1}, p_{c_2}, \dots, p_{c_m}\}$ specifies the popularity of each college in the market. Students' popularity are omitted from the tuple, as they are always assumed to be the same, i.e., $p_s = 1/|S|$ for all $s \in S$.

Given the above assumptions it can be shown that the expected proportion of colleges that can manipulate any stable matching mechanism in this sequence of random markets converges to zero as the number of colleges in the market goes to infinity. Moreover, by adapting a proof presented by [Storms \(2013\)](#) one can show that if in addition this sequence of random markets always has more students than colleges, then the proportion of students that can manipulate any stable match also converges to zero as the number of colleges goes to infinity.⁸ Those results are stated formally in theorems [3.4.1](#) and [3.4.2](#).

Theorem 3.4.1 *Let $(S^m, C^m, P_C^m, q^m)_{m \in \mathbb{N}}$ be a sequence of stochastic markets such that $|C^m| = m$ and $\max(\{q^m\}) \leq \bar{q}$ for every $m \in \mathbb{N}$, where $\bar{q} \in \mathbb{N}$. Also, denote $(\succ_S^m, (R^m, q^m))_{m \in \mathbb{N}}$ as the random sequence of preferences and vacancies associated with this sequence of stochastic markets. For a given stable matching mechanism ψ ,*

$$\alpha(m) = \mathbb{E} \left\{ \#c \in C^m; \psi(\succ_S^m, (\tilde{R}_c, R_{-c}^m), (\tilde{q}_c, q_{-c}^m)) \succ_c^m \psi(\succ_S^m, R_C^m, q^m) \right. \\ \left. \text{for some } (\tilde{R}_c, \tilde{q}_c) \text{ and some stable match } \psi \right\}$$

⁸In his proof, [Storms \(2013\)](#) forgets to mention that, in order to guarantee that the proportion of students that can manipulate a stable match converges to zero, the sequence of random markets must have at least as many students as colleges, which is usually the case in practical situations.

corresponds to the expected number of colleges that can manipulate the mechanism.

Given the above notation and assuming each student only has up to k acceptable choices,

$$U(m) = \frac{(k-1)}{m} + \frac{q}{m} \left[\frac{\left(1 - \left(1 - \frac{\bar{q}}{\bar{q}+1}\right)^{k-1}\right)}{\left(1 - \left(1 - \frac{\bar{q}}{\bar{q}+1}\right)^{k-1}\right) \frac{1}{m-k+1} + \left(1 - \frac{\bar{q}}{\bar{q}+1}\right)^{k-1}} \right]$$

is an upper bound to the proportion of colleges that can manipulate a stable matching mechanism, $\alpha(m)/m$, and it converges to zero as $m \rightarrow \infty$.

The intuition for the proof can be described as follows. First, a result derived by [Storms \(2013\)](#) (which is an improvement to one of the lemmas in [Kojima and Pathak \(2009\)](#)) is used to narrow down the search of possible strategic manipulations from colleges in the SOSM mechanism. Indeed, though in principle there are several different ways a college can misreport its preferences ($|S^m|!$ to be more precise), one can focus on a single type of misrepresentation, which consists on a college reporting all the students that it would ordinarily end up matched with in the SOSM as unacceptable, while reporting its true preferences over the remaining students. The intuition for why the analysis can be restricted to this very specific strategy is because the only reason why a college would want to misreport its preferences in the SOSM mechanism is to receive proposals from students who would ordinarily not do so under truthful reporting. And those offers can be obtained by rejecting acceptable students. Indeed, if a college rejects an acceptable student then that student will propose to his next best option, which can potentially displace another student who will then propose to his next best college, and so on, creating a chain of new proposals

that can potentially reach back the college that triggered this chain reaction. So intuitively, the more students a college rejects, the more likely the college will receive a new offer. But because the deferred acceptance algorithm does not backtrack, the only rejections that need to be considered are the ones that would not naturally arise in the SOSM under truthful reporting, which consists on the rejection of the students that a college ends up matched with under truthful reporting. Therefore a necessary condition for a college to have incentives to manipulate the SOSM is that the rejection of all the students it would ordinarily get matched with under truthful reporting generates at least one new offer for that college. This is formally proven in lemma [B.2.2](#) in the appendix.

Then, it is shown that the probability that this type of strategy actually results in a new offer being made to the college that triggers such chain of rejections is small as the market size increases. While the previous literature has proven this result by showing that for large markets such chain is likely to be disrupted due to a new offer being made to a college with open vacancies before a new offer is ever made to the college that triggered the chain of rejections; I shown instead that those chains of rejections are likely to stop as the result of a student having exhausted all of their acceptable choices. Proving the result through this method results in a faster convergence rate, and it does not rely on assumptions that would cause the market to have many unfilled seats.⁹

Because the SOSM is college pessimal, it can be shown that a necessary con-

⁹[Kojima and Pathak \(2009\)](#) and [Storms \(2013\)](#) impose the restriction that $|S^m| \leq \bar{q}m$ for all $m \in \mathbb{N}$, which is not required in our derivation.

dition for a college to have incentives to manipulate a stable matching mechanism is that it can manipulate the SOSM mechanism, which implies that the upper bound derived for colleges' incentives to manipulate the SOSM also applies to any stable match (lemma B.2.4).

Now moving to the results of the theorem, notice that the term inside the brackets of the upper bound $U(m)$ from theorem 3.4.1 always lies between zero and one. However, because $\bar{q} \geq 1$, this upper bound can potentially be greater than one for small sample sizes.

Figure 3.1 displays $U(m)$ when every agent has the same popularity and when each college has capacity $\bar{q} = 5$ and each student has $k = 5$ acceptable colleges. As illustrated in the figure, this function converges to zero much faster than previous upper bounds derived in the literature.¹⁰

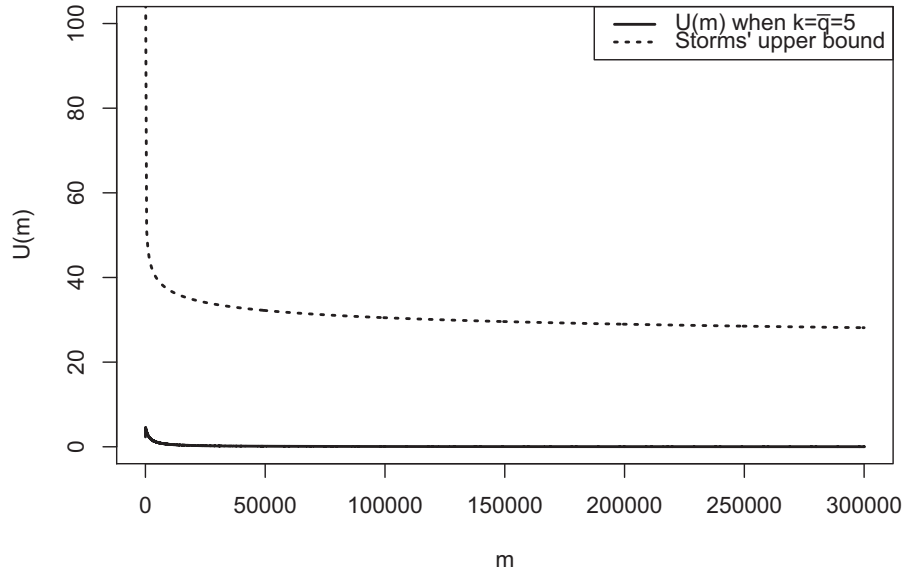


Figure 3.1: Upper bound $U(m)$ compared with the upper bound derived by Storms (2013), when $\bar{q} = k = 5$.

¹⁰The upper derived by Storms (2013) is given by $16\bar{q}k/(\ln(\bar{q}m)) + (\bar{q} + 1) \ln(\bar{q}m)/(2k\sqrt{\bar{q}m})$.

For lower values of k or \bar{q} , the upper bound naturally converges more quickly. In particular, for $\bar{q} = 1$, which corresponds to a one-to-one matching environment, $U(m)$ always lies between zero and one. But by inspecting the equation for $U(m)$, one can see that when k , the number of acceptable choices from each student, is high, $U(m)$ converges very slowly towards zero, even for low values of \bar{q} . This is illustrated in figure 3.2, which displays $U(m)$ computed for different values of k , where upper and lighter lines correspond to higher values of k . This pattern is consistent with the theoretical prediction that incentives to manipulate stable matching mechanisms are high when the market is balanced (i.e., when the number of students is approximately equal to the number of vacancies) and all matches are acceptable to both sides of the market.

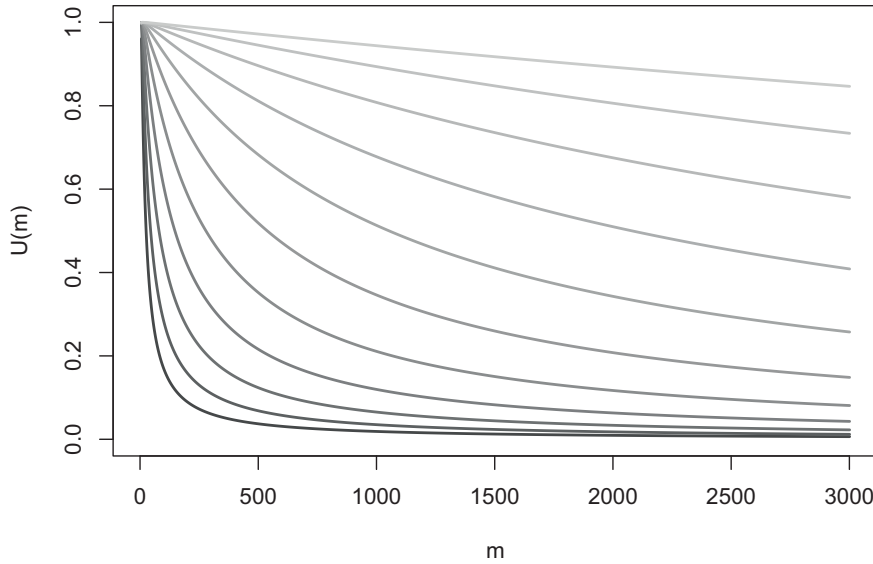


Figure 3.2: Upper bound $U(m)$ for $q = 1$ and different values of k ranging from 5 to 15, where the upper and lighter lines correspond to higher values of k .

Using the result from theorem 3.4.1, one can also show as a corollary that, provided that the market has more students than colleges (which is almost always

true in practical situations) students also do not have incentives to misreport their preferences in large stable matches. This result is formally stated in theorem 3.4.2.

Theorem 3.4.2 *Consider the sequence of markets $(S^m, C^m, P_C^m, q^m)_{m \in \mathbb{N}}$ described in theorem 3.4.1. For a given stable matching mechanism ψ ,*

$$\beta(m) = \mathbb{E} \left\{ \#s \in S^m; \psi((\succsim_s, \succ_{-s}^m), R_C^m, q^m) \succ_s^m \psi(\succ_s^m, R_C^m, q^m) \right.$$

for some \succsim_s and some stable match ψ \},

corresponds to the expected number of students that can manipulate the mechanism.

If $|S^m| \geq m$ for all $m \in \mathbb{N}$, then

$$\frac{\beta(m)}{|S^m|} \leq \bar{q} \frac{\alpha(m)}{m}. \quad (3.1)$$

Because $\frac{\alpha(m)}{m} \rightarrow 0$, as $m \rightarrow \infty$, this implies that the proportion of students that can manipulate a stable matching mechanism, $\frac{\beta(m)}{|S^m|}$, also converges to zero, as $m \rightarrow \infty$.

Kojima and Pathak (2009) proves a corollary stating that the incentives to form prearranged matches also goes to zero as the number of participants in the market increases. The intuition from such corollary can be explained as follows. When preferences are common knowledge, a college can only attract a student through a prearranged match if the student's ranking is strictly worse than all the other students the college would naturally get matched with in the SOSM under truthful reporting. So the only motivation for a college to pursue a prearranged match is to improve its match in the centralized matching mechanism. So a necessary condition for a college to be able to manipulate the mechanism through a prearranged match

is that, if it does so, it receives a new offer in the centralized mechanism. Because the probability that a college can receive a new offer by underreporting its vacancies by one unit is only slightly lower than the probability that a college can get a new offer in the centralized mechanism by forming a prearranged match, the convergence result derived earlier also applies to this type of manipulation.

The practicality of this result is questionable, however, as it requires agents to know in advance exactly what offers they would get in the centralized matching mechanism, when in practice the main reason why agents agree to form prearranged matches is because they are uncertain about the offers they might get (if any) in the centralized market.¹¹

3.5 Equilibrium Analysis

Our previous results show that the proportion of colleges and students that does not have incentives to misreport their preferences or vacancies in a stable matching mechanism converges to zero as the number of colleges in the market goes to infinity, assuming that all other agents report their preferences truthfully. Though this result already gives us an insight as to how agents will behave in practice (i.e.,

¹¹While incentives to manipulate a mechanism through reported preferences or vacancies tend to be the lowest in the presence of perfect information, when it comes to manipulation through prearranged matches, it is the lack of knowledge of what will transpire in the centralized mechanism that usually induces agents to seek an early match. [Sönmez \(1999\)](#) has shown that even in the presence of full information, it is impossible for a stable matching mechanism to be both stable and not manipulable through prearranged matches.

they will probably just report their true preferences and vacancies), it still allows for the theoretical possibility that a few agents may want to manipulate a stable matching mechanism, even in large markets. Moreover, if those few agents were to misreport their preferences, that could cause a chain reaction that would lead more agents to misreport their preferences or vacancies. In this section we show that, with the addition of a technical assumption (namely, assumption 3.5.2), it can be shown that, when the number of participants in the market is sufficiently large, all agents have incentives to report their preferences truthfully.

Because colleges' preferences over students are assumed to be responsive, there exists an additive utility function that represents these preferences. More precisely, there exists an utility function $u_c(\cdot)$ for college c defined over 2^S , the set of all possible subsets of S , such that, for any $S' \subseteq S$,

$$u_c(S') = \begin{cases} \sum_{s \in S'} u_c(s), & \text{if } |S'| \leq q_c \\ < 0, & \text{else} \end{cases},$$

where $u_c(s) > 0 \ \forall s \in S$, $u_c(s) = u_c(\{s\})$, and $u_c(s) > u_c(\tilde{s})$ if and only if $s R_c \tilde{s}$.

Analogously, we can define a utility function $u_s(\cdot)$ to each student s , such that $u_s(c) > u_s(c')$ if and only if $c \succ_s c'$, and $u_s(c) < 0$ if and only if college c is unacceptable to student s .

For our sequence of random markets, we assume that $u_c(\cdot)$ and $u_s(\cdot)$ are both bounded, i.e., there exists a $w \in \mathbb{R}$ such that $\sup u_c(s) < w$ and $\sup u_s(c) < w$, where each supremum is taken over colleges, students, and the size of the market, m .

Assumption 3.5.1 *Let $(S^m, C^m, P_C^m, q^m)_{m \in \mathbb{N}}$ be a sequence of stochastic markets such that $|C^m| = m$, and $\{u_c^m\}_{c=1}^m$ and $\{u_s^m\}_{s=1}^m$ are agents' realized utility functions. We assume that utility functions are bounded, i.e., there exists a $w \in \mathbb{R}$ such that $\sup u_c^m(s) < w$ and $\sup u_s^m(c) < w$ for all $s \in S^m$ and all $c \in C^m$ and all $m \in \mathbb{N}$.*

Assumption 3.5.1 is required to ensure that the potential gains from misreporting preferences or vacancies do not increase faster than the diminishing probability of being able to manipulate the mechanism as the market grows larger.

Assumption 3.5.2 *For the sequence of random markets $(S^m, C^m, P_C^m, q^m)_{m \in \mathbb{N}}$, suppose that:*

$$\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} \rightarrow 0, \quad \text{as } m \rightarrow \infty.$$

Assumption 3.5.2 simply limits how popular a college can be. This assumption ensures that even for the college that has the highest probability of being able to successfully manipulate a stable matching mechanism, the chances from it doing so converge to zero as the sample size increases. The reason this feature is needed is because, while in the previous section a small number agents were allowed to misreport their preferences as the sample size increased, for a truth telling equilibrium the low incentives to manipulate the matching mechanism has to apply to all agents.

Theorem 3.5.3 *If assumptions 3.5.1 and 3.5.2 are satisfied, then for any $\varepsilon > 0$ there exists an $m_0 \in \mathbb{N}$ such that truth-telling by every college and student is an ε -Nash equilibrium for any market in the sequence with more than m_0 colleges.*

So by adding the assumption that agents' utilities are bounded and by imposing restrictions on the maximum popularity from colleges, truth telling becomes

an ε -equilibrium for sufficiently large markets, i.e., all agents in the economy have small incentives to misreport their preferences when the market is sufficiently large.

3.6 Conclusion

This paper derives an upper bound for the probability that agents can successfully manipulate a stable matching mechanism. It also shows that such upper bound converges to zero as the market grows larger. In terms of contributions to the literature, this paper relaxes assumptions that would cause a market to end up with many unfilled seats after the implementation of the SOSM, so that the results presented here can be applied to competitive environments in which virtually all vacancies end up being filled. The speed of convergence for truth-telling is also significantly faster than the ones obtained in previous studies.

While the analysis in this paper is build under a perfect information environment, agents are expected to have even less incentives to manipulate a matching mechanism under imperfect information, given that under such scenario misreporting one's preferences or vacancies is a risky strategy that can potentially worsen one's final allocation. So the actual proportion of agents that would be willing to manipulate a stable matching mechanism is expected to be much lower in practical situations than the figures presented in this paper, which only correspond to an upper bound to those incentives.

Because the results suggest that the SOSM is virtually strategy proof for large markets, and because it satisfies a series of other desirable properties (it is

stable, student-optimal, etc.), this mechanism should be considered more often in practical situations. This is particularly important given that mechanisms like the Boston school choice mechanism, which not only is unstable but also gives students strong incentives to misreport their preferences even in large markets, are still widely popular.

The results from this paper also gives us insight as to which tie breaking rule to apply in situations where schools exhibit indifferences in their preferences. Indeed, though papers like [Kesten \(2012\)](#) advocate in favor of using a single tie breaking lottery in the SOSM in order to increase the correlation of schools' preferences over students, and thus reduce schools' incentives to misreport their true number of seats; our results suggest that a single tie breaking rule may be unnecessary in large stable matching markets, given that in such markets schools are very unlikely to be able to achieve a better outcome by underreporting their capacities, even when schools have completely uncorrelated rankings over students. Moreover, a single tie breaking lottery has the disadvantage of making parents from kids who get a low draw from the lottery to feel as though they have been treated unfairly, which can then lead them to sue the matchmaker. Such disputes could be avoided if each school held its own independent lottery.

Appendix A: Appendices for Chapter 1

A.0.1 Proofs

Proof of proposition 1.3.1: Let $C(X)$ be the set of real bounded continuous functions with the sup norm defined over $[0, \bar{\eta}]$. If $V(q, \cdot) \in C(X)$, then applying the following transformation T to $V(q, \cdot)$:

$$T(V(q, \mu)) \equiv \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv \quad (\text{A.1})$$

$$s.t. \quad \mu' = \frac{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}}}{\mu e^{-\frac{(v-1-\eta(1,\mu))^2}{2\sigma^2}} + (1-\mu)e^{-\frac{(v-\eta(0,\mu))^2}{2\sigma^2}}}, \quad (\text{A.2})$$

we have that $T(V(q, \cdot))$ also belongs to $C(X)$. Indeed, because $\tilde{\eta} \in [0, \bar{\eta}]$, the expression $\lambda\tilde{\eta}^2$ is bounded. In addition, because $q \in \{0, 1\}$, we have that $0 \leq \mu \leq 1$, so that $\omega(\mu') = (1 + \mu)^2/4$ is also bounded. And finally, by assumption, $V(q, \cdot)$ is bounded, which implies that $\delta V(q, \cdot)$ is also bounded. So if we aggregate all these terms to form the function $X(\mu, v, \tilde{\eta}) \equiv \omega(\mu') - \lambda\tilde{\eta}^2 + \delta V(q, \mu')$ defined over $[0, 1] \times \mathbb{R} \times [0, \bar{\eta}]$ (where μ' is obtained by constraint A.2), we have that X is bounded. Therefore, there exists $\underline{x}, \bar{x} \in \mathbb{R}$ such that $\underline{x} \leq X(\mu, y, \tilde{\eta}) \leq \bar{x}$ for any $(\mu, v, \eta) \in [0, 1] \times \mathbb{R} \times [0, \bar{\eta}]$. This implies that

$$T(V(q, \mu)) = \max_{\tilde{\eta}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} X(\mu, v, \tilde{\eta}) dv \in [\underline{x}, \bar{x}], \quad \forall \mu \in [0, 1],$$

so that $T(V(q, \cdot))$ is bounded.

The continuity of $T(V(q, \mu))$ follows from the fact that the function $f : [0, 1] \times [0, \bar{\eta}] \rightarrow \mathbb{R}$ such that

$$f(\mu, \tilde{\eta}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\mu^2 - \lambda\tilde{\eta}^2 + \delta V(q, \mu')] dv,$$

is continuous,¹ and the set of feasible choices for $\tilde{\eta}$, $[0, \bar{\eta}]$, is compact so that, from the maximum theorem,

$$T(V(q, \mu)) = \max_{\tilde{\eta} \in [0, \bar{\eta}]} f(\mu, \tilde{\eta})$$

is continuous with respect to μ .

Now the operator $T : C(X) \rightarrow C(X)$ clearly satisfies the Blackwell sufficient conditions for a β -contraction. Because $C(X)$ is a Banach space, the contraction

¹To show that $f(\mu, \tilde{\eta})$ is continuous, define

$$g(\mu, \tilde{\eta}, v) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} [\mu^2 - \lambda\tilde{\eta}^2 + \delta V(q, \mu')],$$

and let $(\mu_n, \tilde{\eta}_n)_{n=1}^{\infty}$ be a generic sequence defined on $[0, 1] \times [0, \bar{\eta}]$ such that $(\mu_n, \tilde{\eta}_n) \rightarrow (\mu, \tilde{\eta})$. Because $g(\cdot)$ is continuous (since it is the multiplication of continuous functions), the sequence of functions $h_n : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$h_n(v) = g(\mu_n, \tilde{\eta}_n, v), \quad \forall n \in \mathbb{N} \text{ and } \forall v \in \mathbb{R},$$

converges pointwise to $h(\cdot)$ such that

$$h(v) \equiv g(\mu, \tilde{\eta}, v) \quad \forall v \in \mathbb{R}.$$

Moreover, because $|h_n(v)| \leq l(v) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(v-q-\tilde{\eta})^2}{2\sigma^2}} \max\{\bar{x}, -\underline{x}\}$ for all $n \in \mathbb{N}$ and all $v \in \mathbb{R}$, and because $l(\cdot)$ is integrable, we have from Lebesgue's Dominated Convergence Theorem that

$$\lim_{n \rightarrow \infty} f(\mu_n, \tilde{\eta}_n, v) = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g(\mu_n, \tilde{\eta}_n, v) dv = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} h_n(v) dv = \int_{-\infty}^{\infty} h(v) dv = f(\mu, \tilde{\eta}, v).$$

mapping theorem guarantees that the operator $T(\cdot)$ has a unique fixed point in $C(X)$. ■

A.0.2 Jaccard similarity index

To compute the Jaccard similarity index, we first generate all sequences of 4 words from each review. We call those sequences as “shingles”. As an example, consider the following hypothetical review:

“These wireless earphones are the best!”

The shingles from the above sentence are:

1. “These wireless earphones are”
2. “wireless earphones are the”
3. **“earphones are the best”**

Now doing the same process with the following sentence:

“Those earphones are the best I ever had!”,

we get the shingles

- I “Those earphones are the”
- II **“earphones are the best”**
- III “are the best I”
- IV “best I ever had”

The Jaccard similarity between those two reviews is given by the number of shingles that intersect divided by the added number of shingles from each review. So in the current example, one can see that shingles 3 and II are the only ones that match. So the Jaccard similarity between those reviews is given by $2/(3+4) = 0.29$.

While computing the Jaccard similarity index is computationally feasible for a pair of small reviews, doing so for thousands of potentially large reviews is computationally infeasible.² Fortunately, computer scientists have devised clever hashing algorithms that allows one to consistently estimate the actual Jaccard index in a way that is computationally feasible. For the purposes of this research I used the widely popular MinHash algorithm. For more details about how the algorithm works, see [Wang et al. \(n.d.\)](#).

A.0.3 Variables

A.0.3.1 Naïve Bayes estimate of text reliability

As mentioned earlier, text similarity was build by using a Naïve Bayes classifier. At a high level, the process consists on computing the frequency from each word that appears among fake and real reviews, and then using those frequencies to estimate the probability that a certain sequence of words was generated from a legitimate or a fraudulent review. The process can be employed using content from both review text and review title.

More precisely, let $text = (w_1, w_2, \dots, w_n)$ represent a generic sequence of

²For one of my samples, I would need to make 4.17+10 of those computations.

words used to review a product. Then it follows from Bayes' rule that:

$$P(\text{fake}|\text{text}) = \frac{P(\text{text}|\text{fake})P(\text{fake})}{P(\text{text})},$$

and

$$P(\text{real}|\text{text}) = \frac{P(\text{text}|\text{real})P(\text{real})}{P(\text{text})},$$

where the notation is self explanatory.

So conditional on its content, a review is more likely to be fake iff

$$P(\text{fake}|\text{text}) > P(\text{real}|\text{text})$$

$$\iff P(\text{text}|\text{fake})P(\text{fake}) > P(\text{text}|\text{real})P(\text{real})$$

$$\iff \log(P(\text{text}|\text{fake})) + \log(P(\text{fake})) > \log(P(\text{text}|\text{real})) + \log(P(\text{real})). \quad (\text{A.3})$$

Getting an unbiased and consistent estimate of $P(\text{fake})$ is relatively easy: one only needs to compute the fraction of reviews in the sample that are fake (though in practice one actually uses the fraction of reviews in the sample that are *classified* as fake, as it is virtually impossible to perfectly distinguish fake reviews from real ones). But unless one is willing to make restrictions regarding the data generating process (DGP) from review texts, one can not hope to obtain an unbiased and consistent estimates of $P(\text{text}|\text{fake})$ and $P(\text{text}|\text{real})$.

The Naïve Bayes classifier approach simplifies the DGP from review texts by assuming that words are generated randomly and independently. Though this assumption is not very realistic since words need to be put in a logical order in order to convey meaning, it greatly simplifies the process of finding a reliable estimate of $P(\text{text}|\text{fake})$. Indeed, letting $\text{text} = (w_1, w_2, \dots, w_n)$ denote the sequence of words

from a review, this assumption implies that

$$P(\text{text}|\text{fake}) = \prod_{i=1}^n P(w_i|\text{fake}).$$

Because the probabilities $P(w_i|\text{fake})$ can be consistently estimated by computing the proportion of times each word w_i appears on the set of words used to write fake reviews, one can consistently estimate $P(\text{text}|\text{fake})$ by multiplying those estimated probabilities.³ The same approach can be applied to estimate $P(\text{text}|\text{real})$.

So the aforementioned procedure was used to estimate the left and righthand side of inequality A.3. If the estimated $P(\text{real}|\text{text})$ was greater than $P(\text{fake}|\text{text})$, then the dummy variable “*Reliability index from review text*” would assume value 1, else it would assume value 0. The same procedure was used to compute “*Reliability index from review tile*”, but using the contents from the review title as opposed to the review text.

A.0.3.2 Detecting anomalous peaks on the volume of 5 star reviews

Detecting spikes on the number of 5 star reviews received by a seller was done using an STL (seasonal trend decomposition) approach. The process consists on first estimating the expected number of positive reviews that a seller should receive at a particular day as a function of trend, seasonal effects and covariates. If the estimated prediction was sufficiently distant from the realization of positive reviews on that period, a dummy would classify all the 5 star reviews that the seller received

³As a standard approach, *stop words*, such as “I”, “there”, “but”, etc., were removed from the reviews before conducting the Naïve Bayes estimation.

on that day as anomalous.

More precisely, reviews were aggregated on a daily level to create a panel data. Let $X_{i,t,p,s}$ be the number of 5 stars that a product p from seller s received at date t , during its i 'th period since it entered the market (notice that t is the actual date it received a review, whereas i corresponds to the number of days since that product got its first review). $X_{i,t,p,s}$ was regressed against its lagged components, trend, seasonal dummies, and seller fixed effect, likewise:

$$X_{i,t,p,s} = \beta_0 X_{i-1,t-1,p,s} + \beta_1 t + \sum_{j=1}^7 \gamma_j D_{j,t} + \alpha_s + \varepsilon_{i,t,p,s},$$

where $\{D_{j,t}\}_{j=1}^7$ are the dummies for the corresponding days of week, α_s is the seller's fixed effect, and $\varepsilon_{i,t,p,s}$ is an iid random term.

After estimating the model using OLS, it was determined that if a residual term was 4 standard deviations above or below the average residual, then that day for the corresponding seller would be flagged as anomalous, in which case all the 5 star reviews that the seller received on that day would be flagged as anomalous.

Appendix B: Appendices for Chapter 3

B.1 Dropping strategies are exhaustive

A college is said to adopt a dropping strategy if it reports a subset of students as unacceptable while truthfully reporting its order of preferences over the remaining students. Formally, dropping strategies are defined as follows:

Definition B.1.1 *For a given market (S, C, \succ_S, R_C, q) , a preference and vacancy report (R'_c, q'_c) from college c is said to be a dropping strategy if*

1. $\emptyset R'_c s$ implies that $\emptyset R'_c s$ for all $s \in S$.
2. There is a set $D \subseteq S$ such that $s R'_c \emptyset$ and $\emptyset R'_c s$ for all $s \in D$, and $s R'_c s'$ if and only if $s R'_c s'$ for all $s \in S \setminus D$.
3. $q'_c = q_c$.

As stated in the main text, [Storms \(2013\)](#) has shown that we can focus on this specific type of manipulation to analyze colleges' incentives to game the SOSM mechanism. The proof is replicated below.

Lemma B.1.1 *(Dropping strategies are exhaustive) Consider a market $(S, C, \succ_S, \succ_C, q)$ and a stable matching mechanism for ψ for that market. Then, for any*

pair of preference and vacancy report $(\tilde{R}_c, \tilde{q}_c)$ from college c , there exists a dropping strategy (R'_c, q_c) such that

$$\psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c})) \succeq_c \psi(\succ_S, (R'_c, R_{-c}), q).$$

Proof: Consider the preference and vacancy report $(\tilde{R}_c, \tilde{q}_c)$ from college c , and define

$$\mu \equiv \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c})).$$

Let (R'_c, q_c) be the dropping strategy that defines all students in the set $D \equiv \{s \in S; s \notin \mu(c) \text{ or } \emptyset R_c s\}$ as unacceptable, but preserves the same preference ordering as R_c for those students not in D . Then clearly,

$$\psi(\succ_S, (R'_c, R_{-c}), q)(c) \subseteq \{s \in \mu(c); s R'_c \emptyset\}.$$

We want to show that

$$\psi(\succ_S, (R'_c, R_{-c}), q)(c) = \{s \in \mu(c); s R'_c \emptyset\},$$

because, since the set $\{s \in \mu(c); s R'_c \emptyset\}$ equals to $\mu(c)$ excluding all those students in $\mu(c)$ that college c finds unacceptable according to R_c , we must have that

$$\{s \in \mu(c); s R'_c \emptyset\} \succeq_c \mu(c).$$

To show that $\psi(\succ_S, (R'_c, R_{-c}), q)(c) = \{s \in \mu(c); s R'_c \emptyset\}$, consider the match μ_1 such that:

$$\mu_1(c') = \begin{cases} \mu(c'), & \text{if } c' \neq c, \\ \{s \in \mu(c); s R'_c \emptyset\}, & \text{if } c' = c. \end{cases}$$

For college c we have that

$$sR'_c\emptyset \quad \text{for all } s \in \mu_1(c). \quad (\text{B.1})$$

For all other colleges $c' \neq c$, we have $\mu_1(c') = \mu(c)$, which, since μ is stable and thus individually rational in market $(S, C, \succ_S, (\tilde{R}_c, R_{-c}), \tilde{q}_c, q_{-c})$, implies that

$$\mu_1(c') \succeq_{c'} \emptyset, \quad \forall c' \neq c. \quad (\text{B.2})$$

For every student $s \in S$, either $\mu_1(s) = \mu(s)$ or $\mu_1(s) = \emptyset$. In either case, because μ is individually rational in market $(S, C, \succ_S, (\tilde{R}_c, R_{-c}), \tilde{q}_c, q_{-c})$, we must have

$$\mu_1(s) \succeq_s \emptyset. \quad (\text{B.3})$$

Together, conditions (B.1) to (B.3) imply that μ_1 is individually rational in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$.

Now clearly, because $\{s \in \mu(c); sR'_c\emptyset\}$ is the set of all acceptable students from college c according to the preference list R'_c , and because $\mu_1(c) = \{s \in \mu(c); sR'_c\emptyset\}$ we have that college c can not be part of college-student pair that blocks μ_1 in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$. Adding this to the fact that $\mu_1(c) \subseteq \mu(c)$ and $\mu_1(c') = \mu(c') \quad \forall c' \neq c$ and the fact that μ is a stable allocation in market $(S, C, \succ_S, (\tilde{R}_c, R_{-c}), \tilde{q}_c, q_{-c})$, we have that the only student-college pairs that may potentially block μ_1 in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$, are the ones involving a students s such that $\mu(s) = c$ and $\mu_1(s) = \emptyset$ and a college $c' \neq c$.

So lets iteratively generate a stable matching allocation by adopting the following algorithm:

Step 1) Select an arbitrary student s' that can be paired with some college to block μ_1 in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$. Let college c' be student s' most preferred college among all the colleges it can use to form such a blocking pair. Define μ_2 as the matching such that $\mu_2(\tilde{c}) = \mu_1(\tilde{c})$ if $\tilde{c} \neq c'$, and

$$\mu_2(c') = \begin{cases} \mu_1(c') \cup s' & \text{if } |\mu_1(c')| < q_c, \\ \mu_1(c') \cup s' \setminus \{\tilde{s}; s'' R_c \tilde{s} \quad \forall s'' \in \mu_0(c) \quad \text{s.t. } s'' \neq \tilde{s}\}, & \text{else.} \end{cases}$$

Clearly, in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$, μ_2 is individually rational, and only students unmatched under μ_2 can potentially form a pair with some college to block μ_2 . If there is no such student, define $\tilde{\mu} = \mu_2$ and end the algorithm, else, proceed to step 2.

Step k) Select an arbitrary student s' that can be paired with some college to block μ_k in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$. Let college c' be student s' most preferred college among all the colleges it can use to form such a blocking pair. Define μ_{k+1} as the matching such that $\mu_{k+1}(\tilde{c}) = \mu_k(\tilde{c})$ if $\tilde{c} \neq c'$, and

$$\mu_{k+1}(c') = \begin{cases} \mu_k(c') \cup s' & \text{if } |\mu_k(c')| < q_c, \\ \mu_k(c') \cup s' \setminus \{\tilde{s}; s'' R_c \tilde{s} \quad \forall s'' \in \mu_0(c) \quad \text{s.t. } s'' \neq \tilde{s}\}, & \text{else.} \end{cases}$$

Clearly, in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$, μ_{k+1} is individually rational, and only students unmatched under μ_{k+1} can potentially form a pair with some college to block μ_{k+1} . If there is no such student, define $\tilde{\mu} = \mu_{k+1}$ and end the algorithm, else, proceed to step $k + 1$.

Because at each step of the algorithm, colleges forming the blocking pair are always made strictly better off as compared to the previous match, while all other

colleges get the same match as the previous one, we have that the algorithm must eventually stop after a finite number of steps, since each college can only improve a finite number of times in a market with a finite number of students.

By construction, $\tilde{\mu}$ is stable in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$, and since we have that at every step of the algorithm college c can never be part of a blocking pair, we also have that $\tilde{\mu}(c) = \mu_1(c) = \{s \in \mu(c); sR'_c\emptyset\}$.

Because the number of vacancies filled by each college is always the same in any stable match (Roth and Stomayor (1992)), this implies that

$$|\tilde{\mu}(c)| = |\psi(\succ_S, (R'_c, R_{-c}), q)(c)|. \quad (\text{B.4})$$

Since $\psi(\succ_S, (R'_c, R_{-c}), q)(c) \subseteq \{s \in \mu(c); sR'_c\emptyset\} = \tilde{\mu}(c)$, (B.4) implies that

$$\psi(\succ_S, (R'_c, R_{-c}), q)(c) = \{s \in \mu(c); sR'_c\emptyset\},$$

as we wanted to show. ■

B.2 Rejection Chains

Let μ_S denote the student optimal stable match in market (S, C, \succ_S, R_C, q) . In order to simplify notation, we assume every student is acceptable to every college in this market. Then we define a rejection chain for the subset of students $B_c^0 \subseteq \mu_S(c)$ matched with college c as follows:

Algorithm 4 (*Rejection Chains*)

1. Initialize the algorithm with $i = 0$, $H_{c'} = \mu_S(c') \forall c' \in C$, and define $A_{s'}$ as the set of acceptable colleges student $s' \in S$ did not propose to under the SOSM.

2. If $B_c^i = \emptyset$, terminate the algorithm. Else, let $s \in B_c^i$ be such that $s'R_c s \forall s' \in B_c^i \setminus \{s\}$ (i.e., let s be college c 's least preferred student among B_c^i). Then, set $B_c^{i+1} = B_c^i \setminus \{s\}$ and proceed to step 3.
3. (a) If $A_s = \emptyset$, set $i = i + 1$ and go back to phase 2, else:
 - (b) Let c' be student s most preferred college among A_s . Then student s applies to that college and we set $A_s = A_s \setminus \{c'\}$ and $H_{c'} = H_{c'} \cup \{s\}$. If $c' = c$, terminate the algorithm, else:
 - (c) If $sR_{c'} s'$ for some $s' \in H_{c'}$ and $|H_{c'}| \geq q_{c'}$, set $s = s'$, and restart step 3. Else:
 - (d) If $sR_{c'} s'$ for some $s' \in H_{c'}$ and $|H_{c'}| < q_{c'}$, set $i = i + 1$ and return to step 2. Else:
 - (e) If $s'R_{c'} s \forall s' \in H_{c'}$, go back to the beginning of step 3.

Lemma B.2.1 *In market (S, C, \succ_S, R_C, q) , let $X_{c'}(R'_c)$ be the number of students who make a proposal to college c' in the student optimal deferred acceptance algorithm, when college c adopts the dropping strategy (R'_c, q_c) and all other agents report their preferences and vacancies truthfully. Then, if the set of students reported as unacceptable under the dropping strategy (R'_c, q_c) is contained in the set of students reported as unacceptable under the dropping strategy (\tilde{R}_c, q_c) , we must have*

$$X_{c'}(R'_c) \subseteq X_{c'}(\tilde{R}_c).$$

In words, lemma B.2.1 simply states that, the more students a college drops from its preference list, the more offers the college potentially gets under the SOSM

mechanism. The intuition for the proof is that, the higher the number of students that a college c lists as unacceptable, the more likely the college is to reject students, who on their turn will potentially make offers to other colleges, triggering a rejection chain that may potentially generate a new offer to college c . For a detailed proof, see [Storms \(2013\)](#).

Lemma B.2.2 *In market (S, C, \succ_S, R_C, q) , a college c has incentives to manipulate the SOSM mechanism ψ_S only if algorithm 4 with $B_c^0 = \mu_S(c)$ ends with college c receiving an offer (i.e., at the end of step 3b).*

Proof: Consider an arbitrary dropping strategy (R'_c, q_c) from college c . Suppose by contrapositive that algorithm 4 with $B_c^0 = \mu_S(c)$ does not end with a new offer being made to college c (i.e., it ends at step 2). Then, we want to show that

$$\mu_S(c) \succeq_c \psi_S(\succ_S, (R'_c, R_{-c}), q)(c). \quad (\text{B.5})$$

Clearly, a sufficient condition for B.5 to hold is that

$$X_c(R'_c) \subseteq X_c(R_c),$$

i.e., a college won't have incentives to manipulate the SOSM mechanism if the set of offers college c gets under any dropping strategy is always contained in the set of offers college c gets when it reports its preferences and number of vacancies truthfully in the student optimal deferred acceptance algorithm.

Let (\tilde{R}_c, q_c) be the dropping strategy from college c that selects all students in $X_c(R_c)$ as unacceptable, while keeping college c 's original preference list for the

remaining students. Then, because algorithm 4 with $B_c^0 = \mu_S(c)$ does not end with a new offer being made to college c , we clearly have that $X_c(\tilde{R}_c) = X_c(R_c)$.¹

Now let (R_c^*, q_c) be the dropping strategy that selects all students in the matching mechanism as unacceptable. Because students in $S \setminus X_c(\tilde{R}_c)$ never propose to college c after performing the student-proposing deferred acceptance algorithm under preferences and vacancies $(\succ_S, (\tilde{R}_c, R_{-c}), q)$, we have that if, in addition to rejecting students in $X_c(\tilde{R}_c)$, college c also selects students in $S \setminus X_c(\tilde{R}_c)$ as unacceptable, then that should not affect the set of students that propose to college c , which implies that $X_c(\tilde{R}_c) = X_c(R_c^*)$.

Therefore, for any dropping strategy (R'_c, q_c) from college c , we have from lemma B.2.1 that

$$X_c(R'_c) \subseteq X_c(R_c^*) = X_c(\tilde{R}_c) = X_c(R_c),$$

as we wanted to show. ■

Lemma B.2.3 *For a given market (S, C, \succ_S, R_C, q) , consider a stable matching mechanism ψ and a strategy report $(\tilde{R}_c, \tilde{q}_c)$ from college c . Then, letting ψ_S denote the SOSM mechanism, we have that there exists a dropping strategy (R'_c, q_c) from*

¹This is due to the fact that, at the beginning of algorithm 4, all students in $X_c(R_c)$ but not in $\mu_S(c)$ have already been rejected by college c in the student-proposing deferred acceptance algorithm, and then, at the end of algorithm 4, all students in $\mu_S(c)$ are rejected by college c . Therefore, because at the end of algorithm 4, college c has rejected all students in $X_c(R_c)$, and because no new offer is made to college c , and because the order of offers and rejections in the Gale and Shapley algorithm is irrelevant, we have that, if college c adopts the dropping strategy (\tilde{R}_c, q_c) which selects all students in $X_c(R_c)$ as unacceptable and preserves the preference list for all other students, it won't receive any new offer, so that $X_c(\tilde{R}_c) = X_c(R_c)$.

college c such that:

$$\psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c) = \psi_S(\succ_S, (R'_c, R_{-c}), q)(c).$$

Proof: Consider the strategy (R'_c, q_c) from college c that selects all students not in $\psi(\succ_S, (\tilde{R}_c, R_{-c}), (q_c, q_{-c}))(c)$ as unacceptable, while keeping the same preference ordering over the remaining students as in \tilde{R}_c . Then,

$$\psi_S(\succ_S, (R'_c, R_{-c}), q)(c) \subseteq \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c). \quad (\text{B.6})$$

Clearly, because the matching $\psi_S(\succ_S, (R'_c, R_{-c}), q)$ is stable in market $(S, C, \succ_S, (R'_c, R_{-c}), q)$, we have that the matching μ such that

$$\mu(c') = \begin{cases} \psi_S(\succ_S, (R'_c, R_{-c}), q)(c'), & \text{if } c' \neq c \\ \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c), & \text{if } c' = c \end{cases}$$

is also stable in this same market.²

Because the number of vacancies filled by a college is always the same in every stable match, we have that

$$|\psi_S(\succ_S, (R'_c, R_{-c}), q)(c)| = |\mu(c)|,$$

which, by (B.6), implies that

$$\psi_S(\succ_S, (R'_c, R_{-c}), q)(c) = \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c),$$

as we wanted to show. ■

²To see this, notice that the matching μ equals to the matching $\psi_S(\succ_S, (R'_c, R_{-c}), q)$, with the only potential difference being that college c may be better off by being matched to additional acceptable students (according to preferences R'_c) who, on their turn, prefer to be matched to college c over staying unmatched, as in $\psi_S(\succ_S, (R'_c, R_{-c}), q)$.

Lemma B.2.4 *If a college does not have incentives to manipulate the SOSM mechanism, then it does not have incentives to manipulate any stable matching mechanism.*

Proof: Let ψ be an arbitrary stable matching mechanism defined in market $(S, C, \succ_S, \succ_C, q)$, and let $(\tilde{R}_c, \tilde{q}_c)$ be an arbitrary preference and vacancy report from college c . Then, if ψ_S is the SOSM mechanism, we have, from lemma B.2.3 that there exists a dropping strategy (R'_c, q_c) such that

$$\psi_S(\succ_S, (R'_c, R_{-c}), q)(c) = \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c).$$

Assuming college c can not manipulate the SOSM mechanism implies that

$$\psi_S(\succ_S, R_C, q)(c) \succeq_c \psi_S(\succ_S, (R'_c, R_{-c}), q)(c) = \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c). \quad (\text{B.7})$$

But because the SOSM is the worst stable match for colleges (Roth and Stomayor (1992)), we must have

$$\psi(\succ_S, R_C, q)(c) \succeq_c \psi_S(\succ_S, R_C, q)(c). \quad (\text{B.8})$$

Together, (B.7) and (B.8) imply that

$$\psi(\succ_S, R_C, q)(c) \succeq_c \psi(\succ_S, (\tilde{R}_c, R_{-c}), (\tilde{q}_c, q_{-c}))(c).$$

Because the choice of $(\tilde{R}_c, \tilde{q}_c)$ was arbitrary, we have that college c can not manipulate the matching mechanism ψ . ■

B.3 Stochastic student-proposing DA algorithm and stochastic rejection chains

Consider a stochastic market (S, C, P_S, P_C, q) . It can be easily shown that the data generating process (DGP) for the preference list from colleges described in section 3.4, in which all students have the same popularity, is equivalent to the following DGP (in the sense that it generates the same distribution of preference lists):

step 1) For a given college c , and a given differentiable cumulative density function

$F(\cdot)$ select an arbitrary student $s_c^1 \in S$ and draw a random number x_c^1 from distribution F . That number will represent college c 's utility over being matched with student s_c^1 . Defining $H_c^1 = \{s_c^1\}$, we clearly have that, since H_c^1 has a unique element, student s_c^1 must be college c 's favorite student from H_c^1 .

step 2) Next, select some arbitrary student $s_c^2 \in S \setminus H_c^1$ and draw a random number

x_c^2 from the same distribution F , which will represent college c 's utility over being matched with student s_c^2 . Defining $H_c^2 = H_c^1 \cup s_c^2$ we have that s_c^2 will be college c 's most preferred student among H_c^2 iff $x_c^2 > x_c^1$.

\vdots

step k) Select some arbitrary student $s_c^k \in S \setminus H_c^{k-1}$ and draw a random number x_c^k

from the same distribution F , which will represent college c 's utility over being matched with student s_c^k . Defining $H_c^k = H_c^{k-1} \cup s_c^k$ we have that s_c^k will be

college c 's l th preferred student among H_c^k if it provides the l th highest utility to college c among H_c^k .

Continue this process until college c has formed a preference list over all of the students in the market.

It can be shown that the order in which students are added to the preference list does not affect the overall distribution of preferences. Defining the DGP in this manner is useful because it allows us to easily compute the probability that a college accepts a new offer at some step of the student-proposing deferred acceptance algorithm, or at some step of algorithm 4, conditional on the set of offers that the college has received so far.

With that in mind, we have that the following algorithm is stochastically equivalent to drawing entire preference lists for students and colleges according to the data generating process described in section 3.4, and then implementing the student-proposing deferred acceptance algorithm on the realized preferences.

Algorithm 5 (*Stochastic student-proposing DA algorithm*)

1. *Initialization:* Let $l = 1$. For every $s \in S$, let $A_s = \emptyset$, $H_c = \emptyset$. Throughout the algorithm we will denote \tilde{R}_c as the preference list from college c among the students in H_c .

2. *Choosing the applicant:*

- a) If $l \leq |S|$, then let s be the l th student (student s_l) and increment l by one.

b) Else, terminate the algorithm.

3. Choosing the applied:

a) If $|A_s| \geq k$, return to step 2.

b) Else, select c randomly from distribution P_C until $c \notin A_s$, and then define

$$A_s = A_s \cup \{c\}.$$

4. Acceptance or rejection: Draw an utility from a fixed distribution $F(\cdot)$, to represent the college c 's utility from being matched with student s . If the utility is lower than the utility from college c 's q_c most preferred students among the ones that have applied to that college so far, the student is rejected, and we return to step 3. Else, the student becomes tentatively matched with college c ; and if the college was already tentatively matched with q_c students (i.e., operating in full capacity), the college rejects its least preferred student among the students that it was tentatively matched with, and we redefine s as the rejected student and we go back to step 3, else we go back to step 2.

This process goes on until each student is either tentatively matched with a college, or has applied to k colleges.

The resulting match obtained after performing this algorithm, μ_S , is such that $\mu_S(c)$ equals to the most preferred students from college c among the ones that have applied to that college throughout the algorithm, up to college c 's capacity, q_c .

Algorithm 6 (*Stochastic rejection chain*)

Let μ_S be the match obtained after performing the stochastic student-proposing DA algorithm. Define A_s as the set of colleges that student s has applied to throughout the student-proposing DA algorithm, and let $H_{\tilde{c}}$ be the set of students who applied to college \tilde{c} throughout the same algorithm. We start the stochastic rejection chain with each college \tilde{c} tentatively matched with $\mu_S(\tilde{c})$ (i.e., each college \tilde{c} starts tentatively matched with its $q_{\tilde{c}}$ most preferred students among $H_{\tilde{c}}$).

We define a rejection chain for the subset of students $B_c^0 \subseteq \mu_S(c)$ matched with college c as follows:

1. Initialize the algorithm with $i = 0$.
2. If $B_c^i = \emptyset$, terminate the algorithm. Else, let s be college c 's least preferred student among B_c^i . Then, set $B_c^{i+1} = B_c^i \setminus \{s\}$ and proceed to step 3.
3. [a)]

If $|A_s| \geq k$, set $i = i + 1$ and go back to step 2, else:

- (b) Select $\tilde{c} \in C$ randomly from distribution P_C until $\tilde{c} \notin A_s$. Then student s applies to college \tilde{c} and we set $A_s = A_s \cup \{\tilde{c}\}$. If $\tilde{c} = c$, terminate the algorithm, else:
- (c) Redefine $H_{\tilde{c}} = H_{\tilde{c}} \cup \tilde{c}$. Draw an utility from a fixed distribution $F(\cdot)$, to represent the college \tilde{c} 's utility from being matched with student s . If the utility is lower than the utility from college \tilde{c} 's $q_{\tilde{c}}$ most preferred students among $H_{\tilde{c}}$, the student is rejected, and we return to step 3. Else, the student becomes tentatively matched with college \tilde{c} ; and if the college

was already tentatively matched with $q_{\bar{c}}$ students (i.e., operating in full capacity), the college rejects its least preferred student among the students that it was tentatively matched with, and we redefine s as the rejected student and we go back to step 3, else we go back to step 2.

Proof of Theorem 3.4.1: Assume all colleges c_j with $j \leq k - 1$ (i.e., the $k - 1$ most popular colleges) can manipulate a stable match with probability 1. For the remaining colleges c_j such that $j > k - 1$, we will find an upper bound to the probability that algorithm 6 with $B_{c_j}^0 = \mu_S(c_j)$ ends with a new offer being made to college c_j .

Clearly, if at some point of the algorithm a student makes an offer to a college with open vacancies, the chain of new offers immediately stops, and as a result college c_j does not receive a new offer. So we can get an upper bound to the probability that a new offer is made to the college triggering the rejection chain by assuming that, throughout all the steps of algorithm 6, a student making an offer always does so to a college with no open vacancies. This conservative assumption simplifies the computation of our upper bound, and will thus be used throughout the rest of the analysis.

The more likely a college with all vacancies filled accepts a new student in algorithm 6, the more likely the chain of new offers will continue to generate a new offer to college c_j .³ And the less the number of applications a college has received

³This intuitive fact can be verified analytically by replacing $1 - \frac{1}{q_j+1}$ into expression (B.10) by a generic probability, and then conducting comparative statics analysis on U_{c_j} through the usage of the implicit function theorem, to show that U_{c_j} is increasing in this probability.

so far, the more likely it will accept a new one, which is why we assume, throughout the algorithm, that whenever a college is faced with a new offer, it has received a number of offers exactly equal to its capacity: no more and no less. No more because that would increase the probability that a new offer is rejected, and no less because that would cause the algorithm to stop immediately and without a new offer ever being made to the college triggering the rejection chain.

With these assumptions, let $X_{i,c}$ denote the random utility drawn from the cdf $F(\cdot)$, which represents the utility that college c gets from being matched with student i . Let Y_c denote the lowest utility that college c gets from the students it is already tentatively matched with. Then, a college will accept a new offer if and only if $X_{i,c} > Y_c$. Because $X_{i,c}$ has cdf $F(\cdot)$, the expected probability that college c accepts the new offer from student i is given by

$$\mathbb{E}_{Y_c, X_{i,c}} [\text{Prob}(X_{i,c} \geq Y_c)] = \mathbb{E}_{Y_c} [1 - F(Y_c)]. \quad (\text{B.9})$$

To simplify notation, assume college c is currently matched with students $\{s_1, s_2, \dots, s_{q_j}\}$ (and recall that we are assuming that these are the only students who have made an offer to that college so far). Then, the cdf of Y_c is given by

$$\begin{aligned} G(y) &= \text{Prob}(\min\{X_1, \dots, X_{q_j}\} \leq y) \\ &= 1 - \text{Prob}(X_1 \geq y, \dots, X_{q_j} \geq y) \\ &= 1 - (1 - F(y))^{q_j}, \end{aligned}$$

which implies its pdf is given by

$$g(y) = q_j(1 - F(y))^{q_j-1} f(y).$$

Therefore, from expression B.9, the expected probability that a college accepts a new offer at each step of algorithm 6 is given by

$$\begin{aligned}
\mathbb{E}_{Y_c, X_{i,c}} [Prob(X \geq Y_c)] &= 1 - \int_{-\infty}^{+\infty} F(x)g(x) dx \\
&= 1 - \left[F(x)G(x)|_{-\infty}^{\infty} - \int_{-\infty}^{+\infty} f(x)G(x) dx \right] \\
&= 1 - \left[1 + \int_{-\infty}^{+\infty} f(x)[1 - (1 - F(k))^{q_j}] dx \right] \\
&= 1 - \int_{-\infty}^{+\infty} f(x)(1 - F(x))^{q_j} dx \\
&= 1 - \frac{1}{q_j + 1}.
\end{aligned}$$

A student making a new offer does so to college c_j with highest probability when the student has already offered to the top $k - 1$ colleges, excluding college c_j . Therefore, an upper bound to the probability the new offer is made to college c_j , with $j \geq k$ is given by

$$\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}.$$

Clearly the more likely students offer to college c_j at each step of algorithm 6, the more likely the rejection chain will result in a new offer being made to college c_j .⁴ Therefore, we will assume that, throughout the rejection chains, each student currently making an offer does so to college c_j with probability $\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}$.

Therefore, we can get an upper bound U_{c_j} to the probability that a college c_j with $j \geq k$ gets a new offer after rejecting one acceptable student, by solving the

⁴This intuitive fact can be verified analytically by replacing $\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}$ into expression (B.10) by a generic probability, and then conducting comparative statics analysis on U_{c_j} , by using the implicit function theorem, to show that U_{c_j} is increasing in this probability.

following expression:

$$\begin{aligned}
U_{c_j} = & \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m} + \frac{1}{q_j + 1} U_{c_j} \right] \\
& + \frac{1}{q_j + 1} \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m} + \frac{1}{q_j + 1} U_{c_j} \right] \\
& + \left(\frac{1}{q_j + 1}\right)^2 \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m} + \frac{1}{q_j + 1} U_{c_j} \right] \\
& \vdots \\
& + \left(\frac{1}{q_j + 1}\right)^{k-2} \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m} + \frac{1}{q_j + 1} U_{c_j} \right], \tag{B.10}
\end{aligned}$$

where $1 - \frac{1}{q_j + 1}$ corresponds to the probability that a matched college accepts a new offer, and $\frac{1}{q_j + 1}$ is the probability that the college rejects a new offer. If at some point of the rejection chain a college rejects a proposal from a student, that student can still apply to its remaining acceptable colleges. Because each student has only k acceptable choices, and because each student making an offer at some step of the algorithm has already been rejected by at least one college, each student in the rejection chain can not make more than $k - 1$ offers, which is why this summation has only $k - 1$ elements. The term $\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}$ is the probability that the current offer is made to college c_j . Whenever a new offer is not made to college c_j , which happens with probability $1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}$, the rejection chain goes on, which is why we multiply the term $1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}$ by the upper bound U_c .

Expression [B.10](#) can be rearranged as

$$U_{c_j} = \sum_{i=0}^{k-2} \left(\frac{1}{q_j + 1}\right)^i \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m} + \left(1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}\right) U_{c_j} \right],$$

which implies that

$$U_{c_j} = \frac{\left(1 - \left(\frac{1}{q_j+1}\right)^{k-1}\right) \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}}{1 - \left(1 - \left(\frac{1}{q_j+1}\right)^{k-1}\right) \left(1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}\right)}.$$

Because college c_j can reject at most q_j students, we have that, for any $j \geq k$, an upper bound to the probability college c_j can manipulate a stable match is given by $q_j U_{c_j}$. Because $q_j \leq \bar{q}$ and because this upper bound is linearly increasing in q_j , we have that an upper bound to the probability that college q_j obtains a new offer after rejecting all its matched students is given by

$$\bar{q} U_{c_j} = \bar{q} \frac{\left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}}{1 - \left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \left(1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}\right)}.$$

Adding all these upper bounds, including the upper bound of 1 for the $k-1$ most popular colleges, and then dividing the sum by m , we get the following upper bound to the expected proportion of colleges that can manipulate a stable matching mechanism:

$$U(m) = \frac{(k-1)}{m} + \bar{q} \times \frac{m-k+1}{m} \times \frac{\left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}}{1 - \left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \left(1 - \frac{p_{c_j}^m}{\sum_{l \geq k} p_{c_l}^m}\right)}. \quad (\text{B.11})$$

Now recall that, by assumption, $p_{c_j} \geq p_{c_{j+1}} \forall j < m$. Suppose that that $p_{c_j} > p_{c_{j+1}}$ for a certain $j \geq k$. What would happen to the upper bound we just computed if we transferred some popularity from college c_j to college c_{j+1} ? More precisely, what would happen to the upper bound $U(m)$ from equation (B.11) if the popularity from college c_j was changed to $p_{c_j} - \varepsilon$ and the popularity from college c_{j+1} was changed to $p_{c_{j+1}} + \varepsilon$, such that $p_{c_j} - \varepsilon > p_{c_{j+1}} + \varepsilon$, keeping the popularities from all

other students and colleges unaltered? Defining $A \equiv \left(\frac{1}{\bar{q}+1}\right)^{k-1}$, and $B \equiv \bar{q} \times \frac{m-k+1}{m}$,

we have that the variation in the upper bound would be given by

$$\Delta U(m)(\varepsilon) \equiv B \left[\frac{(1-A) \frac{p_{c_j} - \varepsilon}{\sum_{l \geq k} p_{c_l}^m}}{1 - (1-A) \left(1 - \frac{p_{c_j} - \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right)} + \frac{(1-A) \frac{p_{c_{j+1}} + \varepsilon}{\sum_{l \geq k} p_{c_l}^m}}{1 - (1-A) \left(1 - \frac{p_{c_{j+1}} + \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right)} \right].$$

Deriving this expression with respect to ε , we get

$$\frac{\partial \Delta U(m)(\varepsilon)}{\partial \varepsilon} = B \left[\frac{-A(1-A) \frac{1}{\sum_{l \geq k} p_{c_l}^m}}{\left[1 - (1-A) \left(1 - \frac{p_{c_j} - \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right)\right]^2} + \frac{A(1-A) \frac{1}{\sum_{l \geq k} p_{c_l}^m}}{\left[1 - (1-A) \left(1 - \frac{p_{c_{j+1}} + \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right)\right]^2} \right],$$

which is greater than zero if and only if

$$\begin{aligned} 1 - (1-A) \left(1 - \frac{p_{c_j} - \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right) &> 1 - (1-A) \left(1 - \frac{p_{c_{j+1}} + \varepsilon}{\sum_{l \geq k} p_{c_l}^m}\right) \\ \iff p_{c_j} - \varepsilon &> p_{c_{j+1}} + \varepsilon, \end{aligned}$$

which is satisfied by assumption.

Therefore, the upper bound $U(m)$ from equation (B.11) increases the more popularity is transferred from college c_j to college c_{j+1} , so that the upper bound is highest when $p_{c_j} = p_{c_{j+1}}$. Because this is true for any j such that $k \leq j < m$, we have that the upper bound achieves a maximum when $p_{c_j} = p_{c_i} \forall i, j \geq k$, which would entail

$$\frac{p_{c_j}}{\sum_{l \geq k} p_{c_l}^m} = \frac{1}{m-k+1} \quad \forall j \geq k.$$

So the following expression provides a tighter upper bound to the probability that a generic college can manipulate a SOSM

$$U(m) = \frac{(k-1)}{m} + \bar{q} \times \frac{m-k+1}{m} \times \frac{\left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \frac{1}{m-k+1}}{1 - \left(1 - \left(\frac{1}{\bar{q}+1}\right)^{k-1}\right) \left(1 - \frac{1}{m-k+1}\right)}.$$

From lemma B.2.4, a necessary condition for a college not to have incentives to manipulate a stable matching mechanism is that it has no incentives to manipulate the SOSM, which implies that $U(m)$ is also an upper bound to the expected proportion of colleges that can manipulate any stable matching mechanism.

The rest of the proof, namely, that this expression converges to zero, is trivial.

■

Proof of Theorem 3.4.2: From lemma B.2.3 and from the fact that the student optimal stable match is college pessimal, a necessary (but not sufficient) condition for a college not to have incentives to manipulate any stable matching mechanism is that it must be matched with the same set of students in every stable match. Therefore, from theorem 3.4.1, the expected number of colleges that are matched with different students at different stable matches is less than or equal to $\alpha(m)$. Because each college has at most \bar{q} vacancies, the maximum expected proportion of students who get a different match at different stable matches is $\bar{q} \left(\frac{\alpha(m)}{|S^m|} \right)$. Because a sufficient (and necessary) condition for a student not to have incentives to manipulate any stable matching mechanism is that it gets the same allocation in every stable match, we have that the proportion of students who can manipulate a stable matching mechanism, $\frac{\beta(m)}{|S^m|}$, is less than or equal to $\bar{q} \left(\frac{\alpha(m)}{|S^m|} \right)$. If $|S^m| \geq m$ $\forall m \in \mathbb{N}$, we have that

$$\frac{\beta(m)}{|S^m|} \leq \bar{q} \left(\frac{\alpha(m)}{|S^m|} \right) \leq \bar{q} \left(\frac{\alpha(m)}{m} \right) \rightarrow 0, \quad \text{as } m \rightarrow \infty,$$

which implies that $\frac{\beta(m)}{|S^m|} \rightarrow 0$, as $m \rightarrow \infty$. ■

B.4 Simulated Incentives to misreport preferences or vacancies under different DGP

On section 3.4 we make the assumption that all students have the same popularity, i.e., students are equally likely from being considered as top choices from schools. Though this might seem like a very restrictive assumption, we show through simulations that, for cases in which schools' preferences are more correlated they on average have less incentives to fake reviews.

Using the DGP suggested by [Ashlagi, Kanoria and Leshno \(2015\)](#), assume that each agent i has two characteristics, x_i^A and x_i^D . The utility of agent i being matched with agent j is given by:

$$u_i(j) = \beta x_j^A - \gamma(x_i^D - x_j^D)^2 + \varepsilon_{i,j},$$

where x_j^A , x_j^B and $\varepsilon_{i,j}$ are uniformly and independently distributed between $[0, 1]$. γ and β are parameters that capture the level of correlation between agents' preferences, with $\gamma = \beta = 0$ being the case in which agents' preferences are completely uncorrelated, which corresponds to the case in which every agent has the same popularity.

Figure B.1 plots the simulated proportion of cases in which a college can potentially have incentives to misreport its vacancies or capacity in a stable matching mechanism, as a function of parameters β and γ . More precisely, it plots the proportion of colleges that obtain at least one new offer after rejecting all of the students it would ordinarily get matched with in the SOSM with truth telling. As it can be

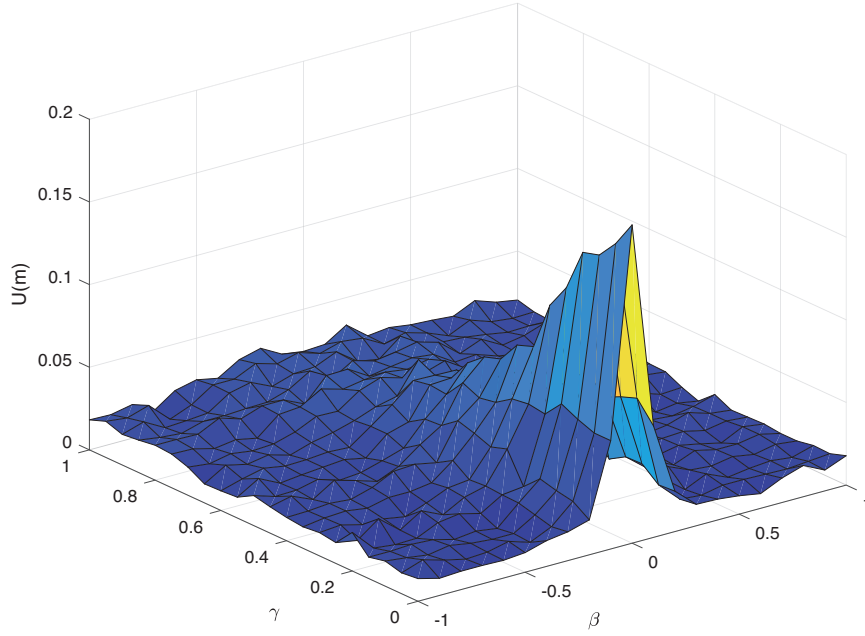


Figure B.1: Simulated proportion of colleges that had potential incentives to misreport preferences or vacancies under the SOSM mechanism, for different combinations of γ and β . The simulations were done assuming there were 45 colleges each offering 5 vacancies, and 225 students, so that the number of students equals to the overall number of vacancies. Each student was assumed to have $k = 5$ acceptable choices. seen from the plot, incentives to misreport one's preferences are highest in the case in which preferences are perfectly uncorrelated, i.e., when $\beta = \gamma = 0$.

It is important to emphasize that theorem 3.4.1 requires students' popularities to be uniform only for tractability purposes. Indeed, the theorem uses the most conservative scenarios to compute the probability that a generic college can receive a new offer after rejecting a student. The more popular a student is, the more likely a college will receive a new offer after rejecting that student. So if we assume that a generic college finds itself in the most conservative scenario, namely, that it is

matched with the most popular students, then a rejection will generate a new offer for that college with high probability, thus inflating the upper bound. However, in reality there can only be so many very popular students, i.e., it is impossible for two or more different colleges to be each simultaneously matched with the top \bar{q} most popular students, as each student can only be matched with a single college at a time. But because it is virtually impossible to take expectations as to who ends up matched with whom in the SOSM (even for relatively small markets, the number of possible different matches can greatly exceed the estimated number of atoms in the universe), the best we can do is make the conservative assumption that every college finds itself matched with the top \bar{q} most popular students, even though in reality that is not possible. But this (unrealistic) assumption then inflates the upper bound, compromising its convergence. So the assumption that every student has the same popularity is used to avoid such technicality.

B.5 Equilibrium Analysis

Lemma B.5.1 *If assumption 3.5.2 is satisfied, then the probability that any college can manipulate a stable matching mechanism goes to zero, as $m \rightarrow \infty$.*

Proof: An upper bound to the probability that a student who has been rejected at some step of the stochastic rejection chain makes a new offer to the college that triggered the rejection chain is given by:

$$\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m}.$$

Therefore, by a similar argument as the one used to prove theorem 3.4.1, an upper

bound U^m to the probability that any generic college gets a new offer after rejecting one of the students it ends up matched with under the SOSM mechanism is given by

$$\begin{aligned}
U^m = & \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} + \frac{1}{q_j + 1} U^m \right] \\
& + \frac{1}{q_j + 1} \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} + \frac{1}{q_j + 1} U^m \right] \\
& + \left(\frac{1}{q_j + 1}\right)^2 \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} + \frac{1}{q_j + 1} U^m \right] \\
& \vdots \\
& + \left(\frac{1}{q_j + 1}\right)^{k-2} \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} + \frac{1}{q_j + 1} U^m \right],
\end{aligned} \tag{B.12}$$

We can rearrange expression B.12 as

$$U^m = \sum_{i=0}^{k-2} \left(\frac{1}{q_j + 1}\right)^i \left(1 - \frac{1}{q_j + 1}\right) \left[\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} + \left(1 - \frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m}\right) U^m \right],$$

which implies that

$$U^m = \frac{\left(1 - \left(\frac{1}{q_j + 1}\right)^{k-1}\right) \frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m}}{1 - \left(1 - \left(\frac{1}{q_j + 1}\right)^{k-1}\right) \left(1 - \frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m}\right)}.$$

Because each college can reject at most \bar{q} students after the execution of the SOSM mechanism, the upper bound to the probability that a college can manipulate any stable matching mechanism is given by $\bar{q}U^m$. From assumption 3.5.2, we have that

$$\frac{p_{c_1}^m}{p_{c_1}^m + \sum_{i>k} p_{c_i}^m} \rightarrow 0,$$

as $m \rightarrow \infty$, which implies that $\bar{q}U^m \rightarrow 0$ as $m \rightarrow \infty$. ■

Lemma B.5.2 *If assumption 3.5.2 is satisfied, then the probability that any student can manipulate a stable matching mechanism goes to zero, as $m \rightarrow \infty$.*

Proof: Notice that the upper bound computed in the proof of theorem B.5.1 to the probability that a college can manipulate a stable matching mechanism did not depend on the particular preferences of the students it rejected after performing the SOSM. Therefore, conditional on having a certain preference, the probability that a student is matched to the same college in every stable match goes to one, as m goes to infinity. Because being matched to the same college is a necessary and sufficient condition for a student not to have incentives to misreport its preferences, we have that the probability that each student can manipulate a SOSM goes to zero as m goes to infinity. ■

Proof of Theorem 3.5.3: From Lemmas B.5.1 and B.5.2, the probability that students and colleges can successfully manipulate a stable matching mechanism converge to zero as the number of participants in the market goes to infinity. Because agents' utilities are assumed to be bounded (assumption 3.5.1), this implies that for a sufficiently large sample the market has an ε equilibrium in which all agents report their preferences and vacancies truthfully. ■

Bibliography

- Abdulkadiroglu, Atila, and Tayfun Snmez.** 2003. "School Choice: A mechanism Design Approach." *American Economic Review*, 93: 729–747.
- Akerlof, George A.** 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *The Quarterly Journal of Economics*, 84: 488–500.
- Ashlagi, Itai, Yash Kanoria, and Jacob D. Leshno.** 2015. "Unbalanced Random Matching Markets: The Stark Effect of Competition."
- Azevedo, Eduardo M, and Eric Budish.** 2019. "Strategy-proofness in the Large." *The Review of Economic Studies*, 86: 81–116.
- Azzimonti, Marina, and Marcos Fernandes.** 2018. "Social Media Networks, Fake News, and Polarization."
- Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch.** 1992. "A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades." *Journal of Political Economy*, 100.
- Campbell, Arthur, Dina Mayzlin, and Jiwoong Shin.** 2017. "Managing Buzz." *Journal of Economics*, 48.
- Dellarocas, Chrysanthos.** 2006. "Strategic Manipulation of Internet Opinion Forums: Implications for consumers and firms." *Management Science*, 52: 1577–1593.
- Esteban, Joan-Maria, and Debraj Ray.** 1994. "On the Measurement of Polarization." *Econometrica*, 62.
- Gentzkow, Matthew, and Jesse M. Shapiro.** 2006. "Media Bias and Reputation." *Journal of Political Economy*, 114.
- Immorlica, Nicole, and Mohammad Mahdian.** 2005. "Marriage, Honesty, and Stability." In Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms.

- Jabr, Wael, and Zhiqiang Zheng.** 2014. “Know Yourself and Know Your Enemy: An Analysis of Firm Recommendations and Consumer Reviews in a Competitive Environment.” *MIS Quarterly*, 38.
- Jindal, Nitin, and Bing Liu.** n.d.. “Opinion Spam and Analysis.”
- Kadam, Sangram Vilasrao.** “On the Large Market Core Convergence Result for Two-Sided Matching Markets.” 2014.
- Kaghazgaran, Parisa, James Caverlee, and Majid Alfifi.** n.d.. “Behavioral Analysis of Review Fraud: Linking Malicious Crowdsourcing to Amazon and Beyond.”
- Kesten, Onur.** 2012. “On two kinds of manipulation for school choice problems.” *Economic Theory*, 51.
- Kojima, Fuhito, and Parag A Pathak.** 2009. “Incentives and Stability in Large Two-Sided Matching Markets.” *American Economic Review*, 99: 608–627.
- Luca, Michael, and Georgios Zervas.** 2016. “Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud.” *Management Science*, 62: iv?vii, 3393–3672.
- Mayzlin, Dina.** 2006. “Promotional Chat on the Internet.” *Marketing Science*, 25: 155–163.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier.** 2014. “Promotional Reviews: An Empirical Investigation of Online Review Manipulation.” *American Economic Review*, 104: 2421–2455.
- Mukherjee, Arjun, Bing Liu, and Natalie Glance.** n.d.. “Spotting Fake Reviewer Groups in Consumer Reviews.”
- Mullainathan, Sendhil, and Andrei Shleifer.** 2005. “The Market for News.” *The American Economic Review*, 95.
- Nimark, Kristofer P, and Savitar Sudaresan.** 2018. “Inattention and belief polarization.”
- Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T Hancock.** n.d.. “Finding Deceptive Opinion Spam by Any Stretch of the Imagination.”
- Papanastasiou, Yiangos.** 2017. “Fake News Propagation and Detection: A Sequential Model.”
- Pathak, Parag A.** 2011. “The Mechanism Design Approach to Student Assignment.” *Annual Review of Economics*, 3: 513–536.

- Roth, Alvin E, and Elliott Peranson.** 1999. "The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design." *American Economic Review*, 89: 748–780.
- Roth, Alvin E, and Marilda Stomayor.** 1992. *Two-sided matching*. Elsevier Science Publishers.
- Shannon, Claude E.** 1948. "A mathematical theory of communication." *Bell System Technical Journal*, 27.
- Shukla, Aishwarya, Guodong Gao, and Ritu Agarwal.** 2018. "How Digital Word-of-Mouth Affects Consumer Decision Making: Evidence from Doctor Appointment Booking."
- Sönmez, Tayfun.** 1997. "Manipulation via Capacities in Two-Sided Matching Markets." *Journal of Economic Theory*, 77: 197–204.
- Sönmez, Tayfun.** 1999. "Can Pre-arranged Matches be Avoided in Two-Sided Matching Markets?" *Journal of Economic Theory*, 86: 148–156.
- Storms, Evan.** 2013. "Incentives and Manipulation in Large Market Matching with Substitutes." Master's diss. Stanford Department of Economics.
- Tennekoon, Vidhura, and Robert Rosenman.** 2016. "Systematically Misclassified Binary Dependent Variables." *Communications in Statistics - Theory and Methods*, 45.
- Tirole, Jean.** 1988. *Industrial Organization*. MIT Press.
- Wang, Jingdong, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen.** n.d.. "A Survey on Learning to Hash."
- Xiang, Yi, and Miklos Sarvary.** 2007. "News Consumption and Media Bias." *Marketing Science*, 26.
- Yildirim, Pinar, Esther Gal-Or, and Tansev Geylani.** 2013. "User-Generated Content and Bias in News Media." *Management Science*, 59.