IMPERIAL COLLEGE LONDON
IMPERIAL COLLEGE BUSINESS SCHOOL

ESSAYS IN FINANCIAL ECONOMICS

Antoine Uettwiller

April 2023

A thesis submitted in partial fulfillment of the requirements of the degree of *Doctor of Philosophy* at Imperial College London.

Declaration

I herewith certify that this thesis constitutes my own work and that all material, which is not my own work, has been properly acknowledged. I confirm that chapter 2 was jointly written with Santosh Anagol, Vimal Balasubramaniam and Tarun Ramadorai, and I contributed 25% of this work. I confirm that chapter 3 was jointly written with Tarun Ramadorai and Ansgar Walther, and I contributed 33% of this work.

Antoine Uettwiller

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial-No Derivatives 4.0 International Licence (CC BY-NC-ND).

Under this licence, you may copy and redistribute the material in any medium or format on the condition that; you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisors, Tarun Ramadorai and Christopher Hansman, for their invaluable guidance, support, and encouragement throughout my doctoral studies. Their mentorship and insights have been instrumental in shaping my research and the skills I have acquired.

Next, I am grateful to my other fantastic co-authors, Vimal Balasubramaniam, Santosh Anagol, and Ansgar Walther, for without their collaboration, two of the chapters in this thesis would not exist as such. I certainly look forward to working on additional projects with them in the future.

I would also like to extend my sincere thanks to the faculty members at Imperial College Business School whose conversations I have greatly benefited from. Special thanks go to Savitar Sundaresan, Marcin Kacperczyk, Franklin Allen, and Rajkamal Iyer for their time, feedback, and support.

I have been fortunate to pursue this Ph.D. alongside individuals who were always willing to offer assistance and ready to bounce ideas off. In particular, I would like to thank Mobeen Iqbal, Nicolas Rapanos, Massimo Dello Preite, Moritz Wiedemann, and Dheeya Rizmie for their invaluable support.

Finally, I owe a debt of gratitude to those that pushed me to pursue this Ph.D.; Edward Morrison, William Wilhelm, and, last but not least, my family.

Abstract

The three chapters in this thesis study agent behavior, whether it be that of retail investors (chapter 1), the reporting behavior of home-buyers for the purpose of taxation (chapter 2), or that of firms choosing what data to collect on their consumers (chapter 3). All three chapters employ data-driven approaches and novel methods to explore heterogeneity within their respective fields of study and draw conclusions that have implications for policymakers, market participants, and other stakeholders.

In the first chapter, I study the heterogeneity of retail investors based on the stocks they chose to discuss. I use Reddit WallStreetBets data totalling more than 1.5 million posts and 35 million comments from 2018 to the end of 2021. I find retail investors to be diverse, with distinct subgroups showing interest in stocks with attention-grabbing features, or fundamental-based investing. I also find those attributes to be connected to echo chamber effects, likely causing confirmation bias, finding that subgroups do not mix. Finally, I find subgroups to be of differential importance in a predictability exercise. Effects commonly attributed to retail investors, such as increased volatility, might be coming from a sub-sample of them but also a sub-sample of stocks.

In the second chapter, we develop a new method to estimate under-reporting and employ it on large, granular administrative data from the Mumbai real-estate market. The approach compares bunching of reported values around government-assessed guidance values with the distribution of a third-party measure of market values. We estimate that 11 percent of Mumbai real-estate value is under-reported between 2013 and 2022. We detect a decline in bunching post-demonetization, but no change in the overall under-reporting rate. Properties with mortgages from banks with high non-performing assets and from public-sector banks exhibit greater under-reporting.

In the third chapter, we scrape a comprehensive set of US firms' privacy policies and study these policies alongside firms' web data extraction behaviour. We find considerable and systematic variation in privacy policies along multiple dimensions including ease of access, length, readability, and clarity, both within and between industries. Surprisingly, firms' data extraction, measured by the extent of cookies they place on consumers browsing their websites, is strongly and positively related to the length and complexity of their policies. Firms with intermediate levels of technical sophistication have longer, more complex policies than firms with very high or low technical sophistication, controlling for a range of other firm attributes. A simple signalling model of firms engaging in data extraction in an economy with both myopic and sophisticated consumers, where privacy is a "shrouded attribute" of firm-consumer interactions, helps to rationalize these findings.

Contents

1 Retail Investor Heterogeneity: Evidence from WallStreetBets **17**

1.1 Introduction . . . 17

1.2 Data . . . 23

1.2.1 Reddit . . . 23

1.2.2 CRSP, Compustat, IBES, & SEC . . . 27

1.2.3 TAQ . . . 28

1.2.4 Summary Statistics . . . 28

1.2.5 Sentiment . . . 30

1.3 Characterising Individuals . . . 31

1.3.1 Stock Talking Matrix . . . 31

1.3.2 Stock Characteristic Matrix . . . 32

1.3.3 Characteristic Talking Matrix . . . 33

1.3.4 Characteristic Co-Talking Matrix . . . 34

1.4 Unobserved Multi-factor Model . . . 36

1.4.1 Principal Components Analysis (PCA) . . . 36

1.4.2 Cross Sectional Heterogeneity . . . 38

1.5 Clientele Effects . . . 40

1.5.1 Regression Model . . . 40

1.5.2 Results & Echo Chambers . . . 41

1.6 Returns, Volatility & Autocorrelation . . . 42

1.6.1 Short Horizon Returns . . . 44

1.6.2 Volatility . . . 45

1.6.3 Longer Horizon Returns . . . 46
1.6.4 Momentum and Reversal . . . 47
1.7 Conclusion . . . 48

2 A Bad Bunch: Asset Value Under-Reporting in the Mumbai Real Estate Market 70

2.1 Introduction . . . 70
2.1.1 Related Literature . . . 78
2.2 Institutional Background . . . 82
2.3 Economic Framework . . . 85
2.3.1 Estimating Under-reporting from Bunching . . . 88
2.3.2 Household Leverage, Cash Payment, and Preferences vs. Beliefs 90
2.4 Data . . . 91
2.4.1 Registrar Data . . . 91
2.4.2 Propequity Data . . . 93
2.5 Baseline Results . . . 96
2.5.1 Under-reporting Estimation in the Presence of Measurement Error in the Market Value Proxy . . . 97
2.5.2 Formal Elasticity Estimation . . . 108
2.5.3 Model Based Estimates of Audit Beliefs . . . 110
2.5.4 Estimated Aggregate Revenue Losses Over Time . . . 112
2.6 Incentives to Under-report: Demonetization, Seller Type, and Mortgage Constraints . . . 112
2.6.1 Impact of India's Demonetization on Under-Reporting . . . 113
2.6.2 Seller Type and Under-reporting Behavior . . . 118
2.6.3 Financial Constraints, Mortgages, and Under-reporting . . . 119
2.7 Conclusion . . . 124

3 Privacy Policies and Consumer Data Extraction: Evidence From U.S. 134

3.1 Introduction . . . 134
3.2 Data . . . 139

3.2.1 Firm data . . . 139
3.2.2 Privacy Policies . . . 141
3.2.3 Data Extraction Behavior . . . 144
3.3 Variation in Privacy Policy Text . . . 145
3.4 Data Extraction Practices and Privacy Policies . . . 147
3.5 Firm Characteristics and Privacy . . . 150
3.6 A Simple Model of Privacy Policies, Visibility and Signaling . . . 153
3.6.1 Environment . . . 154
3.6.2 Separating Equilibria . . . 156
3.6.3 Empirical Predictions . . . 159
3.7 Conclusion . . . 161

Appendices **173**

A Retail Investor Heterogeneity: Evidence from WallStreetBets **173**
A.1 Retail Investors in Academic Research . . . 174
A.2 Tickers removed . . . 175
A.3 Sentiment . . . 176
A.4 Factor Model . . . 178
A.4.1 Traditional Factor Model . . . 178
A.4.2 This Model . . . 180
A.4.3 Why it matters . . . 180
A.5 Robustness Checks Correlations & PCA . . . 181
A.5.1 Stock Level Correlation Matrix . . . 181
A.6 First Principal Component . . . 182
A.7 Talking Share Regressions for all 2,926 stocks . . . 184
A.8 Returns with quadratics . . . 185
A.9 Clustering . . . 187
A.10 Reversal Strategy Returns . . . 191
A.11 Details of the variables . . . 192

A.11.1 Firm Level Variables . . . 192
A.11.2 Reddit Variables . . . 198

B A Bad Bunch: Asset Value Under-Reporting in the Mumbai Real Estate Market 199

B.1 Tables & Figures . . . 200
B.2 Data . . . 228
B.2.1 Transactions Data . . . 228
B.2.2 Circle Rate Scheduled Changes . . . 234
B.3 Simulation to Illustrate Effects of Measurement Error in p . . . 236
B.4 Bunching Elasticity Estimation . . . 237
B.5 Developer Regulation and Under-Reporting . . . 240
B.6 Matching Transactions to Mortgages . . . 241

C Privacy Policies and Consumer Data Extraction: Evidence From U.S. Firms 246

C.1 Proof of Proposition 1 . . . 247
C.2 Dimensions of Expert Evaluation . . . 247
C.2.1 Dimension 1: Data Collection . . . 247
C.2.2 Dimension 2: Consent . . . 248
C.2.3 Dimension 3: Responsible use . . . 249
C.2.4 Dimension 4: Third-parties . . . 250
C.2.5 Dimension 5: User-rights . . . 252
C.2.6 Dimension 6: Overall . . . 253
C.3 Legal Quality: Overall Score and Individual Categories . . . 254
C.4 Variation in Policy Text with LSA . . . 255
C.5 Additional Results on Policy Attributes . . . 256
C.6 Policy Attributes: Service Sector . . . 262
C.7 First and Third-Party Data Extraction . . . 263
C.8 Residuals . . . 265
C.9 Alternative Measure of Legal Clarity . . . 266

List of Tables

1.1 Main Descriptives . . . 65
1.2 Posts Summary . . . 66
1.3 Stock Level Estimates for the Top 1,000 Stocks . . . 67
1.4 Return Predictions . . . 68
1.5 Volatility Predictions . . . 69

2.1 Summary Statistics on Transactions . . . 133
2.2 Elasticity of Reported Value to Transaction Tax Rate . . . 133

3.1 Descriptive Statistics of Firm Characteristics . . . 170
3.2 Descriptive Statistics of Privacy Policy Attributes . . . 170
3.3 Descriptive Statistics of Data Extraction Measures . . . 171
3.4 Policy Attributes: Regressions . . . 172

A.1 Stock Level Estimates . . . 185
A.2 Return Predictions . . . 186
A.3 Clusters Return Predictions . . . 190
A.4 Firm Level Variables . . . 192
A.5 Reddit Variables . . . 198

B.1 Guidance Value Systems in Cities around the World . . . 224
B.2 Under-reporting Before Circle Rate Changes . . . 228
B.3 Circle Rate - Summary Statistics . . . 235
B.4 Data Validation: GIS Tagging . . . 243

B.5 Mortgage Matching Summary Statistics 245

C.1 Overall Score and Individual Categories 254
C.2 Policy Attributes: Controlling for Policy Length 257
C.3 Policy Attributes: Further Firm Controls 258
C.4 Policy Attributes: Excluding Manufacturing Firms 259
C.5 Policy Attributes: Service Sector Only 260
C.6 Policy Attributes: 2-digit SIC Code Fixed Effects 261
C.7 First and Third-Party Data Extraction 263
C.8 Controlling For First and Third-Party Cookies 264
C.9 Policy Attributes and Cookies 265
C.10 Policy Attributes: Alternative Legal Clarity Measure 268

List of Figures

1.1 SIC Codes 50
1.2 Ticker Mentions per Ticker Rank 51
1.3 Average Sentiment and VIX 52
1.4 Method Summary 53
1.5 Individual Talking Matrix 54
1.6 Stock Characteristic Matrix 54
1.7 Characteristic Talking Matrix 55
1.8 Hierarchically Clustered Co-Talking Matrix 56
1.9 PCA Results 57
1.10 Factor Loadings 58
1.11 Principal Component Density 59
1.12 Account Characteristics Importance 60
1.13 Number of Posts per PC1 Terciles 61
1.14 Sentiment and Returns 62
1.15 Longer Horizon Returns 63
1.16 Concurrent Returns and Posts 64

2.1 Simulation Results 126
2.2 Main Results 127
2.3 Heterogeneity Over Time 128
2.4 Reporting Behavior in Days Around Scheduled Guidance Value Increases 129
2.5 Reporting Behavior in Months Around Demonetization 130

2.6 Developer vs. Resale Heterogeneity . . . 131
2.7 Heterogeneity by Mortgage Status . . . 132

3.1 Word Cloud of the Privacy Policy Sample . . . 162
3.2 Distribution of Privacy Policy Attributes . . . 163
3.3 Expert Evaluation of Privacy Policies . . . 164
3.4 Distribution of Data Extraction Measures . . . 165
3.5 Variation in Privacy Policy Text . . . 166
3.6 Average Attributes of Privacy Policies by Quintile of Total Cookies . . . 167
3.7 Average Attributes of Privacy Policies by Firm Size . . . 168
3.8 Average Attributes of Privacy Policies by Knowledge Share . . . 169

A.1 Retail Investors in Academic Research . . . 174
A.2 Accuracy of OpenAI GPT-3 Models . . . 177
A.3 TF-IDF of Posts' text with Positive and Negative Sentiment . . . 178
A.4 Stock Level Correlation Matrix . . . 182
A.5 First Principal Component Loadings . . . 183
A.6 Classification Result . . . 188
A.7 Gaussian Mixture Test Likelihood . . . 189
A.8 Weekly return reversal strategy and VIX . . . 191

B.1 Bunching Estimates for Sao Paulo, Brazil . . . 200
B.2 Sample Comparison to Aggregate Tax Revenues . . . 201
B.3 Propstack-Propequity Match Quality . . . 202
B.4 Heatmap of Transactions in our Final Sample . . . 203
B.5 Correlation between Reported Value, Propequity (Estimated Market) Values, and Guidance Values . . . 204
B.6 Bunching of Reported and Propequity Values Around Circle Values in 0.1% Bins . . . 204
B.7 Robustness: Aggregate Tax Base and Propequity Values by Reporting Behavior Bins . . . 205
B.8 Aggregate Reported and Propequity Values by Reporting Behavior Bins 206

B.9 Truthful Reporting Hypothesis: Measurement Error Contribution by Reporting Behavior Bin . . . 207
B.10 Estimated Measurement Error Assuming Truthful Reporting: Full Sample 208
B.11 Exact Match: Bunching of Reported and Propequity Values Around Circle Values . . . 209
B.12 Exact Match: Aggregate Reported and Propequity Values by Reporting Behavior Bins . . . 210
B.13 Exact Match: Under-Reporting Rate by Reporting Behavior Bins . . . 211
B.14 Agreement Date Backdating . . . 211
B.15 Bunching Estimates for Guidance Value Bins . . . 212
B.16 Loss Function: $\hat{\pi}$ . . . 213
B.17 Model fit . . . 213
B.18 Model-Implied Reported Value Distribution . . . 214
B.19 Efficiency Frontier . . . 215
B.20 Example Price Sheet from Propequity . . . 216
B.21 Weekly Total Estimated Transaction Value . . . 216
B.22 Weekly Estimated Amount Under-Reported (US$ Millions) . . . 217
B.23 Weekly Estimated Under-Reporting Rate . . . 217
B.24 Fraction of Sample with $(r-c)/c < -.05$ . . . 218
B.25 Contribution of Bin to under-reporting in (r-c)/c Bins 180 Pre- and Post-Demonetization . . . 219
B.26 Developer vs. Resale Heterogeneity (Since Jan 2019) . . . 220
B.27 Above vs. Below Median Circle Value Heterogeneity . . . 221
B.28 Reported Value Elasticity to Transaction Tax Rate . . . 222
B.29 Incidence of Mortgages and Under-reporting . . . 222
B.30 Equilbrium Relationship between $\frac{R}{P}$ and $TLTV$ . . . 223
B.31 Estimating Total Revenue Loss for the Government . . . 223
B.32 Original Transaction Document: An Example . . . 232
B.33 Translated Transaction Document: An Example . . . 233
B.34 Coverage Validation . . . 234

B.35 Sub-Zones 236
B.36 Mortgage Balance Tests 245

C.1 Variation in Policy Text in 100 Latent Dimensions 255
C.2 Privacy Policies and Cookies 256
C.3 Privacy Policies and Cookies 262
C.4 Performance of Classifiers 267

Chapter 1

Retail Investor Heterogeneity: Evidence from WallStreetBets

1.1 Introduction

The participation of retail investors in the stock market has increased in recent years[1], but their role and sophistication remains disputed in the finance literature. A recent example is offered by Welch (2022) and Barber, Huang, Odean, and Schwarz (2022) who study the same dataset and arrive at different conclusions. More generally, while retail trades have been shown to predict future stock returns (Kaniel, Saar, and Titman (2008); Kelley and Tetlock (2013); Boehmer, Jones, Zhang, and Zhang (2021)), many have shown that retail investors trade too much (Odean (1999)), have biases (Barber and Odean (2008)), increase volatility (Foucault, Sraer, and Thesmar (2011)), and in some cases affect market liquidity (Barrot, Kaniel, and Sraer (2016); Eaton, Green, Roseman, and Wu (2022)). This literature has usually studied retail investors in an

[1]Martin, K., & Wigglesworth, R. (2021, March 9). "Rise of the retail army: the amateur traders transforming markets". Financial Times.

aggregated fashion, assuming away heterogeneity, thus suggesting a large portion of retail investors share those characteristics, even though some have challenged this assumption (Kumar (2009); Dorn and Sengmueller (2009); Fong, Gallagher, and Lee (2013)).

In this paper, I use conversations on the WallStreetBets subreddit (WSB) between January 2018 and December 2021 to ask whether retail investors have systematic differences. The goal of this exercise is to understand if retail investors can be separated into distinct subgroups based on their interests for certain stocks at specific times, by expanding on a new technique, akin to a factor model, specifically tailored to root out heterogeneity. This method reveals three distinct subgroups.

WSB provides an interesting dataset as it has generally been portrayed as a disreputable forum of gamblers and thrill-seekers[2]. If diversity and sophistication, can be found there, it does make the case for retail heterogeneity even in settings in which retail investors self-select. Most notably, I show a surprising fact: despite the forum's reputation, a large portion of Redditors show interest in stocks with strong fundamentals.

This social media dataset is crucial in answering this question, by providing a large and high-frequency cross-section, with over 1.2 million accounts, as research on individual investors has been hindered by lack of data[3]. One could have turned to administrative data, but those are not only rare but also hard to access[4]. Surveys suffer from satisficing and social desirability bias (Krosnick (1991)) as people tend to provide

[2]Weisenthal, Joe; Alloway, Tracy (March 5, 2020). "How a Profane Subreddit Moved the Market (Podcast)". Odd Lots (Bloomberg)

[3]Grinblatt and Keloharju (2001): "Up until now, the empirical analysis of what makes investors trade has been hindered by limited and incomplete data about the financial markets. Work [...] either focuses on a small segment of the market that may not be representative and/or limits the analysis of trading to single issues, like contrarian behavior or the aversion to losses." This statement remains largely true as of 2022.

[4]For obvious reasons, as demonstrated by Narayanan and Shmatikov (2008).

answers that either match the interviewers' expectations, or social norms[5]. WSB allows for a different approach, looking at what individuals say - not what they answer, and therefore in a setting in which such concerns are unlikely to apply. Specifically, the posts and comments from WSB reveal individuals' preferences for talking about certain stocks. Thus, I know the *who* (who posts), the *what* (what they post about), the *when* (when they post), the *where* (on Reddit), the *how* (how they post) and a new method, pioneered by Balasubramaniam, Campbell, Ramadorai, and Ranish (2022), provides a framework in which I can make headway on the *why* (why they post about what post) and thus, infer interests. Expanding on this methodology unearths a remarkable amount of heterogeneity.

I first show that the overall sentiment of the forum correlates with the VIX, lending validity to the overall exercise. To do so, I extract the sentiment of WSB posts and classify them as either positive, neutral, or negative using the GPT-3 language model developed by OpenAI. I then proceed to create a profile for each of the Redditors who participate in conversations on WSB, characterising them by attributes such as age of their accounts, number of posts, or the percentage of words misspelled, but also by their preference for talking about stocks with certain attributes. To do so, I build on the technique introduced by Balasubramaniam, Campbell, Ramadorai, and Ranish (2022), wherein the characteristics of a stock at the time an individual chose to talk about it provides information about the individual's interests. I find that there is considerable diversity amongst those Redditors, with three clusters clearly visible. The main differentiation is in their respective interest in either stocks with strong fundamentals - especially in terms of growth - and those with high volatility. Subsequent differentiation relates to firm size and age, cash flow volatility, R&D spending, and

[5]This was additionally recently expressed by Richard Flynn, Charles Schwab's UK managing director, finding that two-thirds of respondents to a survey said that ethical stocks make for good investments, while only a small percentage actually invested in ethical strategies (Source: https://www.trustnet.com/news/13299392/esg-not-a-priority-for-retail-investors,-study-finds).

finally indebtedness. Those differences are also connected to individual attributes that can be gathered from their Reddit accounts directly: age of the account, the number of firms they talk about, the sophistication of their sentences, and the propensity to make spelling errors. Individuals interested in volatility have younger accounts, appear less sophisticated, with simpler sentences, and more spelling errors. They also tend to participate in discussion during trading hours and to be more interested in Reddit posts related to news events than their counterparts. I also show that those characteristics matter for the stocks individuals talk about. As such, I find that the WallStreetBets subreddit, commonly regarded as an entity in and of itself, is composed of several groups whom, for the most part, do not mix.

I then explore the consequences of such differences by analysing the performance of subgroups; by checking whether one could use their conversations in a predictability exercise. I find that the posts of a first subgroup, individuals with an interest in volatile stocks, are particularly useful at predicting future returns. This result continues to hold even after I add to the set of predictive variables the measure of retail order imbalance as defined by Boehmer, Jones, Zhang, and Zhang (2021). The fundamental investors also appear useful. Furthermore, I show that both those subgroups show a spike reversal-pattern (Engelberg, Sasseville, and Williams (2012)), with the returns reversing over the course of the following weeks. I also show that both fundamental and volatility investors talk about stocks whose volatility will increase in the near future, with those interested in volatility showing, once again, a much stronger coefficient. Finally, I show that all subgroups discuss both momentum and reversal strategies (Nagel (2012)).

Those results are interesting on several dimensions; the first of which is methodological. The approach taken to gather sentiment has implications for the text analysis and sentiment literature (Tetlock (2007); Loughran and Mcdonald (2011)). To my knowl-

edge, it is the first, in the finance literature, to make use of language models to extract sentiment, as opposed to more traditional methods such as dictionaries and other statistical approaches. Most notably, the language model I use requires few training examples (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever, and Amodei (2020)) and in this case reaches an 86% accuracy rate, well over what could have been obtained through traditional techniques (Antweiler and Frank (2004); Li (2010); Chen, Despres, Guo, and Renault (2019)).

This paper also has implications for the literature on retail investors, which has expanded over the years, with, as of August 2022, well over one in ten of recently published papers mentioning "retail investors" at least once[6]. One should certainly not expect this trend to reverse in the coming years, especially in light of the GameStop events of 2021. Indeed, internet chat boards - and more particularly Reddit and WallStreetBets - have been at the centre of attention, with numerous papers making use of those public data (Hu, Jones, Zhang, and Zhang (2021); Pedersen (2022); Fedyk (2022); Bryzgalova, Pavlova, and Sikorskaya (2022)). First, the characteristics that separate different types of Redditors are reminiscent of Baker and Wurgler (2006) who define securities with subjective valuations as either extreme growth, high volatility, small, young, or distressed. Those are the very attributes that separate Redditors. As it should: if sentiment has larger effect on securities whose valuation are highly subjective, those are the very securities that should be discussed.

Additionally, some characteristics are connected with others, in particular suggesting that the gambler in Kumar (2009) is the same as the sensation-seeker of Grinblatt and

[6]For papers published in the Journal of Finance, Journal of Financial Economics, and Review of Financial Studies. See Section A.1 of the appendix for a graph showing the number of mentions since 2000.

Keloharju (2009) and the attention-driven of Barber and Odean (2008). Those results, in particular the correlation between interest in volatility (and other daily changing characteristics) and sophistication, may also link cognitive ability to excessive trading: if the less sophisticated are more likely to discuss and be interested in volatile or attention grabbing stocks, which changes at a higher frequency than fundamentals, it would explain why they trade more (Odean (1999)), but would also explain other biases (Grinblatt, Ikäheimo, Keloharju, and Knüpfer (2016)). Those traits also add to the debate on the level of sophistication of retail investors (Odean (1998, 1999); Grinblatt and Keloharju (2001); Kaniel, Saar, and Titman (2008); Kelley and Tetlock (2013); Boehmer, Jones, Zhang, and Zhang (2021); Barber, Huang, Odean, and Schwarz (2022)), in which the results differ, by highlighting a cluster of individuals interested in stocks with strong fundamentals.

Furthermore, the fact that investors with different interests, and different account level attributes, discuss different stocks is further proof of the existence echo chambers (Cookson, Engelberg, and Musslins (2021); Pedersen (2022)), in which only Redditors with similar interests discuss certain stocks. Those echo chambers also matter for the results on predictability, which themselves contribute to the literature on retail generated research (Antweiler and Frank (2004); Chen, De, Hu, and Hwang (2014); Jame, Johnston, Markov, and Wolfe (2016); Farrell, Jame, and Qiu (2020)). Indeed, I find effects to depend on who is talking, and therefore what is being talked about. In particular, those interested in volatility show stronger coefficients in a predictability exercise, suggesting that the value, in terms of predictability, of retail investors matters for a limited number of stocks. Those interested in volatility also predict higher future volatility than the average, again suggesting that the results of Foucault, Sraer, and Thesmar (2011) and Eaton, Green, Roseman, and Wu (2022) are stronger for a sub-sample of investors and stocks.

Finally, the fact that all groups discuss both momentum and reversal strategies (a finding which is consistent with Welch (2022); Fedyk (2022)) raises further questions with respect to their market impact, especially in terms of liquidity. In particular, Eaton, Green, Roseman, and Wu (2022) find Robinhood outages to be associated with increased market liquidity, which may not be the case during times of financial market turmoil (Nagel (2012); Barrot, Kaniel, and Sraer (2016)).

The remainder of the paper is organized as follows. Section 1.2 describes the data, in particular the WSB data, and provides basic descriptive statistics. Section 1.3 describes the method employed to reveal the heterogeneity of retail investors while Section 1.4 characterizes the main types. Sections 1.5 and 1.6 explore the importance of separating individuals into those broad types. Finally, Section 1.7 concludes.

1.2 Data

1.2.1 Reddit

Reddit is a message board website created in 2005 by Alexis Ohanian, Steve Huffman, and Aaron Swartz. Members can submit content the website, which are then commented and voted upon by other members. As of October 2022, it is the 19th most visited site in the world, which and average of 1.7B monthly visits[7]. The website is divided in into more than 130,000 user created communities known as "subreddits". The largest of which, r/Announcements has - as of October 2022 - more than 177 million subscribers. There are multiple subreddits with over 1 million subscribers that discuss the stock market: r/Stocks, r/Investing, r/StockMarket, etc. In this paper, I

[7]Reddit.com Traffic Analytics & Market Share. Similarweb. Retrieved October 25, 2022, from https://www.similarweb.com/website/reddit.com/.

focus on what is now the largest of those communities r/WallStreetBets, also known as "WSB". This community was created in 2012 by Jaime Rogozinski and has seen its subscriber base increase to 12.8 million as of October 2022.

Reddit & Pushshift API

I gather the posts and comments from WallStreetBets between January 2018 and December 2021 in two manners. The first is through the use of the Pushshift API[8]. This API collects all data and metadata related to a post or comment at the time of its posting. For each submission, the API returns the title, text, poster user id, number of comments along with metadata such the time of submission.

I then augment those data using the Reddit API itself. While the Reddit API does not allow for the download of all posts for a given day - which is why I first rely on the Pushshift API - it does allow to search for an individual post or comment by using its ID. This allows for the gathering of other metrics such as the score of the post (number of upvotes - number of downvotes), the ratio of upvotes and whether said comment or post was subsequently deleted (either by moderators or by the user itself).

Getting the Tickers

Unfortunately, there is no standardized manner in which users express the ticker or the name of the company or companies their post relates to. Given that two letter words might be misclassified as tickers, I rely on Part-of-Speech (POS) tagging, also called grammatical tagging in order to capture the tickers. I first mark the words in a given post as corresponding to a particular POS (noun, verbs, etc.). I then limit the search for tickers to words classified either as nouns or pronouns with no more than five

[8] https://www.pushshift.io

characters. The list of tickers to search for is taken from CRSP. Finally, tickers that are also commonly used as abbreviations by Redditors are removed. Those include tickers such as "YOLO" (which, perhaps unsurprisingly, is also the ticker of a Cannabis ETF), "DD" (Due Diligence, a type of post, but also the ticker of DuPont), "RH" (Robinhood, which is frequently mentioned due to its commission-free trades - but is also a trading symbol for a different company), "CEO", etc. The full list of ticker removed as well as the reason for their removal is available in Section A.2 of the appendix.

For the comments, I proceed in the following three manners. First, I search for tickers in the comment itself. Second, I assign to the comment the same tag as the post it is commenting on. Finally, since comments are hierarchical, I also assign the tickers of any parent comment. The latter is particularly necessary for the most popular type of post, categorized as "Daily Discussion" and created once a day, before the US market opens, by the moderators. This post will not feature any ticker. However, Redditors are invited to start their own discussion in the comments, often mentioning a ticker.

Sentiment

Ever since text data became available to researchers, measuring sentiment has been critical to many academic papers and this paper is no different. The literature has generally used two approaches: dictionary and statistical, both limited by their simplicity, both lacking the ability to understand context. The dictionary approach was further developed by Loughran and Mcdonald (2011), who found that traditional dictionaries (such as the Harvard IV-4) are not adapted to the financial context (or at least, to the financial context of 10-Ks) and misclassify common words such as "liability", suggesting the need to create a dictionary for any given context. This was further proved by Chen, Despres, Guo, and Renault (2019) who found an 86% accuracy rate - on two

classes classification, positive and negative - when devising a lexicon specifically for cryptocurrency tweets (considerably higher than any other off-the-self dictionary they tried, including the Loughran and McDonald). This rather long and tedious approach is still problematic for longer texts, wherein context and preceding sentences do matter. It is also time dependant: if either the producers of text or the vocabulary used change, the dictionary is outdated. This is particularly relevant here since the WallStreetBets community has significantly increased in size over the time period studied. The statistical approach, on the other hand, has generally focused on using simple models such as Naive Bayes. Li (2010), classified 30,000 sentences and used Naive Bayes only to find a cross validation accuracy no higher than 67% (on three classes classification) while Antweiler and Frank (2004) did not report any out of sample metrics but reported an in-sample one of no more than 77% on their minority classes. Again, this method is tedious - since it requires the manual classification of a substantial training sample - with relatively high error rates.

Removing the limitation of both approaches is desirable and language models can do just that. Those probabilistic models are trained so as to predict the next word in a sequence given the words that precede it. While they have been developed with only this specific task in mind, they can be further fine-tuned on a specific task. This fine tuning is however far less tedious than the aforementioned methods as those models have been shown to have one particularly desirable feature: they are few shots learners, meaning they do not need many examples to perform well (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever, and Amodei (2020)). One of those language models, the one I use in this paper, is GPT-3 (Generative Pre-trained

Transformer 3), and was developed by OpenAI[9]. In particular, I fine tune one of their model to characterize reddit posts as either positive, negative or neutral. The full procedure is explained in Section A.3 of the appendix.

1.2.2 CRSP, Compustat, IBES, & SEC

The returns data comes from CRSP. I keep only common stocks (CRSP share codes 10 and 11), meaning ADRs (American Depositary Receipts) such as BABA (the NYSE ticker for Alibaba, which is mentioned on WallStreetBets) will not be featured. I augment those data with Compustat to allow for the calculation of firm characteristics at the annual and quarterly levels. I also add a summary of analyst estimates available through IBES. Finally, given that the Reddit dataset is at the daily level, I make those data daily, forward filling the missing values.

In order to link properly Compustat to the moment the information is available to investors I use SEC data, and more particularly the timestamp at which the annual and quarterly forms are released (such forms mainly include 10-K, 10-Q, N-CSR, and N-CSRS, along with their amended equivalents). For those that I am unable to match, I either use a 90 days offset for annual values, or a 45 days one for quarterly ones.

I consider a large collection of 149 stock characteristics - mostly following Gu, Kelly, and Xiu (2020), which is based on the cross-section of stock returns literature - the full list of which is available in the Section A.11 of the appendix.

[9] https://openai.com/api/

1.2.3 TAQ

Finally, I also use TAQ data to calculate retail order imbalance, following the method of Boehmer, Jones, Zhang, and Zhang (2021). In particular, I keep all trades reported to a Financial Industry Regulatory Authority (FINRA) trade reporting facility (TRF) - those will have "D" as their exchange code in TAQ - and given a transaction price $P_{i,t}$ for stock i at time t, I create $Z_{i,t} = (100 * P_{i,t})\ mod\ 1$. I then classify all trades with $Z_{i,t} \in (0.6, 1)$ as retail buy and with $Z_{i,t} \in (0, 0.4)$ as retail sell. Then, for stock i on day t I create marketable retail volume order imbalance:

$$OrderImbalance_{i,t} = \frac{RetailBuyVolume_{i,t} - RetailSellVolume_{i,t}}{RetailBuyVolume_{i,t} + RetailSellVolume_{i,t}}$$

Note that From 2016 to September 2018, the SEC implemented a tick size pilot program (TSPP), which, as described in Boehmer, Jones, Zhang, and Zhang (2021), limited the ability of brokers to provide price improvement for affected stocks. This time period overlaps with the one in this paper, between January 2018 and September 2018.

1.2.4 Summary Statistics

Posts

The final dataset consists in over 1.5 million posts and over 35 million comments. The years 2020 and 2021 saw a considerable increase in the popularity of the WallStreetBets subreddit, with the number of subscribers - those who subscribe will see WSB posts on their Reddit home page - increasing from less than five hundred thousand in 2018 to more than ten million in 2021. While I only observe users that chose to participate in the subreddit by way of posting or commenting, the sample consist in a little over 1.2

million unique authors, talking about 6,038 tickers[10]. As shown in Table 1.1, the vast majority of posts were written in 2021, with the comments almost evenly split between 2020 and 2021. About 2/3rd of the posts are deleted by either the author himself or moderators for a variety of reasons, such as not having grasped enough interest within minutes after posting, or not following posting guidelines.

Posts are tagged by their creator into subcategories, called flairs, which are thoroughly enforced by the moderators and informs the reader about the content of the post and its general seriousness. The most popular flairs are Discussion, YOLO, Meme, Gain and Loss, News and Due Diligence (DD). Posts under "Discussion" are generally an idea or article the user would like to talk about. DD are posts about research the user has done about a particular company. By design they tend to be longer in length, should be high effort, include source and citations. "YOLO" are generally screenshots of trades with substantial amounts at play while Gain and Loss show either a substantial winning or losing trade. I classify the top 11 flairs into 3 categories: news, information and entertainment. Table 1.2 shows summary variables of the posts. In this setting over 87% of all posts are assigned to a class, with the majority being entertainment posts. As expected, information post are the longest, with an average of 24 words. Among the 23% of posts featuring a ticker, almost 70% are positive, and 12% are negative.

Tickers

Figure 1.1 plots the Industry classification of the firms mentioned in the 23% of posts that mention one ticker. The repartition between industries is constant across time, save for the GameStop (GME) period in early 2021. In normal times, manufacturing

[10]This number will be further reduced in Section 1.3, where I will only keep tickers individuals are actually allowed to talk about, as per WallStreetBets guidelines

companies make the bulk of the posting, with services and retail trade completing the top 80%.

The constant distribution is perhaps unsurprising given that few tickers dominate the discussion on WallStreetBets. Figure 1.2 plots the rank of a given ticker against the number of times it appears in a post or comment. Very few tickers appear in more than 1,000 posts, with only the 500 most mentioned tickers being so over 100 times. This fact, also documented by Semenova and Winkler (2021) will be of particular importance in Section 1.5.

1.2.5 Sentiment

Considering the approach used in order to create the sentiment metric, a language model, and the fact that the training labels have been decided by only one individual, validating the sentiment metric is crucial. One approach would be to juxtapose it to other measures of sentiment. Figure 1.3 compares the monthly aggregate sentiment of Reddit posts with tickers[11] against the VIX. The correlation is rather striking and proves two points. First, the language model appears to work extremely well, which should not be surprising since out-of-sample accuracy is greater than 86%, as shown in Section A.3 of the appendix. Second, it casts doubts as to the alleged unsophisticated of Redditors. They indeed appear to be either attentive to the overall market or have access to similar sources of information. This lends validity to the exercise.

[11] A value of +1 is given to positive posts, and -1 to negative ones. Averaging the sentiment at the monthly level yields the sentiment score.

1.3 Characterising Individuals

In this section, I expand on the method introduced by Balasubramaniam, Campbell, Ramadorai, and Ranish (2022) and study the heterogeneity of the cross-section, setting up the framework with which I will distinguish investor types. Essentially, instead of characterizing redditors by the stocks they talk about, I characterize them by the characteristics of the stocks that they talk about at the moment they talk about about them. Figure 1.4 sums up the method: an individual i talks about n different stocks (x_n) at different dates (t), all of which have different k characteristics on those dates ($C^n_{t,k}$), which in turn, when averaged, yield the characteristic tilts of an individual i (C^i_k).

1.3.1 Stock Talking Matrix

I first construct the stock talking matrix shown in Figure 1.5. This matrix is a collection - through time - of two-dimensional I by N matrices which represent the number of posts or comment each individual makes about a particular ticker on a given day, $x^i_{t,n}$ with n being a particular stock, i the individual in question, and finally t being the date. Those matrices are extremely sparse and would offer a clearer picture if averaged over multiple days - or the entire sample. I build them at the daily level because those matrices will be multiplied by the corresponding stock characteristics defined in section 1.3.2 which will incorporate variables constructed at the daily level. In section 1.5, I will introduce the talking share $Q_{i,n}$ which sums those matrices through time and divides each row (individual) by its total. I also define x^i_t as the row vector of the talking of individual i at time t.

The size and sparsity of this matrix is the very problem this setting is made to deal

with. Given the objective to group individuals, one could perfectly imagine doing so on an averaged version of this matrix. There would however be two problems with this approach. Firstly, clustering algorithms struggle with sparse data and reducing the dimension is generally advised: this is precisely what this technique does. Secondly, and perhaps more importantly, the characteristics of a given stock vary through time, meaning Apple in 2001 was not the same as Apple in 2020. And GameStop in 2020 - pre short squeeze - was certainly not the same as it was in 2021.

1.3.2 Stock Characteristic Matrix

I then create the stock characteristic matrix shown in Figure 1.6. This matrix is again a collection through time of two-dimensional matrices. Those matrices are of N by K dimensions with N being again the number of stocks in our sample and and K being the stock characteristics. On any given day, I rescale each of those characteristics to have mean zero and support $[-0.5, 0.5]$, essentially ranking the stocks, an approach undertaken by both Balasubramaniam, Campbell, Ramadorai, and Ranish (2022) and Gu, Kelly, and Xiu (2020).

This matrix has for entry $C^n_{t,k}$ with with n being a particular stock, k a particular characteristic, and finally t being the date. I also define the vectors $C_{t,k}$ being the column vector of a particular characteristic of all stocks in the sample.

Given that the data spans three years - from 2018 to 2021 - some information will be missing for some stocks. For example, Wish, which is mentioned in over 2,800 posts, went public through its parent company ContextLogic in December 16th 2020 and its characteristics are therefore missing for some parts of the sample[12]. I ignore those

[12]Compustat does have data going back to 2018, but such information was only released as part of their S-1 form, which was filed in November 2020 and therefore not available before then.

stocks for the calculation of the scaling and subsequently set the missing characteristics to zero. Since the data is unavailable to us, it is likely it was also unavailable to Redditors when discussing the stock, meaning those attributes are unlikely to have impacted their decision to discuss the stock to begin with.

Additionally, Redditors on WallStreetBets are limited to talking about stocks above a certain market capitalization. Any post about stocks under this threshold will be considered spam and promptly removed. The threshold itself varies over time, having been \$1B for the most part of the sample[13]. It was recently increased to \$1.5B[14]. Given this, I limit the sample to stocks with over \$1B in market capitalization, which will impact how the characteristics are rescaled, thus limiting it to 2,926 stocks.

1.3.3 Characteristic Talking Matrix

Finally, I create the characteristic talking matrix shown in Figure 1.7 by multiplying the previous collections of two-dimensional matrices together (one by one, at the date level). This matrix is of size $I \times K \times T$. Those matrices have for entry $y^i_{t,k} = x^i_t C_{t,k}$. Note that this matrix is not yet normalized, meaning the attributes will be higher for individuals who talk about more stocks.

Computing those matrices is not trivial. In fact, the dimensions of the stock talking matrix is why this approach is needed in the first place, starting from a $I \times N \times T$ matrix and arriving at a $I \times K \times T$ matrix, with $K \ll N$. In practice, I do not compute stock talking matrix defined in section 1.3.1 as the number of cells it contains would be over 5 trillions[15]. Instead, on any given day, the dimensions are reduced to $I_t \times N$ where I_t is the number of individuals that participate to the conversation on day t.

[13] https://web.archive.org/web/20210401081427/https://www.reddit.com/r/wallstreetbets/about/rules/
[14] https://web.archive.org/web/20220322110051/https://www.reddit.com/r/wallstreetbets/about/rules/
[15] $1,211,650$ individuals $\times 2,926$ stocks $\times 1461$ Days $= 5,179,665,621,900$.

This makes the data manageable.

Collapsing the time dimension, summing the characteristics talking matrices for all individuals, and dividing each row by the total number of relevant posts (i.e. with ticker) individuals yields the characteristic talking matrix with $z_{i,k} = \frac{1}{\sum_T \sum_N x^i_{t,n}} \sum_T y^i_{t,k}$ in its entries. Similarly to Balasubramaniam, Campbell, Ramadorai, and Ranish (2022), this $I \times K$ matrix describes the talking-weighted average tilts towards K characteristics. Further removing the characteristics of the average individual and creating $\tilde{z}_{i,k}$ such that $\tilde{z}_{i,k} = z_{i,k} - \frac{1}{I} \sum_{i=1}^{I} z_{i,k}$ creates the object which I will use in the rest of this paper, an object particularly suited to study the heterogeneity of retail investors. Any similarity to a factor model is not unintentional and section A.4 of the appendix explores this angle and further shows why this method is appropriate for this particular exercise.

1.3.4 Characteristic Co-Talking Matrix

I am then able to create the $K \times K$ characteristic co-talking matrix which will take in its entries $\sum_{i=1}^{I} \tilde{z}_{i,k_1} \tilde{z}_{i,k_2}$, akin to a variance co-variance matrix. Further normalizing this matrix along the diagonal yields a correlation matrix which is shown in Figure 1.8. The heat map reflects the intensity of the correlations between the characteristics tilts across Redditors, with red and blue representing positive and negative correlations respectively. Two characteristics will be positively correlated if both are of interest to a large number of Redditors.

In plotting this matrix, and for ease of interpretation, I arbitrarily take a subset of variables. First, I keep a set of fundamental variables: book to market, beta, asset growth, return on assets, market cap. Second, I keep a set of stock price related variables: skewness, volatility, momentum, previous day return squared as well as price itself.

I also keep analyst estimates: mean price target to price. Finally, I include a dummy variable which takes the value of one in the week following earnings' release.

The ordering of this matrix is not random: it is hierarchically clustered. This hints at different types of investors by way of correlation between variables - asset growth, return on assets and employee growth rate for example. Indeed, tilts towards fundamental variables are correlated with one another; tilts towards returns on assets, price target, asset growth, employee growth rate, week of earnings and beta are grouped together. However, those are negatively correlated with tilts towards momentum, skewness, volatility. The negative correlation of the latter cluster with price is reminiscent of the propensity of some individuals to invest in lottery type stocks found by Kumar (2009). Note that this correlation should decrease in the future with the recent advent of fractional shares (see Bartlett, McCrary, and Maureen (2022))[16]. Interestingly, individuals with a tilt towards stocks with large gains or loses, either due to attention catching (Barber and Odean (2008)) or sensation seeking (Grinblatt and Keloharju (2009)), are highly correlated with the aforementioned cluster, suggesting this type of investor might in fine be the same as Kumar (2009).

The objective when designing and studying this matrix is to show that, at the individual level and in opposition to the stock level, correlations exist - either positive or negative - between different attributes. One possible concern is that those characteristics might already be correlated at the stock level, which would then be reflected here. I deal with this issue in two ways. First, I remove individuals who only talk about one stock so as not to capture those stock level correlation. Second, I plot in Section A.5.1 of the appendix the correlation matrix done at the stock level. This matrix shows that none of those correlation either exist or exist as strongly as they do in Figure 1.8.

[16]Interactive Brokers started offering fractional trading on November 25, 2019, Robinhood on December 12, 2019,Fidelity on January 29, 2020 and Charles Schwab on June 9, 2020 for stocks in the S&P500.

1.4 Unobserved Multi-factor Model

Many of the account level or characteristic tilts variables we have considered so far are correlated with one another - momentum, skewness and volatility for example. For the next part, and more particularly when dividing individuals into different groups, it would be advantageous to reduce that number so the analysis may be tractable. I therefore now construct an unobserved multi-factor model using PCA, again leaving out individuals who talk only about a single stock.

1.4.1 Principal Components Analysis (PCA)

The principal components are the eigenvectors of the covariance matrix partially showed in Figure 1.8. Figure 1.9a shows how much of the variance each principal components explains, with the first explaining over 25% of the total variance and the first four explaining over 50%. Plotting the eigenvalue ratios (ER) as defined by Ahn and Horenstein (2013), where $ER(k) = \frac{\tilde{\mu}_k}{\tilde{\mu}_{k+1}}$ with $\tilde{\mu}_k$ being the k^{th} eigenvalue, yields Figure 1.9b. They suggest limiting one's analysis to the factor k for which the ER is maximized unless there is a dominant factor. In this application, it would mean stopping after the first PC, which I do in Section 1.6. Nonetheless, the next two principal component are instructive and of almost equal importance (the eigenvalue ratio of the second principal component to the third is almost equal to 1).

Figure 1.10 plots world clouds of variable names based on their t-statistics from a univariate regression on the first three PCs. The left panels show variables with a positive loading on the PC of interest while the right panels are those with negative loadings. Both the size and colour intensity of the writing reflect the relative importance of each variable with the most important ones appearing larger and darker.

The first principal component (PC 1) differentiates between the two types we previously mentioned in Section 1.3: investors interested in fundamentals, as shown in the world cloud to the left of Figure 1.10a whom have a positive loading on the first principal component and those interested in lottery like stocks as defined in Kumar (2009): stocks with high volatility, high skewness, high earnings volatility, who will have a negative PC 1 loading. Since having a liking for stocks with either low or high returns on the previous day (the variable "Previous Day Return Squared"), is connected to this factor, it is also further confirmation that this penchant is connected to Barber and Odean (2008) and Grinblatt and Keloharju (2001), as suggested by Welch (2022).

The second principal component (PC 2), as shown in Figure 1.10b, seems to capture a size and age effect. On one hand, there are those interested in well established stocks, with large market capitalization whom will have a negative loading on PC 2. On the other hand, there are those who are interested in newer stocks with high R&D spending relative to sales, and high cash flow volatility.

Finally, the third principal component (PC 3), as shown in Figure 1.10c, seems to capture some indebtedness effect. Individuals with a negative loading on this component are interested in more leveraged firms, with a lower share price, but which remain profitable as suggested by the importance of their cash flow to price and a loading on returns on equity. Those with a positive loading on this component are more interested in firms with large but productive cash holdings.

A possible concern here is the number of volatility related metrics included. If such metrics are correlated with one another, which they are, one should expect the first principal component to indeed be connected to volatility as it would explain a large amount of variation within the cross-section. Section A.6 of the appendix plots the first PC when excluding keeping only volatility and skewness over 252 days, and mo-

mentum only over 90 days. The first principal component remains the same.

1.4.2 Cross Sectional Heterogeneity

Looking further into the first three principal components, there appears to be clusters of individuals. Figure 1.11 shows the densities of the first principal component with respect to the second and third. The top figure, which shows the density of the first and second principal components shows at the very least three clusters, one loading heavily on the first PC - individuals interested in fundamentals - and the other two loading negatively on it, showing individuals mostly interested in stocks with gambling like features. Those two clusters load differently on the second PC, with one loading positively, showing an interest for newer stocks of which the cash flow is more volatile, and the second loading negatively, showing an interest in large and older stocks. Additionally, the third principal component, as show in the bottom graph of Figure 1.11 shows a further divide among those interested in volatility: a group loading positively on the third PC, interested in firms with large amounts of debt and an another focusing mainly on firms which appear more stable - more profitable with larger cash holdings.

Those clusters are reminiscent of Baker and Wurgler (2006) in which they argue that investor sentiment has larger effects for securities whose valuations are highly subjective and difficult to arbitrage. Consequently, those securities are those that investors should discuss the most. The fundamentalist cluster (high PC1) load positively on extreme growth stocks, a type of stock harder to value and for which sentiment should matter more according to Baker and Wurgler (2006). The other two load on volatility, another type of stocks whose valuations are subjective. But amongst those two, one loads positively on the second principal component, focusing more on stocks that are small, young and rely on external founding (change in shares outstanding). A fur-

ther differentiation also comes in the form of those showing further interest in stocks with large amounts of debt but also intangibles, thus harder to value (with a negative loading on the third principal component).

The account characteristics do not appear particularly important in Figure 1.10. However, they remain interesting when attempting to further establish a profile for individuals in the different clusters. Looking at the first principal component and the different account characteristics the data allows to establish, they paint a clearer picture of the individuals interested in stocks with high volatility. Figure 1.12 shows the value of the t-statistics coming from univariate regressions on the first PC. Individuals with a negative loading on the first PC - those that like volatile stocks - tend to post and comment just before the market opens, tend to be particularly interested in news posts, tend to misspell words more and finally post more during trading hours than their counterpart, those interested in fundamentals. On the other hand, fundamentalists talk about more stocks, comment more, make for longer texts all of which can be explained by the fact that their accounts tend to be older. But their texts are also harder to read as shown by the SMOG and Fog indexes.

This also connects to the literature, in particular with Barber and Odean (2008): those who are interested in stocks with attention grabbing features, and in particular extreme one-day returns, are also more likely to be interested in news. It also shows that some retail investors are not impacted by this, ignoring such discussions. Additionally, this provides further evidence of a channel through which one might explain why the less sophisticated trades more, as in Odean (1999) (if the less sophisticated do indeed use discount brokerage accounts, as suggested by Fong, Gallagher, and Lee (2013)). If those individuals are more likely to discuss and be interested in volatile or attention grabbing stocks, which are attributes that change at a higher frequency than fundamentals, it would explain why they trade more.

1.5 Clientele Effects

Individuals are now characterized not by the stocks they talk about but by the characteristics of those stocks at the moment they talk about them. In this section, I show that those characteristics matter for the stocks individuals talk about. I show here that individuals with certain characteristics are more likely to discuss certain stocks. As such, I find that this subreddit, commonly regarded as an entity in and of itself is composed of several groups whom, for the most part, do not mix.

1.5.1 Regression Model

For each stock n, I can run the following regression:

$$Q_{i,n} = \alpha_n + \sum_{k=1}^{K} \beta_{n,k} F_{i,k} + \epsilon_{i,n}$$

Where:

- $Q_{i,n}$ is the talking share of individual i for stock n
  - This represents the number of times individual i has participated in conversations about stock n (x_n^i), divided by the total number of times individual i has participated in conversations about stocks ($\sum_N x_n^i$).
  - As an example, if an individual participates in three conversations about Apple, and one about AMC then:
    - $Q_{i,AAPL} = 0.75$
    - $Q_{i,AMC} = 0.25$
    - $\forall n \notin [AAPL, AMC], Q_{i,n} = 0$
- $F_{i,k}$ is the account characteristic k for individual i, which can be:

- An account characteristic, such as total number of posts or comments.
- A characteristic tilt ($z^{\sim}_{i,k}$). If the individual only talks about one stock, I set those characteristics to zero and add dummy variable.

1.5.2 Results & Echo Chambers

Table 1.3 summarizes the results for the top 1,000 stocks of the 2,926 regressions ran on the talking share of each of the stocks in the sample. I chose to focus arbitrarily on the top 1,000 as the vast majority of stocks are only mentioned a few times and only a small fraction of stocks dominate the discussion, as shown in Figure 1.2. Section A.7 of the appendix shows the result for the 2,926 regressions. For ease of interpretation, I only consider the characteristics showed in the correlation matrix in Section 1.3.4, which are those linked to the first principal component showed in Section 1.4. The table reports the average t-statistic of those variables along with percentiles. It also reports the percentage of regressions showing the variable in question to be significant at the 5% level.

The vast majority of the account characteristics are significant on average with the single stock dummy taking the lead. The Reddit account age, the propensity to post or comment on weekends or during the week for entertainment or information as well as the linguistic complexity are significant as well. On the characteristic tilts front, all variables are significant, with market capitalization, returns on assets, book-to-market, earnings volatility leading.

Importantly the different variables have cross-sectional variation (as in Balasubramaniam, Campbell, Ramadorai, and Ranish (2022)). Not only are most of the t-statistics significant on average, but their respective values vary widely across regressions. Most metrics shown, have t-statistics of which the sign changes across regressions:

they have a 10th percentile t-statistic that is negatively significant at 5%, and a 90th percentile one that is positively significant at 5%. First, this shows further subgroups, and that this subreddit is not to be regarded as an entity as a whole. Second, this is also further proof of echo chambers (Cookson, Engelberg, and Musslins (2021); Pedersen (2022)), with only individuals with similar characteristics discussing certain stocks.

1.6 Returns, Volatility & Autocorrelation

In this section, I first seek to understand how, when omitting individual heterogeneity, both posts and comments might predict returns. Following Antweiler and Frank (2004) I also look at the impact of both posts and comments on future volatility. In doing so, I run the following regression:

$$F_{i,w+1} = \beta_1 Log\ Posts_{i,w} + Controls_{i,w} + \beta_w + \beta_i + \epsilon_{i,w} \tag{1.1}$$

Where:

- $F_{i,w+1}$ is either:
  - $R_{i,w+1}$, the cumulative return of stock i during week $w+1$
  - $V_{i,w+1}$, the volatility of stock i during week $w+1$
- $Log\ Posts_{i,w}$ is the log number of posts (plus one) about stock i during week w, which will be further separated into positive and negative for the returns regressions.
- $Controls_{i,w}$ will include the log number of comments (plus one) about stock i during week w, as well as $F_{i,w}$ and retail order imbalance (Boehmer, Jones, Zhang, and Zhang (2021)).

Additionally, in order to show that different type of investors have different predictive capabilities, I also run another regression:

$$F_{i,w+1} = \sum_{Type_k} \left[\beta_{1,k} Log\ Posts_{i,w,k} + Controls_{i,w,k}\right] + \beta_w + \beta_i + \epsilon_{i,w} \tag{1.2}$$

Where:

- $Log\ Posts_{i,t,k}$ is the log number of posts (plus one) about stock i during week w from type k, which will once again be further separated into positive and negative for the returns regressions.
- $Controls_{i,t,k}$ include aforementioned controls as well as the log number of comments (plus one) about stock i during week w from type k.

In this setting, individuals are separated into terciles based on their first principal component. The first tercile represents investors that almost exclusively talk about volatile stocks, as shown in Figure 1.10: those with a negative loading on the first principal component. The third tercile represents individuals interested in firms with strong fundamentals, especially in terms of growth. An alternative specification would have been to create clusters of individuals based on the densities plotted in Figure 1.11. This is done in appendix Section A.9 of which the conclusions are similar.

The number of subscribers on WallStreetBets has substantially increased over the sample period studied, thus affecting the distribution of posts per tercile. Figure 1.13 shows the number of posts each tercile makes in any given month. For the most part, the first tercile does not exists before 2020, reflecting the large increase in the popularity of WallStreetBets over the course of the pandemic and later on, the GameStop episode[17]. This also coincides with an influx of new and inexperienced retail in-

[17]This is not to say that individuals on WallStreetBets shied away from volatility pre-2020 but rather that some in 2020 and 2021 embraced it further (Section A.4 of the appendix explains why this interpre-

vestors[18]. The third tercile has been the one posting the most historically, while the second tercile has also become more prevalent, increasing from around 100 posts per month pre-2020 to over 1,000 as of the end of the sample, December 2021.

1.6.1 Short Horizon Returns

Plotting returns against bins of sentiment for individuals in the three terciles of the first principal component yields Figure 1.14. This figure shows in an univariate fashion that all three subgroups tend to mention positively stocks that do well in the following week[19]. The performance difference is rather striking with the first tercile showing much higher returns than the average. Although it appears the second tercile would be useful as a signal in predicting stock returns, adding controls will show that they are not.

Table 1.4 shows the results of regressions 1.1 and 1.2 on returns. The first two columns show the result of regression 1.1, without dividing individuals into groups. I find a significant positive (negative) relationship between the log number of positive (negative) posts in a given week and stock returns over the following week, with number of comments having a negative relationship. This relationship holds true, even when accounting for retail order imbalance, defined as per Boehmer, Jones, Zhang, and Zhang (2021) and the returns in the previous week. The second column, in which the standard errors are clustered at the firm and month levels, shows a drastic reduction in significance, with neither effects remaining so.

The following two columns, in which I further divide posts and comments by the

tation is the correct one).

[18] As noted by Bryzgalova, Pavlova, and Sikorskaya (2022) and a survey by FINRA (Lush, Fontes, Zhu, Valdes, and Mottola (2021))

[19] Very few stocks have an overall sentiment that is negative.

quantiles of their loading on the first principal component as per regression 1.2, yields additional insights. First, individuals who almost exclusively talk about volatile stocks would appear to be able to select stocks that do perform extremely well or extremely poorly in the following week. The effects on returns of this particular group is significantly larger than the average: over four times larger for positive posts, and seven times larger for negative posts. The effect also remains for individuals in the third tercile as shown in the fourth column of Table 1.4. However, individuals who cannot be classified into either extremes - those in the second quantile - do not appear to be particularly useful as a signal. Again, the effects are reduced once clustering the standard errors at the firm and month levels but they remain significant at 95% for individuals in the third quantile, further proving their capacities beyond that of the average poster.

There is not particular reason the effect of posts should be log-linear. In fact, differentiation in terms of signal intensity would be further proof of heterogeneity. Appendix Table A.2 shows the quadratic effects with now significant effects for individuals in the first tercile, even when clustering at the firm and month levels, with once again different effects depending on the group of individual. Those differential effects are interesting because, together with the results of with Section 1.5 and echo chamber effects, they hint that previous findings in terms of usefulness of retail investors (Kaniel, Saar, and Titman (2008); Kelley and Tetlock (2013); Boehmer, Jones, Zhang, and Zhang (2021)) might be driven by a sub-sample of them, and by a sub-sample of stocks.

1.6.2 Volatility

Table 1.5 shows the results of the second regression on volatility. Similarly to the previous regression, the effects vary by groups. Overall, the number of posts predict higher volatility; result which is consistent with Foucault, Sraer, and Thesmar (2011)

who find that retail activity has a positive effect on volatility. However, the first tercile group - the group who likes to talk about already volatile stocks - shows a stronger coefficient, twice as much as the average, and nearly four times as much as individuals with the third tercile of the first principal component. This later fact is consistent with Eaton, Green, Roseman, and Wu (2022) who find that Robinhood investors, who have been shown to like volatility, or at least attention grabbing stocks (Welch (2022); Fedyk (2022)), increase return volatility. Again, this interesting because together with Section 1.5, it hints that the volatility effect of retail investor, commonly attributed to the average one, might be driven by a sub-sample of them, and by a sub-sample of stocks.

1.6.3 Longer Horizon Returns

In this section, I ask whether the returns showed in Section 1.6.1 are permanent. In doing so, I run the following regression:

$$R_{i,w+1} = \sum_{\tau} \sum_{k=1}^{3} \big[\beta_{1,k,\tau} Log\ Positive\ Posts_{i,k,\tau} + \beta_{2,k,\tau} Log\ Negative\ Posts_{i,\tau,k} + Controls_{i,\tau,k}\big] + \beta_w + \beta_i + \epsilon_{i,w} \tag{1.3}$$

Once again, the dependant variable is the returns in the following week ($R_{i,w+1}$), and the dependant variables include positive and negative posts from either the current week, the preceding four weeks (weeks 2 to 5), and 6 to 9 weeks ago (represented by the sum over τ), for the three terciles of Redditors (represented by the sum over K). Again, controls include comments, weekly return ($R_{i,w}$), retail order imbalance

(Boehmer, Jones, Zhang, and Zhang (2021)), week and firm fixed effects. Figure 1.15 shows the value for the coefficients on the number of posts, positive and negative ($\beta_{1,\tau,k}$ and $\beta_{2,\tau,k}$), for the different groups at different horizons as well as their 95% confidence intervals (non clustered standard errors).

Engelberg, Sasseville, and Williams (2012) find a spike-reversal pattern for stocks following an attention shock. I do find a similar pattern of reversal here, with the reversal being particularly significant for individuals in the first tercile. Note, however, that the reversal pattern exists for all subgroups, and not just those that are mainly interested in stocks with attention grabbing features. This extends the findings of Engelberg, Sasseville, and Williams (2012) to stocks with high retail interest.

1.6.4 Momentum and Reversal

Table 1.4 shows a strong and negative autocorrelation of stock returns during the period studied. This raises questions as to what strategies retail investors employ. In particular, I ask whether they follow momentum strategies or risk reversal ones; and if so, which type of investor use those strategies. This has, most notably, implications for market liquidity and relates to Eaton, Green, Roseman, and Wu (2022) who find that participation by Robinhood investors, whom they qualify as inexperienced, lead to lower market liquidity and higher return volatility. If those investors however follow reversal strategies - especially during turbulent times - they could provide liquidity when it is most needed, as explained by Nagel (2012) who does find short-term reversal strategies to perform especially well during periods of financial markets turmoil, a finding which I can and do replicate on this sample as shown in Section A.10 of the appendix.

One could have certainly expected retail investors to generally follow momentum

strategies, given their reputation for buying high, and selling low, thus talking positively (negativity) about a firm when its stock prices goes up (down). Figure 1.16 shows it not to be the case. This figure plots the weekly return of a firm again the log number of posts written by the three terciles of retail investors, as defined previously. It shows that all of them have a propensity to talk about stocks of which the returns have either been low or high. This is linked to the "Previous day return squared" variable showed in Figure 1.10a: individuals who load negatively on the first principal component have a propensity to talk more about firms with extreme returns; but, as this plot shows, all three types show this pattern. More to the point, they talk both positively and negatively about those firms: all three types therefore show both momentum and reversal.

1.7 Conclusion

This paper shows that retail investors are extremely diverse, even in settings in which they self-select (here, participating in discussions on Reddit, and more specifically, on WallStreetBets), and that the "retail investor" designation is but one of convenience. To show this, I use the characteristics of stocks, as of the day of which they are talked about by Redditors, to construct individual characteristics. The characteristic tilts reveal different groups of investors with, most notably, some interested in volatility while others are more interested in fundamentals. Their differences relate to the literature, in particular with Baker and Wurgler (2006), in which retail investors discuss stocks of which the valuation is highly subjective: extreme growth stocks, volatile stocks, young stocks, and distressed stocks.

This paper also highlights that several biases found in the literature are connected. The investors described by Kumar (2009) - investors who prefer stocks with lottery

like features - are the same as the one described by Welch (2022), Barber and Odean (2008), and Grinblatt and Keloharju (2001) - sensation-seekers and those interested in stocks with attention grabbing features - and, whom are also interested in news, and here in particular posting and commenting more about news posts. This is a single characteristic, with many consequences and contrary to most of the literature and similarly to Welch (2022), I do not find the bias described above to hinder the performance of their bearer - at least over the time period studied here.

I find performance to be different depending on the type of investor with both fundamental and volatility investors proving useful in a predictability exercise, with different intensities, thus showing further heterogeneity. This paper also shows that those interests are related to echo chambers (Cookson, Engelberg, and Musslins (2021); Pedersen (2022)), in which investors are subject to confirmation bias, due to only being exposed to opinions similar to their own. I show here that most stocks are being discussed by individuals of the same type. This has further implication for the literature as effects attributed to retail investors seem to be coming from a sub-sample of them but also a sub-sample of stocks. Those subgroups also matter for the results on volatility. While all Redditors increase volatility (in accordance with Foucault, Sraer, and Thesmar (2011)), Redditors interested in stocks with attention grabbing features predict higher future volatility than their counterpart.

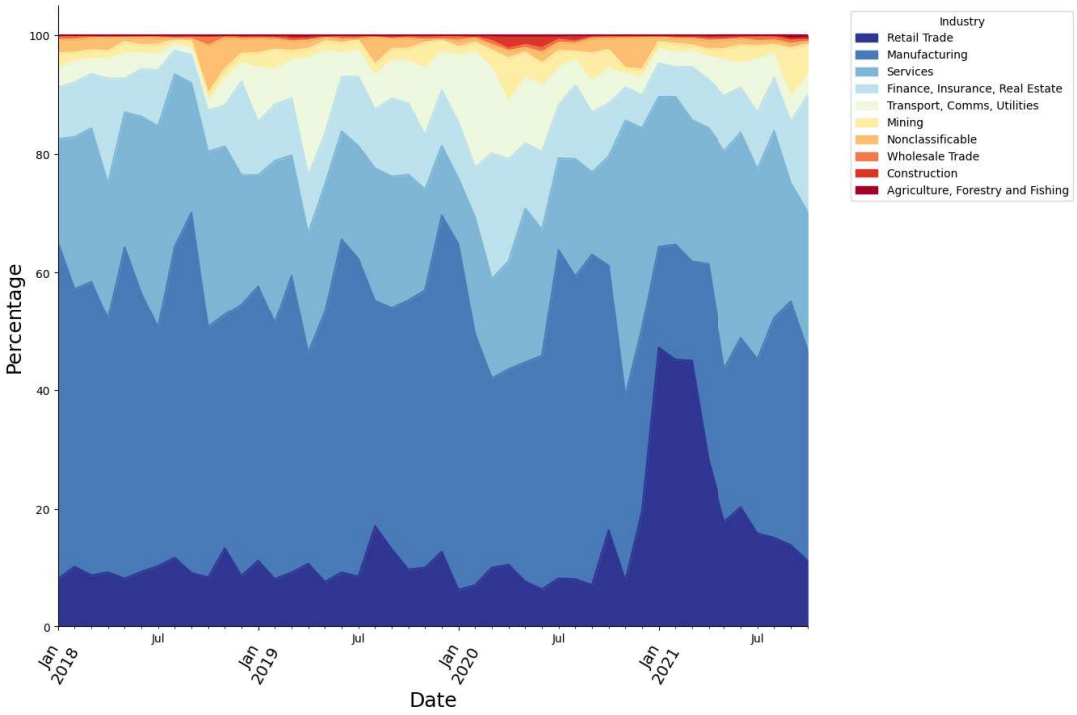


Figure 1.1. SIC Codes

This figure plots the first two digits of the SIC codes of the tickers mentioned in posts a given month. Tickers are weighted by the number of times they are mentioned.

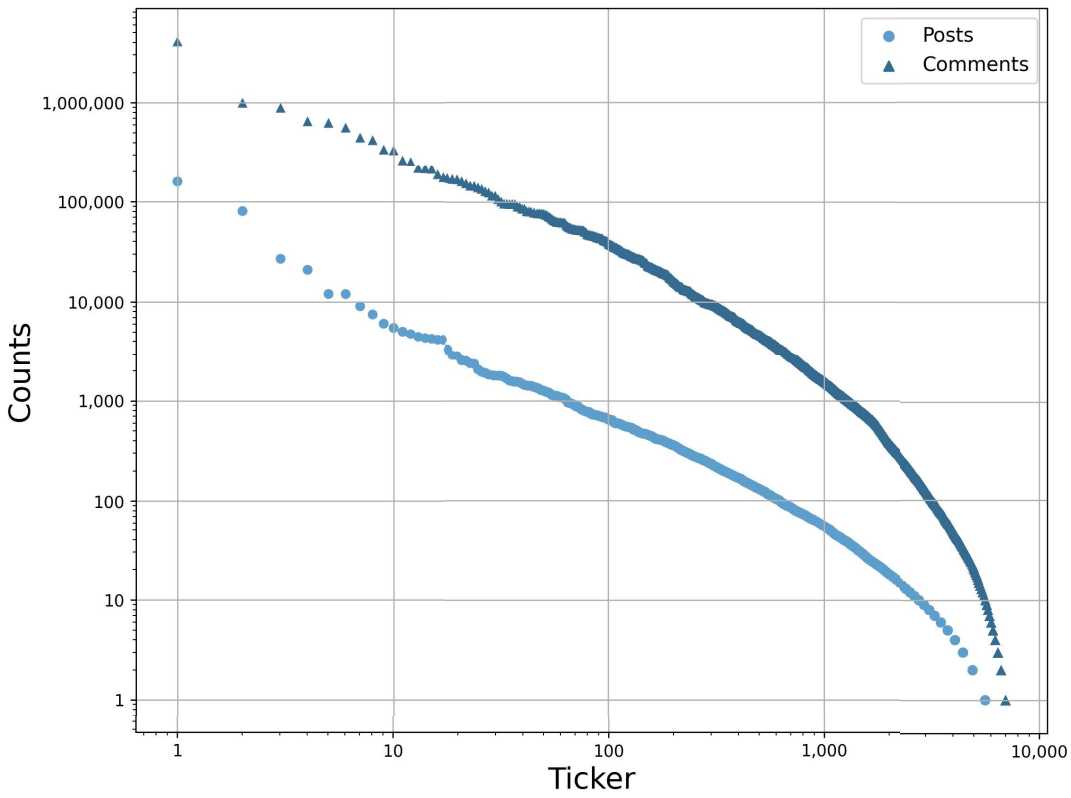


Figure 1.2. Ticker Mentions per Ticker Rank

This figure plots the popularity ranking of each ticker against the number of time each is mentioned. The first curve shows this for the posts while the second shows the same ranking for the comments. Both scales are log.

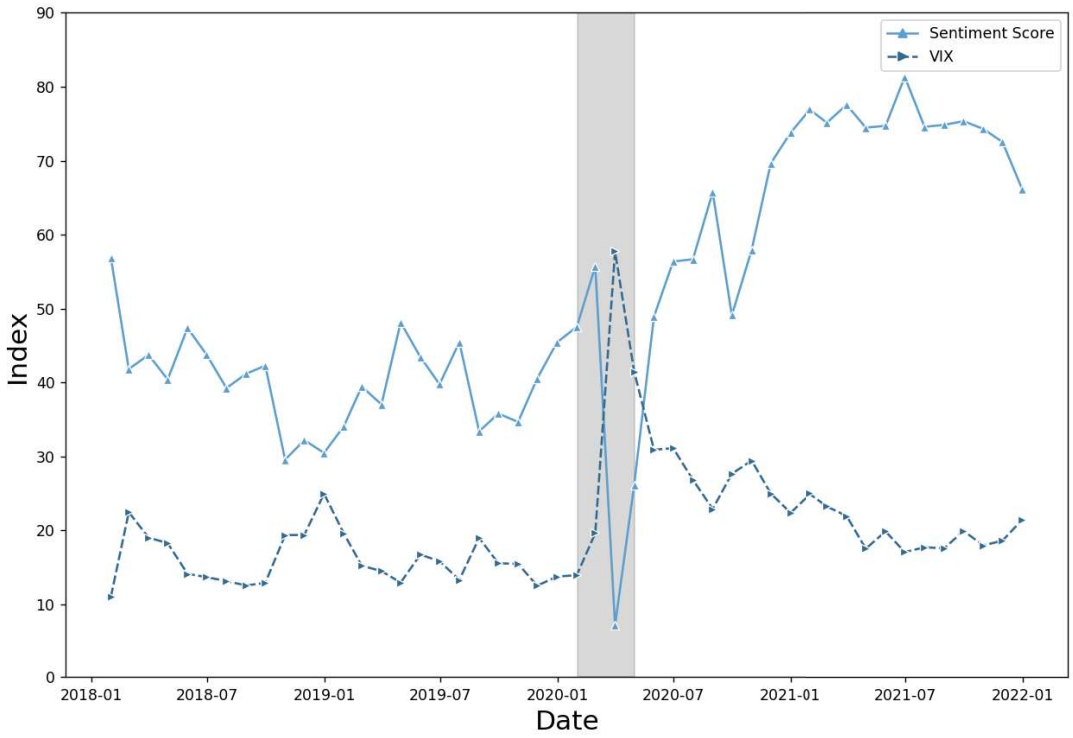


Figure 1.3. Average Sentiment and VIX

This figure plots the overall sentiment of Reddit posts with tickers against the VIX over time. A value of +1 is given to positive posts, and -1 to negative ones. Averaging the sentiment at the monthly level yields the sentiment score. The grey shading represents the 2020 NBER recession period.

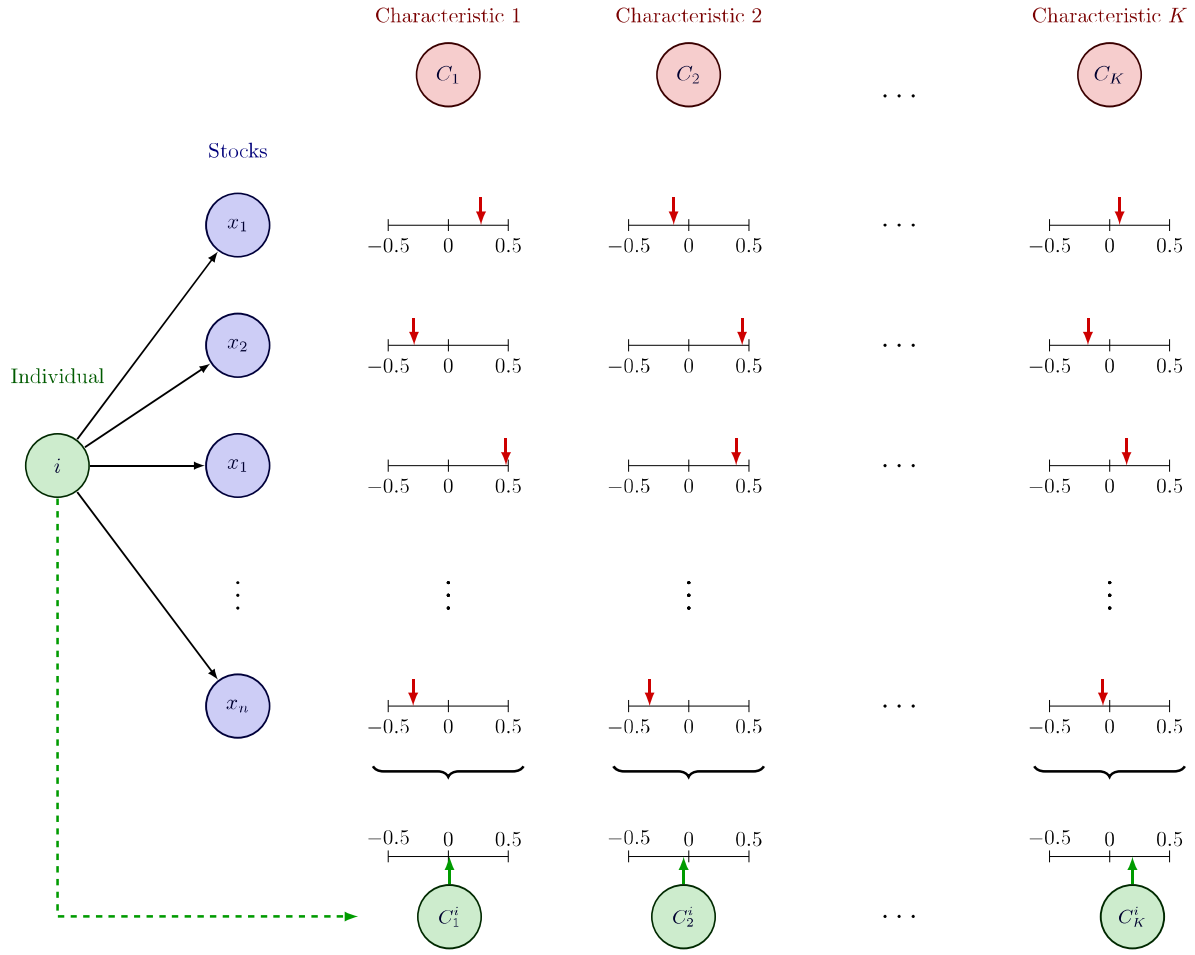


Figure 1.4. Method Summary

This graph summarizes the method by which preferences are inferred. An individual i talks about different stocks (x_n) at different dates (t), all of which have different characteristics on those dates ($C^n_{t,k}$), which in turn, when averaged, yield the characteristic tilts of an individual i (C^i_k).

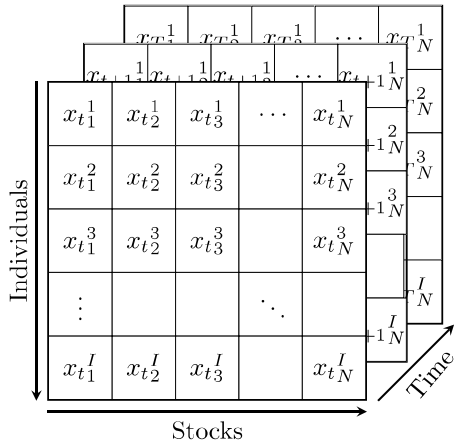


Figure 1.5. Individual Talking Matrix

This figure represents the individual talking matrix. This three-dimensional matrix has for dimensions I x N x T where I is the number of individuals, N is the number stocks, and T the dates. Each cell takes into its entry $x^i_{t,n}$, which represents the number of times individuals i talks about stock n on date t.

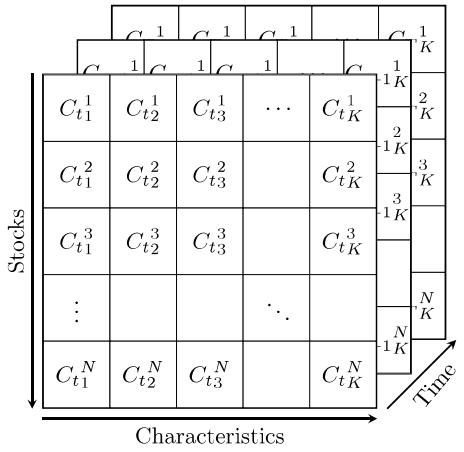


Figure 1.6. Stock Characteristic Matrix

This figure represents the stock characteristic matrix. This three-dimensional matrix has for dimensions N x K x T where N is the number stocks, K is the number characteristics, and T the dates. Each cell takes into its entry $C^n_{t,k}$, which is a relates to the rank of characteristic k of stock n on date t with respect to all other stocks on that day.

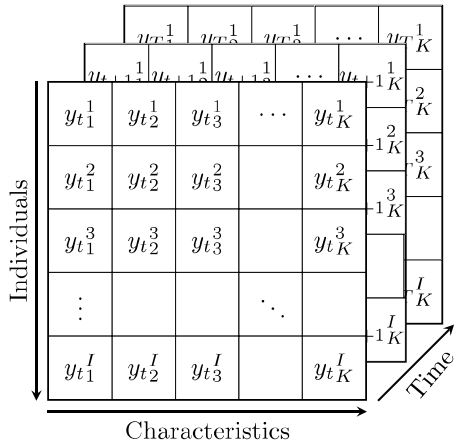


Figure 1.7. Characteristic Talking Matrix

This figure represents the characteristic talking matrix. This three-dimensional matrix has for dimensions I x K x T where I is the number of individuals, K is the number characteristics, and T the dates. Each cell takes into its entry $y^i_{t,k}$, which comes from the multiplication of the two aforementioned matrices.

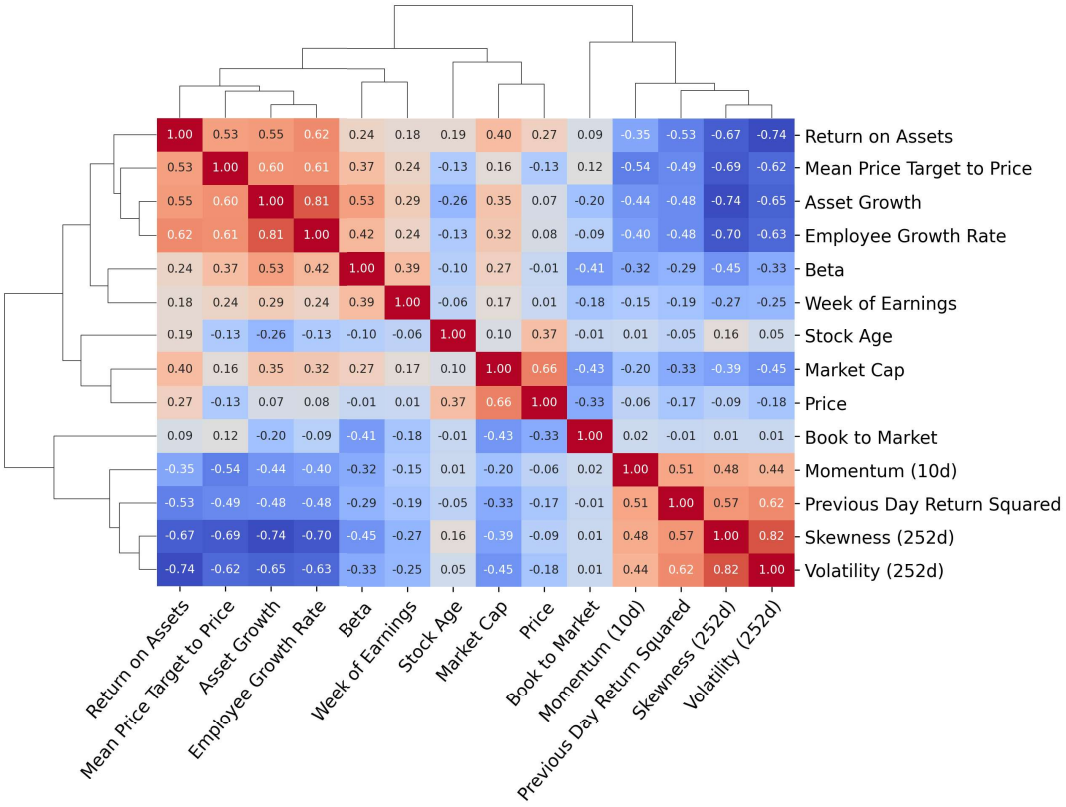
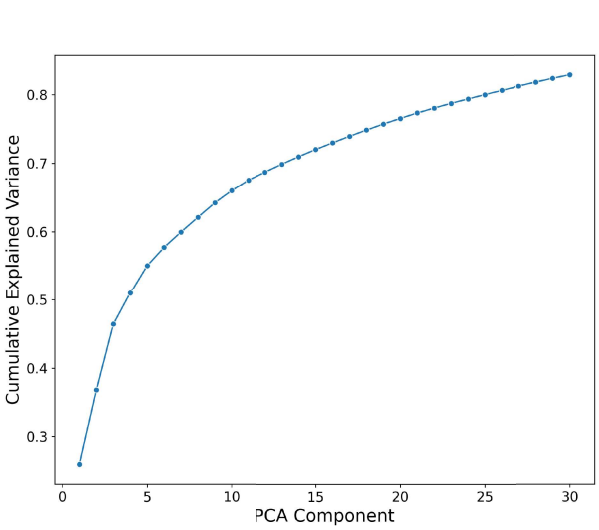


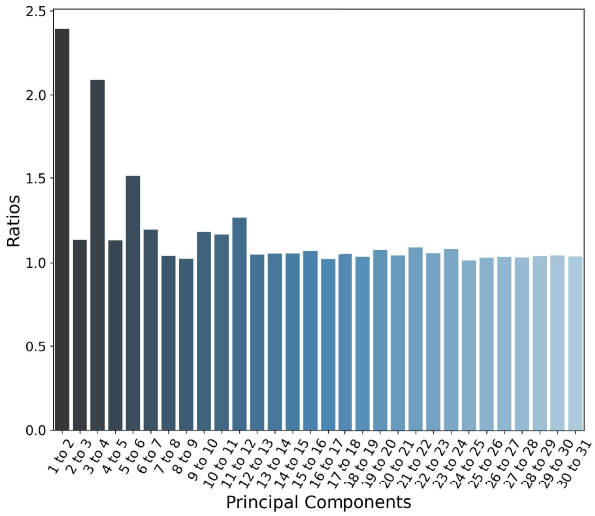
Figure 1.8. Hierarchically Clustered Co-Talking Matrix

This figure plots the characteristic co-talking matrix of selected characteristics. The heat map reflects the intensity of the correlations between the characteristics tilts across Redditors, with red (blue) signifying positive (negative) correlation.



(a) Explained Variance

This figure plots the cumulative explained variance (in the individual characteristic tilts) against the first thirty principal components.



(b) Eigenvalue Ratios

This figure plots the eigenvalue ratios of the thirty first principal components (PC), where for each PC k, $ER(k) = \frac{\tilde{\mu}_k}{\tilde{\mu}_{k+1}}$

Figure 1.9. PCA Results



(a) First Principal Component



(b) Second Principal Component



(c) Third Principal Component

Figure 1.10. Factor Loadings

Those figures plot world clouds of variable names based on their t-statistics from a univariate regression on the first three principal components (PCs). The left panels show variables with a positive loading on the PC of interest while the right panels are those with negative loadings. Both the size and colour intensity of the fonts reflect the relative importance of each variable with the most important ones appearing larger and darker.

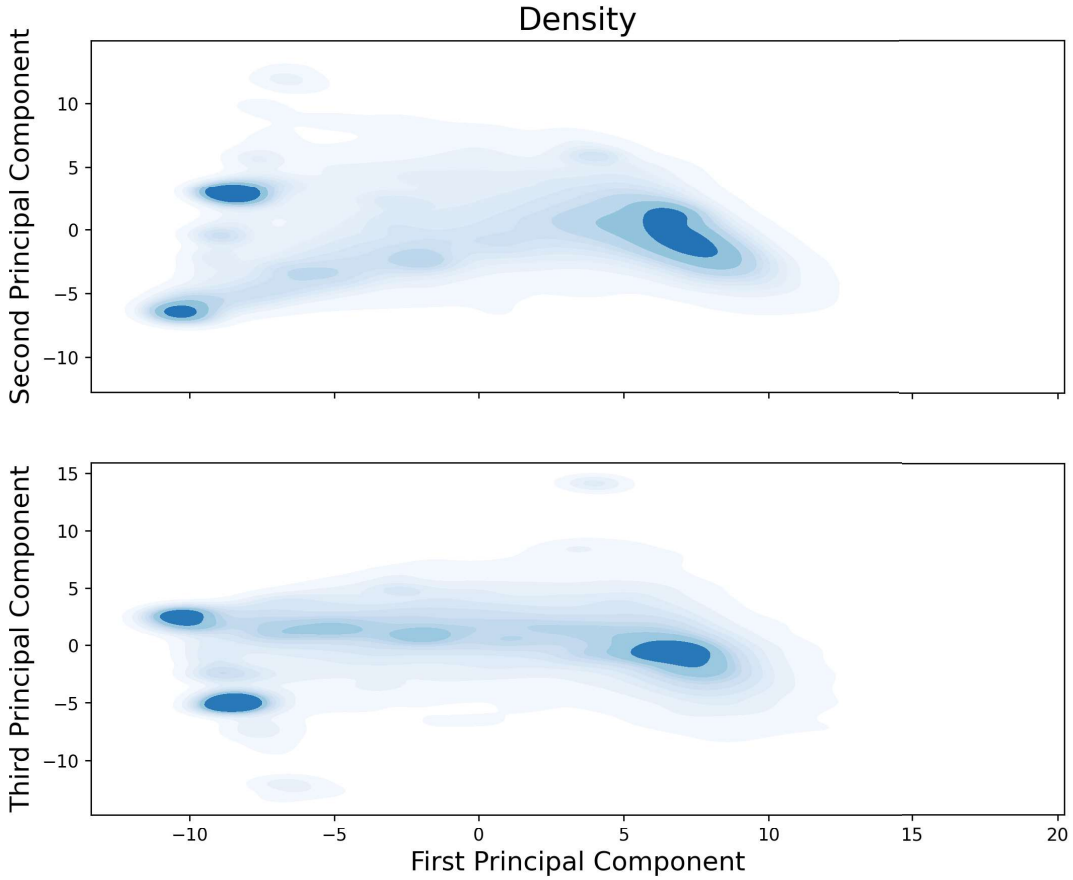


Figure 1.11. Principal Component Density

Those figures show the densities of the first principal component with respect to the second and third for all authors on WallStreetBets, with at the very least three clusters clearly visible.

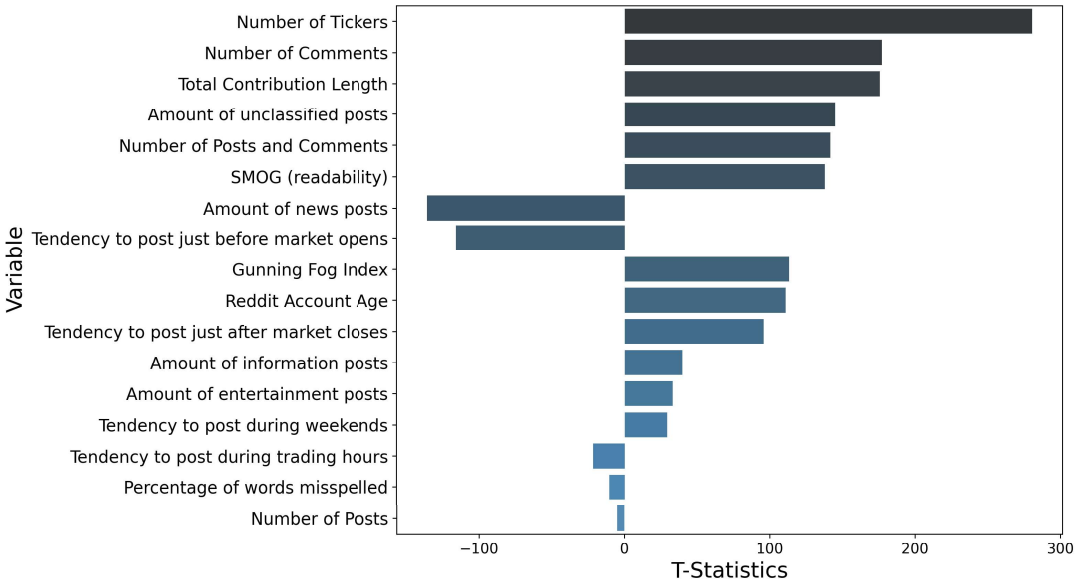


Figure 1.12. Account Characteristics Importance

This figure shows the value of the t-statistics coming from regressing the account level characteristics on the first principal component in a univariate manner. The t-statistics (in absolute value) in a descending order with the larger ones at the top. All of them are significant.

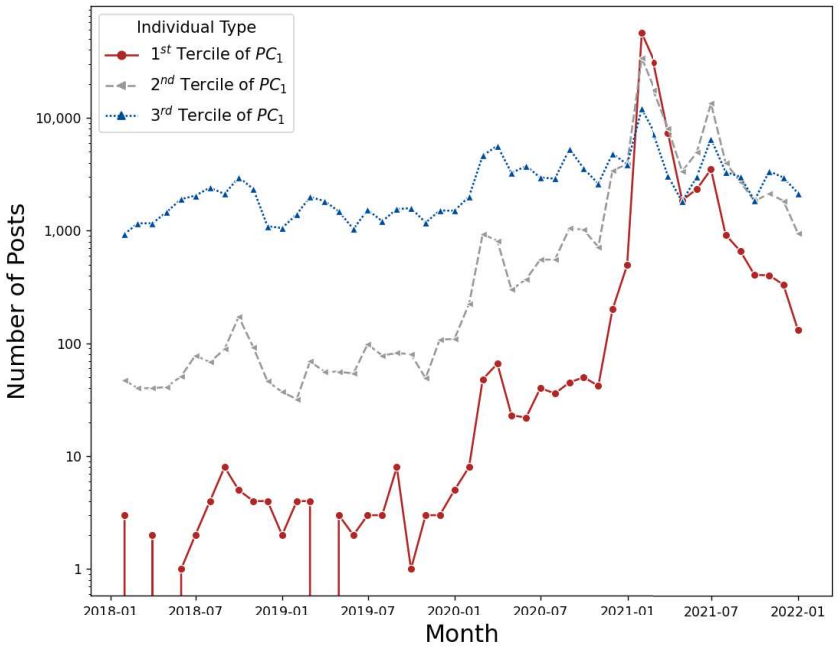


Figure 1.13. Number of Posts per PC1 Terciles

This figure shows the number of posts per month created by individuals aggregated by the terciles of their first principal component. The first tercile represents individuals interested in volatility while the third represents individuals who prefer to discuss stocks with strong fundamentals.

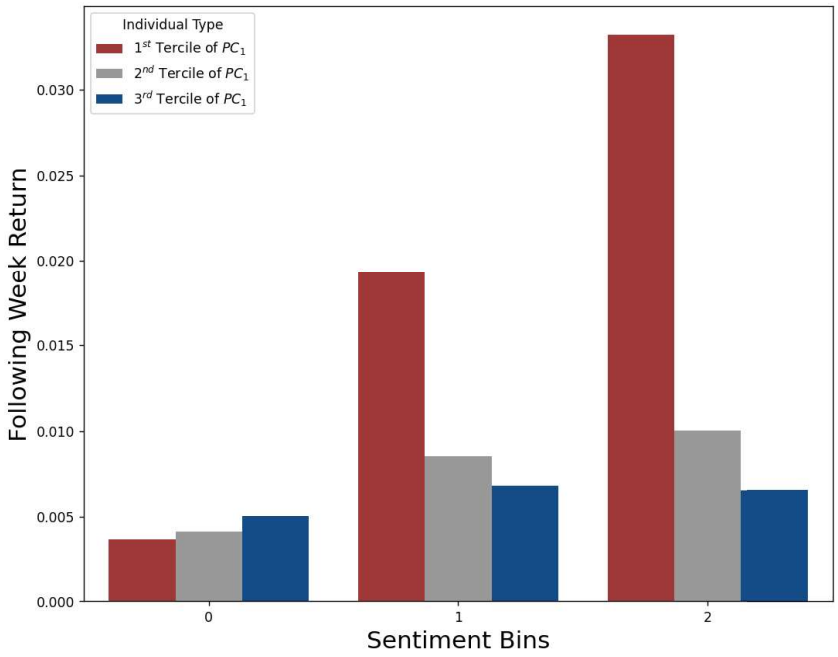


Figure 1.14. Sentiment and Returns

This figure plots the binned averaged sentiment coming from individuals in the three terciles of the first principal component against the returns of those stocks in the following week. A value of +1 is given to positive posts, and -1 to negative ones. Averaging the sentiment at the weekly level for each stock yields the sentiment score, which is then binned.

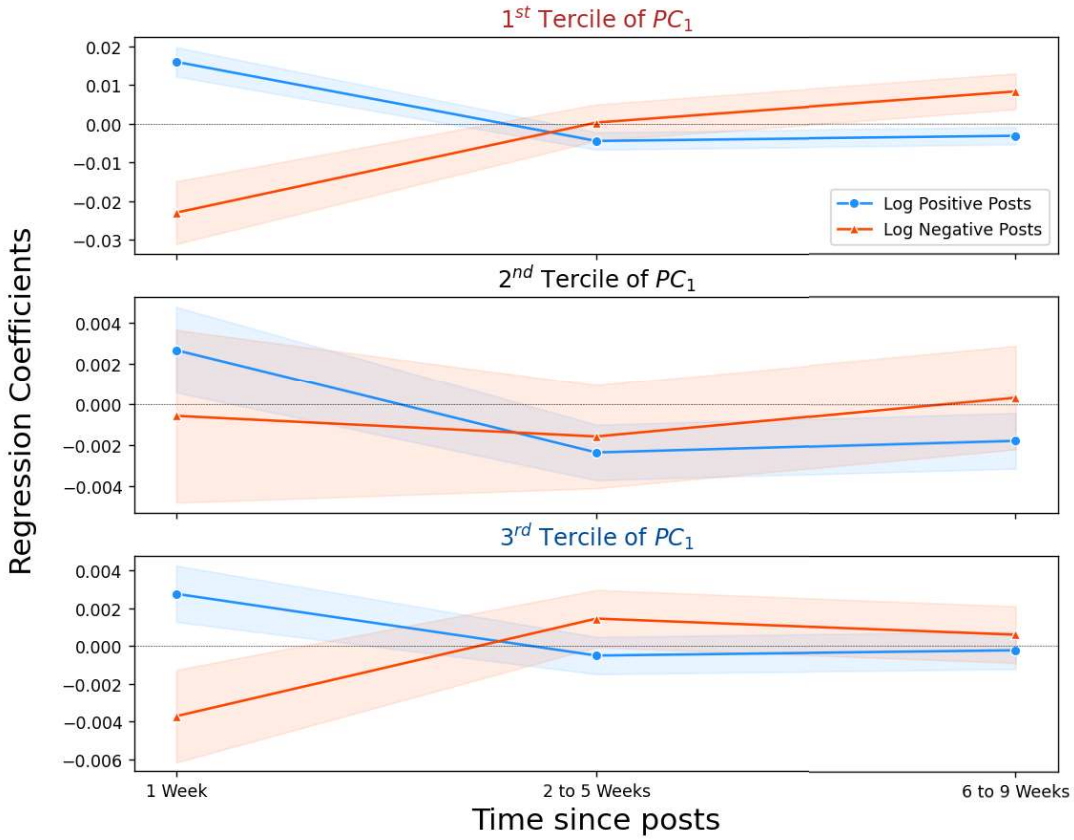


Figure 1.15. Longer Horizon Returns

This figure plots the coefficients (β) and 95% confidence intervals for the following regression: $R_{i,w+1} = \sum_\tau \sum_{Type_K} \left[\beta_{1,k,\tau} Log\ Positive\ Posts_{i,k,\tau} + \beta_{2,k,\tau} Log\ Negative\ Posts_{i,\tau,k} + Controls_{i,\tau,k}\right] + \beta_w + \beta_i + \epsilon_{i,w}$. Where the dependant variable is the returns in the following week, and the dependant variables include positive and negative posts from either the current week, the preceding four weeks (weeks 2 to 5), and 6 to 9 weeks ago (represented by the sum over τ), for the three terciles of Redditors (represented by the sum over K). Controls include comments, weekly return, retail order imbalance, week and firm fixed effects. Standard errors are not clustered.

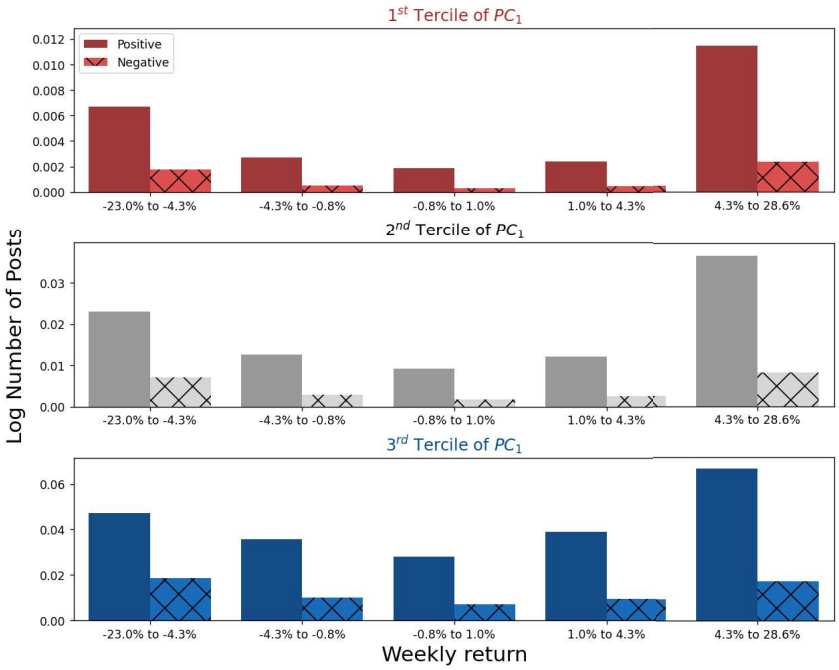


Figure 1.16. Concurrent Returns and Posts

This figure plots the binned returns of stocks against the log number of positive and negative posts in the same week, grouping individuals in the sample by their tercile of their value of the first principal component.

Table 1.1. Main Descriptives

Year	2018	2019	2020	2021
# of Subscribers	454,898.0	772,358.0	1,755,666.0	10,803,666.0
# of Posts	88,383.0	71,567.0	300,420.0	1,213,626.0
- # of non Deleted Posts	45,763.0	31,352.0	110,497.0	381,767.0
- % of non deleted Posts with ticker	34.3	27.8	30.5	32.4
# of Unique Authors	69,898.0	123,315.0	335,285.0	1,199,446.0
- Avg # of Posts per Author	1.3	0.6	0.9	1.0
- Avg # of Comments per Author	20.2	21.4	42.4	11.9
# of Comments (non deleted posts)	1,358,835.0	2,773,217.0	15,213,104.0	20,324,748.0
- Avg # of Comments (non deleted posts)	29.7	88.5	137.7	53.2
# of Tickers Mentioned	2,216.0	3,127.0	3,681.0	4,996.0
- Avg # of Mentions of Ticker	15.1	8.9	26.9	102.7

This table provides summary statistics for the subreddit WallStreetBets. It reports the number of subscribers, number posts, unique authors, comments and tickers mentioned on a yearly basis. Note that some posts are deleted soon after being posted by moderators for a variety of reasons.

Table 1.2. Posts Summary

	Count	Mean	Std	1%	25%	50%	75%	99%
% of Information Posts	1,695,445	37.0	-	-	-	-	-	-
% of Entertainment Posts	1,695,445	43.7	-	-	-	-	-	-
% of News Posts	1,695,445	7.2	-	-	-	-	-	-
# of Words	1,695,445	15.4	63.4	0.0	3.0	5.0	10.0	202.0
- # of Words if Information Post	628,141	24.7	88.4	1.0	3.0	6.0	14.0	352.0
- # of Words if Entertainment Post	741,304	7.8	21.6	0.0	2.0	4.0	7.0	69.0
- # of Words if News Post	121,724	9.5	33.2	1.0	3.0	6.0	9.0	75.0
% with tickers	1,695,445	23.4	-	-	-	-	-	-
- % one ticker	1,695,445	19.9	-	-	-	-	-	-
- % with 2 or more tickers	1,695,445	3.5	-	-	-	-	-	-
Upvote Ratio of the Post	1,695,445	0.9	0.2	0.3	1.0	1.0	1.0	1.0
Score of the Post	1,695,445	92.9	1789.1	0.0	1.0	1.0	2.0	869.0
% Positive Sentiment	351,237	69.8	-	-	-	-	-	-
% Neutral Sentiment	351,237	17.8	-	-	-	-	-	-
% Negative Sentiment	351,237	12.4	-	-	-	-	-	-

This table provides summary statistics for all posts between January 2018 and December 2021. Posts are classified into three categories (information, entertainment, and news), based on user defined categories called flairs. The upvotes and scores are determined by Redditors. The sentiment is only calculated for posts with tickers, and for which the text was available. The sentiment was gathered via one of OpenAI's GPT-3 language models.

Table 1.3. Stock Level Estimates for the Top 1,000 Stocks

(a) Panel A: Account Characteristics

	Avg. $\lvert t-stat \rvert$	10%	50%	90%	% Significant (5% level)
Single Stock Dummy	9.99	-20.42	-6.04	-0.86	81.80
No. Stocks Talked about	3.41	-2.13	0.28	8.69	43.80
Reddit Account Age	3.01	-3.44	0.01	4.44	45.30
No. Posts	1.30	-2.05	0.21	1.75	18.40
No. Comments	2.19	-5.88	-0.43	1.11	31.60
Pct. on Weekends	2.59	-3.37	-0.68	2.81	34.30
Pct. During Trading Hours	1.77	-2.59	-0.22	2.19	26.50
Pct. Entertainment	3.15	-4.48	-0.96	3.28	46.80
Pct. News	2.44	-4.07	-1.13	1.84	38.60
Gunning Fog Index	2.78	-3.75	-0.03	4.12	44.40
Spelling Errors	1.03	-0.85	0.00	2.24	12.70

(b) Panel B: Characteristic Tilts

	Avg. $\lvert t-stat \rvert$	10%	50%	90%	% Significant (5% level)
Stock Age	7.90	-9.56	0.39	11.88	60.70
Return on Assets	7.39	-9.23	0.11	10.67	63.10
Mean Price Target to Price	6.32	-4.47	1.71	11.04	62.60
Asset Growth (quarterly)	6.88	-8.28	-0.25	9.41	59.40
Book-to-Market	7.71	-9.20	0.16	12.24	63.00
Beta	4.37	-6.07	0.21	6.66	49.90
Week following Earnings	3.00	-2.36	0.34	5.27	38.50
Log Market Cap	9.24	-13.70	-1.98	10.16	72.70
Log Share Price	7.90	-9.60	1.10	11.57	66.60
Momentum (30d)	3.53	-4.92	-0.26	4.34	47.20
Skewness (252d)	7.57	-13.59	-0.89	9.31	65.70
Returns Volatility (252d)	7.76	-12.18	-0.65	9.12	65.20
Previous Day Return Squared	3.10	-5.13	-0.87	2.23	40.70
Earnings Volatility	9.17	-12.93	0.48	14.69	69.10

This table shows the average absolute value, as well as the 10%, 50% and 90% percentiles of the t-statistic coming from 1,000 stock level regressions (the 1,000 most popular stocks in terms of the number of posts that mention them) in which the dependant variable is the talking share Q_i regressing how much an individual i talks about a given stock n, in percentage terms. The regression in question, for each stock n is $Q_{i,n} = \alpha_n + \sum_{k=1}^{K} \beta_{n,k} F_{i,k} + \epsilon_{i,n}$. The last columns shows for which percentage of the regression the variable in question is significant at 5%.

Table 1.4. Return Predictions

	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$
Log Positive Posts	0.003***	0.003		
	(0.001)	(0.002)		
Log Negative Posts	-0.003*	-0.003		
	(0.001)	(0.002)		
Log Comments	-0.001***	-0.001		
	(0.000)	(0.001)		
Log Positive Posts$_{1^{st} \text{ Tercile of } PC_1}$			0.014***	0.014
			(0.002)	(0.009)
Log Negative Posts$_{1^{st} \text{ Tercile of } PC_1}$			-0.022***	-0.022
			(0.004)	(0.018)
Log Positive Posts$_{2^{nd} \text{ Tercile of } PC_1}$			0.001	0.001
			(0.001)	(0.004)
Log Negative Posts$_{2^{nd} \text{ Tercile of } PC_1}$			-0.002	-0.002
			(0.002)	(0.004)
Log Positive Posts$_{3^{rd} \text{ Tercile of } PC_1}$			0.003***	0.003*
			(0.001)	(0.001)
Log Negative Posts$_{3^{rd} \text{ Tercile of } PC_1}$			-0.003*	-0.003*
			(0.001)	(0.001)
$r_{i,w}$	-0.040***	-0.040***	-0.041***	-0.041***
	(0.001)	(0.010)	(0.001)	(0.010)
Retail Order Imbalance	0.001*	0.001*	0.001*	0.001*
	(0.000)	(0.001)	(0.000)	(0.001)
Number of Comments Controls	-	-	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Clustering	-	Firm & Month	-	Firm & Month
Obs	784099	784099	784099	784099
R-squared	0.002	0.002	0.002	0.002

Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$

The dependant variable is the return in the following week, $r_{i,w+1}$. The independent variables are the log number of positive and negative posts (to which a 1 was added) made in the current week, either fully aggregated or aggregated based on terciles of the individual's principal components. Controls include the number of comments (per type of individual), the returns in the current week, as well as retail order imbalance (Boehmer, Jones, Zhang, and Zhang (2021)). All regressions include firm and time fixed effects. The firm GameStop was removed from those regression.

Table 1.5. Volatility Predictions

	$\sigma^2_{i,w+1}$	$\sigma^2_{i,w+1}$	$\sigma^2_{i,w+1}$	$\sigma^2_{i,w+1}$
Log Posts	0.003***	0.002***		
	(0.000)	(0.000)		
Log Comments		0.001***		
		(0.000)		
Log Posts$_{1^{st}\text{ Tercile of }PC_1}$			0.003***	0.004***
			(0.000)	(0.000)
Log Posts$_{2^{nd}\text{ Tercile of }PC_1}$			0.001***	0.002***
			(0.000)	(0.000)
Log Posts$_{3^{rd}\text{ Tercile of }PC_1}$			0.003***	0.001***
			(0.000)	(0.000)
Log Comments$_{1^{st}\text{ Tercile of }PC_1}$				-0.001***
				(0.000)
Log Comments$_{2^{nd}\text{ Tercile of }PC_1}$				-0.000*
				(0.000)
Log Comments$_{3^{rd}\text{ Tercile of }PC_1}$				0.001***
				(0.000)
$\sigma^2_{i,w}$	0.129***	0.129***	0.130***	0.129***
	(0.001)	(0.001)	(0.001)	(0.001)
Time Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Obs	766376	766376	766376	766376
R-squared	0.019	0.019	0.019	0.019

Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$

The dependant variable is the volatility in the following week, $\sigma^2_{i,w+1}$. The independent variables are the log number of positive and negative posts (to which a 1 was added), as well as the number of comments made in the current week, either fully aggregated or aggregated based on terciles of the individual's principal components. Controls include the volatility in the current week. All regressions include firm and time fixed effects. The firm GameStop was removed from those regression.

Chapter 2

A Bad Bunch: Asset Value Under-Reporting in the Mumbai Real Estate Market

With Santosh Anagol, Vimal Balasubramaniam and Tarun Ramadorai

2.1 Introduction

Low-income countries collect less tax revenue than high-income countries as a fraction of GDP, and a perennial question is the extent to which this gap is attributable to differences in the under-reporting of economic activity (Gordon and Li, 2009; Kleven, Kreiner, and Saez, 2016). The problem of under-reporting is particularly vexing when attempting to use property values as a tax base either for property or transaction taxes, as properties are heterogeneous along multiple dimensions. This renders objective valuation difficult, providing strong incentives to property owners to under-report

valuations, with substantial tax revenue losses in developed and developing countries alike (Harberger, 1965; OECD, 2007). This issue is also important in light of calls to increase property taxes in discussions about progressive taxation as an antidote to rising wealth and income inequality (Stiglitz, 2015; Piketty, 2015).

A solution proposed to control such under-reporting in many countries is to create formulaic assessments of property value based on the physical location of properties, and to set the tax base as the higher of this government-assessed value and the sales price reported by the buyer. This is a commonly used method around the world that to our knowledge has not yet been carefully analyzed by economists; based on our review of transaction tax policies in the 82 largest cities in the world, 35 of these cities employ this specific system, and variants of this policy that involve some form of "model-based" appraisal of property values are even more widespread.[1]

We propose a new method to detect the under-reporting of property transaction values in such settings, and apply the method on a large and granular administrative dataset tracking the universe of residential property transactions in Mumbai, India between 2013 and 2022, uncovering an 11% under-reporting rate (95% confidence interval = [10.8%, 11.31%]) over the sample period. This is material, as property taxes are an important source of revenue for governments. Real estate transaction taxes constitute approximately 20% of state government revenues in India, and property taxes constitute over 5% of all government tax revenues in OECD countries, on average.[2] While these aggregate numbers are useful indicators of the importance of this issue, it is also worth noting that the decision of how much value to report is a high-stakes decision for individual households. Home values are typically multiples of house-

[1] Appendix Table B.1 presents the cities that employ this guidance value system along with their transaction tax rate. A detailed spreadsheet of valuation systems for the top 82 cities of the world can be found here. We briefly discuss the broader context of cross-country variation in systems of property appraisals for tax purposes later in the paper.

[2] See https://data.oecd.org/tax/tax-on-property.htm, accessed March 2022.

hold income, and reported home values serve as the basis for transaction taxes, capital gains taxes, and annual property taxes.[3] Moreover, for decades, the Indian authorities have suspected that real estate buyers under-report valuations as a way of laundering unreported income, such as cash earnings and bribes (so called "black money"), and economically massive policy interventions, such as India's 2016 demonetization, which we study using our new approach, have been motivated by the desire to reduce such black money.

The statutory requirement in the Mumbai setting is that property purchasers report the true transaction value to the government, to serve as the tax base for the transaction tax and other relevant taxes. If the buyer's reported value falls below the government assessed value (also known as the "guidance" or "circle" value), then the tax base is set equal to the government assessed value. An important mode of evasion in this system is that buyers and sellers can collude to minimize transaction tax liabilities by under-reporting the true transaction price to the government.[4]

Using the universe of all reported property transactions between 2013 and 2022 in Mumbai, we uncover prominent bunching of self-reported property transaction values at government-assessed guidance values: 10.2% of reported transaction values bunch within 1% of the guidance value, and an additional 22.3% report more than 1% below the guidance value. Interestingly, 67.5% of transactions are reported at 1% or more above the guidance value, suggesting that penalties and/or moral concerns provide substantial disincentive to simply report the minimum government value for a large part of the transacting population—26.5% of transaction tax revenues are based on

[3] In Mumbai, there is a 5% stamp tax, a 1% registration tax, and (small) property taxes are levied in certain sub-regions.

[4] The Indian tax administration (and anecdotal reports) discuss that the difference between the reported value and the true transaction price is often transferred from buyer to seller in currency notes, to avoid detection of tax evasion through the formal financial system. See Indian Department of Revenue's "White Paper on Black Money," 2012, https://dor.gov.in/sites/default/files/FinalBlackMoney.pdf.

values reported above the guidance values. In Appendix Figure B.1 we show similar bunching of reported transaction values at guidance values in data from Sao Paulo, Brazil, suggesting our methods are applicable elsewhere [5]

Standard practice in the bunching literature, as developed by Saez (2010) and surveyed in Kleven (2016), is to use an optimizing model to translate such bunching behavior at points where marginal tax rates change to infer the elasticity of under-reporting and real behavior changes in response to tax rates. Our context does feature a "kink," in the sense that the marginal tax rate below the guidance value is zero, but then discretely increases to 5% (the transaction tax rate) for the marginal rupee reported above the guidance value. Viewed through this lens, the bunching of transactions prices at guidance values appears consistent with under-reporting, and we do estimate "evasion elasticities" using this approach. The problem, however, is that such bunching could also be consistent with truthful reporting for at least two reasons. First, while it is unlikely, infrequently updated government-assessed property values could be extremely accurate and timely estimates of true underlying transactions prices. Second, buyers and sellers might perfectly anchor transactions at guidance values.[6] If true underlying market values were observable, we could easily distinguish bunching due to under-reporting versus bunching due to these alternative explanations, but the existence of under-reporting itself renders market values difficult to observe.

To address this challenge we match our administrative data on reported and guidance values to a third-party provided price dataset of new buildings developed and sold during our sample period. Multiple real estate analytics firms in India collect and sell

[5] We replicate this Sao Paulo plot based on exhibits originally shown to us by Thiago Scot as part of their working paper Rocha, Scot, and Feinmann (2023, mimeo).

[6] While we are unaware of direct evidence that market prices anchor on government assessed values Genesove and Mayer (2001) and Andersen, Badarinza, Liu, Marx, and Ramadorai (2022) show evidence that property sellers anchor on original purchase prices, and Garmaise and Moskowitz (2004) suggest that assessed values may be informative about market prices.

these data, and we expect such data exists or will emerge in many developing country cities in the coming years. These "mystery shopping" data serve as a proxy for market values, and these data are based on collecting pricing sheets and other marketing materials directly from developers, developer sales offices, and mail flyers sent by developers. Unlike the distribution of reported property values around government-assessed guidance values, we find that this third measure of market values is smooth and exhibits no bunching. The visible difference between the distributions of self-reported and proxy market values constitutes the basis of our new technique for detecting under-reporting.

While these third-party price data are purchased and used by banks, developers, and investors in the real estate space, we nevertheless expect the data to be an imperfect proxy for market values. We therefore adapt our approach to account for the possibility of measurement error in these proxy values. First, we consider the case of "classical" measurement error, in the sense that mystery-shopping prices are noisy but unbiased measures of true market prices. We show using simulations that the presence of mean-zero noise can in theory rationalize bunching in the presence of truthful reporting (assuming that guidance values perfectly aligned with a large set of market values) even though the proxy does not bunch at guidance values. The insight that allows us to adapt our approach and surmount the challenge of classical measurement error is that aggregation helps to "smooth out" the noise inherent in imperfect proxies for market values. Bunching in the case of truthful reporting and guidance values that perfectly track market prices should result in small differences between *aggregated* reported transactions values and *aggregated* (even if noisily measured) market value proxies. In contrast, bunching that arises from under-reporting produces a distribution of aggregated reported values that lies below the distribution of aggregated market value proxies by the extent of under-reporting. Second, we consider the

possibility that the proxy for market value is both a noisy and a biased measure of true market value. Here we argue that the form of measurement error that we estimate that is required to rationalize the patterns in the data is implausible, requiring a very specific and unusual pattern of discrepancies between reported values and true market values. Moreover, when we further refine our matching strategy and study reporting behavior around revisions in government guidance values, as we describe below, we find additional evidence that is strongly consistent with incentives to under-report.

Using the approach, we estimate under-reporting rates of approximately 20% for "buncher" transactions, and roughly the same rate for transactions with reported values less than guidance values. These estimated under-reporting rates decline linearly for transactions where the buyer chose to report more over the guidance value; for buyers reporting greater than 50% above the guidance values we estimate zero under-reporting. Aggregating across all transactions, we estimate an 11% under-reporting rate (95% confidence interval = [10.8%, 11.31%]). Assuming these under-reporting rates are representative for Maharashtra state, this translates into a large annual revenue loss for the state government: in 2021, a loss of US$ 475mn lower tax revenues from property transactions (or 1% of total revenue from this source).

Governments have strong incentives to set accurate guidance values. That said, it is unlikely that they are perfect measures of true underlying value, given the fairly broad geographical regions over which they are set (this is especially true in our context, given the density of physical properties in Mumbai). Moreover, these values are infrequently updated, which leads to both staleness and inaccuracy. We use this insight to uncover additional behavioral responses using our approach, studying the revision of government-assessed values (which occur in three of the seven years in our sample) across multiple neighborhoods. We find large spikes in the volume and value of registered transactions in the days and months directly before guidance values are

scheduled to change, suggesting substantial gaming behavior. We also find that market participants rush to register transactions right before guidance values change, and uncover evidence that they back-date transactions to exploit lower guidance values. These spikes translate into approximately 6%-12% higher estimated under-reporting rates in months immediately before guidance value changes. We argue these time-series patterns are unlikely to be driven by anchoring, since if bunching comes from transactors setting market prices by anchoring on the guidance value, we would expect similar amounts of bunching over time—it is unclear why such behavior would suddenly become more frequent in the month before a guidance value change. We also argue that the time-series evidence is not consistent with the hypothesis that bunching is primarily determined by guidance values perfectly tracking true market prices. The third-party data suggest that market prices have drifted upwards over time, so we would expect guidance values to be least accurate in the months right before scheduled guidance value changes, and most accurate in the month after these changes, meaning that this explanation predicts greater bunching *after* rather than *before* guidance value changes, contrary to the data.

We further study the effect of economic incentives to under-report, starting by analyzing how India's "demonetization" changed bunching and under-reporting behavior. In theory, demonetization should have made the acquisition of cash more difficult (in the short run), and raised the potential cost of accepting cash in the long run. Consistent with this, we do observe a sharp drop in bunching of reported values immediately after demonetization; quantitatively, however we find that this drop in bunching behavior does not meaningfully change the total estimated amount of under-reporting, suggesting that counterparties may employ alternative strategies to under-report (such as misinvoicing part of the consideration paid for the property as an ongoing service contract).

We also find that under-reporting rates in secondary market resale transactions are higher than those in primary-market developer sale transactions. This finding is consistent with the theory in Kleven, Kreiner, and Saez (2016), who argue that it is more difficult for large organizations to maintain the collusive agreements that underpin tax evasion, as secondary market transactions only require a single buyer and seller to maintain such an agreement. Moreover, we find that under-reporting rates are lower for higher-value transactions and vice-versa, likely in response to higher perceived detection probabilities for more visible transactions with greater absolute economic penalties for under-reporting.

Finally, for a smaller subset of reported transactions, we are able to merge administrative data on mortgage values. With the caveat in mind that the match rate allows us to match roughly a third of (possibly selected) transactions, we find the interesting pattern that reported values of transactions with low loan-to-value (LTV) ratios exhibit the greatest extent of bunching at government-assessed values, while transactions with high LTV ratios exhibit the least bunching. We also estimate the greatest under-reporting for properties with mortgages from cooperative and public-sector banks, and least for those with private and foreign banks; mortgages from banks with high overall non-performing loans are also associated with high bunching behavior. As we demonstrate in a simple model, these patterns can be rationalized by borrowers trading off higher reported values to relax financial constraints associated with mortgage credit against the greater tax bill associated with doing so; we note that assortative matching between household evasion types and bank screening technologies can also rationalize these patterns.

The remainder of this paper is organized as follows. The remainder of this section reviews related literature. Section 2.2 describes the institutional background for property valuation for tax purposes in our setting and elsewhere. Section 2.3 sets up a sim-

ple model to guide our empirical work. Section 2.4 describes our data. Section 2.5 documents our baseline results. Section 2.6 documents how measured under-reporting varies in settings and sub-samples with different economic incentives to under-report, and Section 2.7 concludes.

2.1.1 Related Literature

Our paper contributes to the large literature on tax evasion. Our approach is most similar to previous work that uses third-party derived estimates of economic activity to study evasion behavior. Fisman and Wei (2004) estimate tariff evasion on imports into China from Hong Kong by comparing reported imports in China to the more accurately measured exports from Hong Kong to China. Slemrod (2007) discusses randomized audits, conducted by tax authorities, as a method of estimating aggregate U.S. income tax evasion, and Kleven, Knudsen, Kreiner, Pedersen, and Saez (2011) review the impact of audits on evasion. Pissarides and Weber (1989) pioneered examining consumption expenditure as an indicator of true income, finding that the self-employed have higher rates of consumption relative to their reported incomes, and Braguinsky, Mityakov, and Liscovich (2014) more recently apply this method to administrative data on car ownership in Russia combined with reported earnings. Artavanis, Morse, and Tsoutsoura (2016) use bank determined credit capacity as an independent signal of true income, finding relatively greater credit limits conditional on reported income for the self-employed.

In terms of possible remedies, Pomeranz and Vila-Belda (2019) survey research with tax authorities, much of which focuses on policy interventions aimed at increasing tax revenues. To our knowledge this work has not studied real estate under-reporting, especially in contexts where agents can choose to report at or above government-

assessed values, though Casaburi and Troiano (2016) study the political economy consequences of an Italian national reform that aimed to force property owners to register their land so as to enter the tax base (an extreme form of asset value under-reporting is hiding property ownership from the government).

Our work also relates to the literature on transaction taxes, which has typically focused on advanced economies (e.g., Best and Kleven, 2018; Kopczuk and Munroe, 2015; Dachis, Duranton, and Turner, 2012), and has typically not estimated the importance of asset value under-reporting, presuming that third party reporting by mortgage lenders, real-estate agents, and other market participants eliminates the ability of buyers and sellers to under-report transaction prices. For example, Kopczuk and Munroe (2015) finds no evidence of evasion regarding a mansion transfer tax in New Jersey, and while Slemrod, Weber, and Shan (2017) finds evidence of house price manipulation to avoid a higher average transaction tax rate in Washington D.C., they note that major tax evasion or avoidance behavior is unlikely in their setting given the relatively small change in average transaction tax rates studied. Sood (2020) notes that high transaction taxes are one component of generally high land market frictions in India (and likely other developing countries), that may hinder productivity overall.

A related literature studies ongoing property taxation and finds that assessed values for property taxes often diverge from recent transaction prices in systematic ways, with important distributional consequences (Avenancio-León and Howard, 2022). In this context, we are the first to analyse under-reporting in a system of government-assessed values. A key difference in our context is the statutory obligation for homeowners to report the true market value, while in typical advanced economy property tax contexts, homeowners are required to pay tax on the government's assessed value, even if they know the assessed value is different from the market value (Amornsiripanitch, 2020, reviews this literature).

Our work is also related to a small but growing literature studying under-reporting behavior in China, though these papers do not study the effect of government guidance values. Fan, Wang, and Zhang (2022), Agarwal, Li, Qin, Wu, and Yan (2020), and Agarwal, Kuang, Wang, and Yang (2020) cleverly use data on underlying transaction prices collected by Chinese real-estate brokerages (which serve as the basis for brokerage commissions). In the Mumbai setting, similar to many cities in developing countries, such administrative data are not recorded on true underlying transaction prices. The method we develop can be applied more broadly to detect under-reporting whenever the government sees reported and guidance values and can source (as we do) measures of market value or listing prices from analytics companies or online listings portals—an example of the broader agenda to use technology to improve tax collection in developing countries (Okunogbe and Santoro, 2022).

Our study of property value under-reporting around demonetization connects us to early analyses of black money in Indian real estate, which studied the government's pre-emptive purchase provision, under which the central government tax authority was allowed to purchase any property at 15% above the reported value, creating strong incentives for accurate reporting.[7] In statute, the government was supposed to randomly select property transactions to determine whether to exercise this right, though most sources suggest the sampling was not conducted randomly. National Institute of Public Finance and Policy (1995) estimate 44.8% under-reporting using a small sample of Mumbai transactions under this policy,[8] and the same study conducts

[7]This system appears to have been proposed in the economics literature by Harberger (1965), although Taiwan had a similar, largely unsuccessful, system implemented around the same time period Chang (2012). See Posner and Weyl (2019) for other examples of such "self-assessment" based mechanisms. A challenge to these systems is that those in charge of implementing the policy may be bribed to avoid exercising the government's right on certain properties.

[8]National Institute of Public Finance and Policy (1995) does not directly report the sample size for this estimate, however Table 3.1 in that study counts 46 properties purchased in Mumbai under this program.

a survey of real estate brokers and concludes from this evidence that approximately 60% of true transaction values were under-reported (for earlier small-sample estimates see, Tandon, 1987; Gopalakrishnan, 1986). In a survey article on Indian transaction taxes, Panchapagesan (2017) notes that there have been no more recent aggregate estimates of black money in Indian real estate transactions.

While our mortgage results are based on a limited sample of transactions, they highlight the general point made in Basu (2015) that under-reporting of home values can have potentially important macro-prudential implications. This connects our work to the broader literature on housing collateral value misrepresentation during the global financial crisis (Piskorski, Seru, and Witkin, 2015; Griffin and Maturana, 2016), as well as to Montalvo, Piolatto, and Raya (2020), who estimate transaction tax evasion in Spain using data from one housing brokerage to focus on a buyer's trade-off between under-reporting to avoid transaction taxes and over-reporting to obtain greater mortgage credit. Our results on mortgage bank ownership and under-reporting also connect our work to studies of credit screening differences across public and private sector banks (La Porta, Lopez-de Silanes, and Shleifer, 2002; Mishra, Prabhala, and Rajan, 2022).

Finally, our estimate of the property value under-reporting rate is potentially useful for developing country studies on how neighborhood change, transportation infrastructure, and zoning reforms affect real estate prices, as these studies often use government assessed property values in lieu of frequently unavailable market price data (Anagol, Ferreira, and Rexer, 2021; Tsivanidis, 2019; Gechter and Tsivanidis, 2020; Harari and Wong, 2018).

2.2 Institutional Background

Our paper focuses on the valuation of property for taxation purposes. Systems for property valuation vary around the world, and can be broadly classified into two types. The first type of system is one in which taxation authorities generate "decentralized" or property-specific assessed values. Such decentralized systems can vary in the way that property-specific values are created. In some cases, assessors determine property valuations using some combination of site visits and comparable analysis to determine a valuation. In other cases, assessors determine broad features of the property (i.e., whether it is residential or commercial, square footage, and so on) which are then used as inputs into a hedonic model to determine the guidance value. For example, the Danish system of tax assessor valuation at different points in time adopted both property-specific and model-based property assessments (Andersen, Badarinza, Liu, Marx, and Ramadorai, 2022). In the United States, depending on the location and property type, local governments often conduct comparable-sale based assessments to determine the tax base.[9]

Using assessors to periodically evaluate individual properties has the benefit of producing more accurate assessments, which is particularly valuable given the inherent heterogeneity of (and therefore potential for unobserved quality in) real estate even within small regions. There are two major challenges with individual assessments, however. First, they are costly to implement, given the large number of assessments needed, and the relatively small number of qualified assessors. Second, property owners may bribe assessors to lower their assessments, as carefully documented in Khan, Khwaja, and Olken (2016).

[9]See, for example, this description of New York state property taxes, accessed February 2023.

Typically, in systems with decentralized assessment, the statutory tax base is the government's assessed value, meaning that the owner has no obligation to report the true market value. An alternative method of decentralized valuation is the so-called "self-assessment" method as proposed in Harberger (1965). These systems encourage truthful self-assessments of property values by giving the state the right to purchase the property at the property owner's self-assessed value. Chang (2012) argues that even with these incentives property owners in Taiwan greatly under-reported property values, because the probability of the state actually exercising the right was too low.

Centralized systems of property valuation for tax purposes such as the Indian context we study here are present in many jurisdictions (this includes, among others, Brazil, Colombia, Mexico, Thailand, Indonesia, Philippines, and New Zealand—see Appendix Table B.1). In such systems, the authorities assign location-specific (usually per-square-foot) valuations as a lower bound tax base for all properties in the physical location, and periodically update these valuations as market prices evolve. Such centralization is more cost-effective than property-specific assessments, and can lower the probability of captured assessors. However, these centralized valuations (henceforth "guidance values") can be inaccurate measures of market value, especially if guidance values are infrequently updated.[10]

In such systems the statutory requirement is that owners report the true market value of their property, with the tax base set as the maximum of the guidance value and the owner-reported value. The owner faces a penalty, typically a multiple of the amount of tax avoided if they report a value lower than the true market value (note that the true market value can be different from either or both of the guidance and reported values), but generally have some form of (costly) recourse available to prove that their

[10]In parts of the U.S., tax authorities apply formulaic growth rates to historical assessed values, which is a similar approach (e.g., California's Proposition 13, see http://Santa-Clara-property-assessments accessed February 2023.)

lower valuation can be justified.

Conversations with market participants, and reports from regulators in the Indian centralized tax assessment system suggest that under-reporting follows a typical pattern.[11] The buyer (usually an individual) and seller (either an individual or a real estate developer) of an apartment agree on a transaction price. If this price is higher than the guidance value, to avoid taxes, they also agree to under-report the transaction price on the registration document (a frequent choice is to report exactly the guidance value; interviews with market participants suggest auditing is virtually non-existent as long as the reported value equals or exceeds the guidance value.).[12] To prevent detection of the under-reporting by paper/digital trail, the gap between the transaction price and the reported price is paid in currency notes (part of the true transaction price can also be misinvoiced as a higher "service charge," especially in individual-developer contracts). This ensures that the bank records associated with the transaction are in agreement with the reported value on the registration, which is important, since buyers and sellers are required to report their tax-identification numbers to enable cross-verification with transactors' bank accounts. Typical methods of obtaining large sums of cash include accrued currency from business operations, a sequence of

[11] See, for example, https://timesofindia.indiatimes.com/blogs/law-street/black-money-does-the-devil-lie-in-real-estate/. A similar discussion of under-reporting real estate transactions in rural France is recounted in Mayle (2000): "This is the two-price purchase, and a typical example would work as follows: Monsieur Rivarel, a businessman in Aix, wishes to sell an old country house that he inherited. He wants a million francs. As it is not his principal residence, he will be liable for tax on the proceeds of the sale, a thought that causes him great distress. He therefore decides that the official, recorded price—the prix déclaré—will be 600,000 francs, and he will grit his teeth and pay tax on that. His consolation is that the balance of 400,000 francs will be paid in cash, under the table. This, as he will point out, is an affaire intéressante not only for him, but for the buyer, because the official fees and charges will be based on the lower, declared price. Voila! Everyone is happy."

[12] While the law states that guidance values should be formulaic, following centrally assigned "circle rates," it is possible that the tax authority manually enters a valuation at their discretion. While this practice was not mentioned in our interviews with market participants, Comptroller and Auditor General of India (2016) discusses a few large transactions where the guidance value formulas were not followed. As we have administrative data on guidance values, our method also detects under-reporting arising from such inspector discretion.

smaller withdrawals from bank accounts over time, or writing a check as a "gift" to a friend or relative in exchange for the cash. In some cases, cash funds can also be sourced or earned completely outside the tax net; this is referred to as "black money" in the Indian context.

While market participants state that under-reporting is common, the general sentiment is that market prices are well understood, especially by developers and real estate brokers. Brokers are even known to quote market prices based on the appetite for under-reporting (i.e., a lower overall price for a higher cash share). However, it is generally not an easy task (and one to our knowledge has not been undertaken) for the tax authority to acquire systematic data on market prices (by interrogating brokers and/or developers, or scraping and painstakingly matching data from online housing listings) and conduct the kinds of comparisons that we undertake in this paper.[13]

2.3 Economic Framework

A household purchases a property for market price m and then chooses the amount to report to the government r. The property has an associated guidance value which we denote by c (for "circle rate;" the nomenclature used in the Indian context), and τ is the transaction tax rate. The incentives for reporting r differ based on whether the market price m is greater than or less than c. We first discuss the case when $m \geq c$, and subsequently explore the setting when $m < c$.

When $m \geq c$, the transaction tax liability is $\tau \times \max(r, c)$, i.e., the tax base is the maximum of the reported value and the government assessed value. It is fair to say that in the empirical setting, the tax implications of under-reporting are more complex than

[13]Major listing websites include www.housing.com, www.99acres.com and www.proptiger.com.

this simple setup. One simplification in the Indian context is that annual property taxes are only a function of the government's assessed value, not the reported value, so concerns about annual property taxes should not affect household's reporting behavior at the time of transaction.[14] However, there is a potential incentive to "over-report" to reduce future capital gains taxes.[15] Ultimately the method we develop measures under-reporting as the joint effect of all of these incentives, but we focus our discussion on the transaction tax incentive because it is the most commonly cited reason for under-reporting in our empirical context. With these assumptions, and without countervailing incentives associated with penalties for under-reporting, the model predicts that all households wish to report $r = c$ to minimize their tax burden.

Now let π_1 be the perceived probability that the reported value r is verified by the tax authorities, either through a physical audit or other mechanisms, and for now assume that the probability of verification/detection is orthogonal to the reported property value, and results in the perfect discovery of the market price m. The penalty in the case of under-reporting being detected is an n multiple payment of the avoided transaction taxes, and in our empirical setting, $n = 4$, i.e. the maximum penalty is four times the original transaction tax amount.

When $m \geq c$, the buyer chooses r to minimize the expected tax burden:[16]

[14] See https://housing.com/news/guide-paying-property-tax-mumbai/ for details.

[15] A buyer that plans to sell their property soon may prefer a higher cost basis so that their ultimate capital gain is smaller. Note, however, that sellers would want a lower reported value at the time of transaction to reduce their current capital gains (or in the case of a developer, their corporate income tax bill), meaning that capital tax incentives cut in opposite directions for buyers and sellers.

[16] Technically, the transaction tax liability is $\tau \times \max(r, c)$. For simplicity, we assume here that any household with $m > c$ will always report $r \geq c$ because they are liable for the guidance value c if they report a value less than c. We can therefore replace the $\max(r, c)$ term with r. Furthermore, there is no incentive to report $r < c$ and appeal, because the government will discover when evaluating the appeal that the true m is greater than c, resulting in the original tax amount plus a penalty payment. This formulation rules out the possibility that reporting $r < c$ can be used to avoid under-reporting detection. We discuss empirical evidence on reporting of $r < c$ later in the paper.

$$\min_{r} \tau[(1-\pi_1)r + \pi_1(r + n(m-r))] \tag{2.1}$$

This optimization problem assumes risk-neutrality. If households are risk averse over expected tax and penalty payments, this would move them towards reporting higher amounts, because the variance of any tax payments will be lower if they do so. We also note that the penalty term is independent of the guidance value c, as is the case in Mumbai.

The minimization problem in equation 2.1 has two corner solutions. Consider a household starting at the true market value m and deciding whether to under-report by one rupee. The marginal benefit of under-reporting one rupee is the transaction tax saved τ, and the marginal cost is $\pi_1 n\tau$ (the expected penalty for one rupee under-reported). If $\tau\pi_1 n < \tau$ then agents will under-report all the way down to the guidance value, i.e., all transactions will bunch at the guidance value. In contrast, if $\tau\pi_1 n > \tau$ then the agent reports the full market value $r = m$, because the marginal cost of under-reporting by one rupee always exceeds the marginal benefit.

Next, let π_2 be the perceived probability of a successful appeal to the tax authority that $m < c$. In this case, the buyer's true market value is less than the guidance value c, and they can appeal that the tax burden is $\tau \times r$ rather than $\tau \times c$. Assume that t is the per-unit transaction costs of engaging in the appeals process. t captures the legal fees and costs of repeated interactions with the registrar that are required when fighting an appeal, and nests the possibility of bribe or expediting/facilitation payments that may also be incurred; from anecdotal reports, this cost burden on the buyer scales with the distance that r lies below c. In this case the buyer once again chooses r to minimize the expected tax burden:

$$\min_r \tau[(1-\pi_2)r + \pi_2(r + t(c-r))] \tag{2.2}$$

If the expected marginal cost of appeal is low enough (in practice, π_2 is small), buyers have an incentive to report $r = m$ and appeal, and if it is large, buyers simply report $r = c$.

This simple model delivers corner solutions; a more complex version of the model could allow π_1 and π_2 to vary with either r or the gap between r and c. We later pursue this in Section 2.5.3.

2.3.1 Estimating Under-reporting from Bunching

On the face of it, this simple framework suggests that we should be able to infer households' perceived π values by inspecting the distribution of r around c. However, as the model shows, accurate inference also depends on the relationship between c and the true market value m. If guidance values are perfectly aligned with market values, and reported values are completely truthful, we would still see bunching exactly at c even without any under-reporting. This is our main identification challenge, i.e., the need to distinguish between bunching arising from under-reporting and bunching because government guidance values coincide with market values.

If we had access to a transaction-level measure of m, we could directly estimate under-reporting with no need for a bunching strategy. That said, understanding how a bunching-based strategy to detect under-reporting would work with perfectly observed m is helpful as a benchmark for evaluating how the strategy would operate in the more realistic case of observing a proxy for m.

To fix ideas, let θ be the fraction of households that believe that $\pi_1 n < 1$. We refer to

such households as "under-reporters," who report c irrespective of the true transaction price m; the remaining $(1-\theta)$ "truthful-reporting" households believe that $\pi_1 n \geq 1$, and report $r = m$. Our goal is to extract θ from observed bunching behaviour.

Let $f^j(m)$ be the probability density function of market values for all households of type j, where $j \in \{\text{under-reporter,truthful-reporter}\}$. From the model, observing household $i's$ triple (r, c, m) reveals its (unobservable) type perfectly. If $r = c$ and $m > r$, then $j =$ under-reporter. If $r = m$, then $j =$ truthful reporter. θ (the share of under-reporters) is then simply the fraction of all households with $r = c$ and $m > r$. The *aggregate* amount of under-reporting is the difference between aggregate m and aggregate r for households with $r = c$, because aggregate under-reporting for truthful reporters equals zero, and all under-reporters bunch at $r = c$.

Figure 2.1a simulates the distributions of r and m around c assuming that $\theta = 40\%$, $m - c$ is $\sim \mathcal{N}(\mu = 1, \sigma^2 = 10)$, and c is $\sim \mathcal{N}(\mu = 10, \sigma^2 = 1)$. The x-axis is $\frac{r-c}{c}$ for the r distribution, and $\frac{m-c}{c}$ for the m distribution. The figure reveals substantial bunching of r around c, with 40% of households with $m \geq c$ choosing to report c. The underlying θ parameter can be backed out by inspecting how the (bunched) distribution of r around c differs from the (smoother) distribution of m around c.

As mentioned, the extent to which the government-assessed values c track market values m is a key confound. To evaluate this issue, we need an independent measure of market value m which is neither r nor c. For now, Figure 2.1a is drawn under the assumption that we have access to a measure p which is a perfect estimate of m. More realistically, p is an imperfect proxy for m, as we discuss below.

2.3.2 Household Leverage, Cash Payment, and Preferences vs. Beliefs

The model emphasizes buyers' subjective assessments of detection probabilities and under-reporting penalties as the main drivers of reporting decisions. Clearly there are other factors influencing household reporting decisions. First, banks typically set LTVs using the officially registered property value. This creates an incentive for the borrower to report more than the guidance value to obtain additional mortgage credit. The buyer then trades off the marginally higher transaction tax associated with a higher reported value against the marginal benefit of increased mortgage credit (coming from the ability to increase property size, or fund other investment/consumption). We later extend our model to account for this motivation and evaluate its predictions by conditioning bunching on whether or not a transaction has an associated mortgage.

Second, one mechanism for under-reporting emphasizes buyers' need for cash to fund the gap between r and m. There is likely heterogeneity across households' in access to such (untracked) cash. Households who do not report sources of income to the tax authority may already have low-cost access to "black money". In contrast, salaried employees at large companies (e.g., formal sector workers in say the IT sector) may have difficulty accessing large sums of cash without additional scrutiny, given the direct deposit of their income into bank accounts. On the sell-side, similar heterogeneity is likely: anecdotally, builders that make payments to workers and other suppliers in cash are more willing to accept large currency payments. In contrast, individual salaried sellers will find it more difficult to deposit/launder large cash receipts without attracting government scrutiny. In line with this motivation, we later condition bunching behavior on whether the seller is a corporate entity or an individual.

Third, households or other property sellers may also prefer to honestly report on ethical grounds, or because there are pecuniary or non-pecuniary returns to ethical reporting behavior. One implication is that the model's interpretation of the mass of under-reporters in terms of "beliefs about detection" could be replaced with a different interpretation.[17] Another potential implication is that some real estate developers condition sales on the requirement that buyers report the true market price. Anecdotal evidence suggests that large developers require honest buyer reporting because they prefer sales records to reflect the true income received from these sales, to facilitate transparent revenues for audit certification and other accounting purposes. For example, a publicly listed developer, or a developer planning an IPO will have strong incentives to have all income transparently recorded because of significant public scrutiny (see, e.g., Coffee Jr, 2002, and references therein). While such motivations might affect the interpretation of the underlying mechanisms driving under-reporting behavior, the main estimation ideas go through regardless of whether preferences or beliefs are at work. We next turn to describing and analyzing the data.

2.4 Data

2.4.1 Registrar Data

Our dataset comprises real estate transaction registration documents from the Inspector General of Registration and Controller of Stamps (IGR), Department of Revenue, Government of Maharashtra, India. For our analysis, the important information in these documents is: 1) the reported property value; 2) the government assessed value;

[17] One way to modify our setup to reflect such ethical motivations is to model a higher perceived penalty n incurred upon detection to capture the experienced costs of ethical failure or shame for truthful reporters and vice-versa.

3) the transaction tax paid; 4) the property's floor space area; 5) information about buyer and seller type (corporate or individual); 6) the transaction date; 7) the registration date. These data are publicly available from the registrar's website, and cover all available transactions over the period from 2013 through 2018. See Appendix Section B.2.1 for a complete description of this data and example documents.

We augment the IGR administrative data with data provided by Propstack Analytics, a for-profit real estate firm that uses transactions to provide pricing and ownership information via its Zapkey data platform.[18] In addition to the IGR variables described above, Propstack Analytics also provides us: 1) an indicator for whether a property was sold by the developer directly (a "primary" sale) or sold by an individual 2) the number of buyers 3) the unit number 4) the floor number of the apartment 5) the name of the real estate project associated with the transaction, and 6) the latitude and longitude of each project location in Mumbai. Propstack covers the universe of all transactions reported in the IGR admin data, thus resulting in no loss of information for our analysis. We further confirm that the overlapping data from the two sources are identical, using a 10% random sample of Propstack transactions between 2013–2018 which we find perfectly match IGR reports. Figure B.2 presents a comparison of aggregate counts of transactions and tax revenue from the IGR data, Propstack Analytics, and the aggregate numbers of transactions reported by IGR for the Mumbai Metropolitan Region—a region larger than the coverage of our sample but the closest level of aggregation for comparison.[19]

Between the IGR data and Propstack Analytics, we cover data from Mumbai and Mumbai suburban areas from 2013–2022 worth US$106.92 billion. This region is the

[18] See https://www.propstack.com/ and https://www.zapkey.com/ for details.

[19] The government guidance value is determined by multiplying the circle rate (set on a per square meter basis for a given sub-zone × year within the city) by the area of the property. Additional adjustments to the circle rate are made based on other features, such as the floor on which the property is located or whether a parking spot is included with the property.

most important metropolitan area in the state of Maharashtra, a state that generates a quarter of India's GDP. Our region of study remits approximately 30% of the state's total stamp tax revenues.[20] There is also comprehensive spatial information available for this region, which we use in our empirical analysis.

2.4.2 Propequity Data

Our independent source of price data, i.e., the proxy p, comes from Propequity, a real estate analytics firm that maintains a subscription real estate information portal for the Indian real estate market.[21] Propequity is a for-profit analytics firm that primarily earns revenue by selling access to its data products. The subscribers are real estate public and private equity investors, banks and real estate developers. The primary use case is to understand trends in local prices and quantities for new residential projects being developed.

Propequity aims to provide data on all new real estate projects in India with potential revenues over 10 million rupees (roughly US$ 200,000), with coverage varying over locations. Over the time period 2013 to 2022 this dataset includes information on approximately 11,930 real estate projects (each such project has multiple apartment buildings which in turn have multiple apartment units) from the Mumbai and Mumbai suburban regions. For each project we observe the following time-invariant characteristics: longitude, latitude, a masked developer ID, the number and format of apartment units and amenity information, date project units started being sold, date project construction was completed, luxury status, and a few other features, in addition to an estimated current sales price of the apartments in the project, which is the

[20] Region-wise stamp tax revenue sourced from https://igrmaharashtra.gov.in/dashboard_Data_ArticlewiseAndYearwise.aspx?GvData=maharashtra.

[21] https://www.propequity.in/

main variable for our purposes. The quarterly price data are reported as the price per square foot for a "base" level apartment in the building (i.e., excluding optional amenities like parking spaces, higher floor levels, etc.).

These price data are sourced through two major methods: 1) physical visits to developers' sales offices to collect pricing sheets for projects and 2) collecting developer-emailed advertisements of projects, which report prices. Developers typically market their apartments at a per-apartment price. The data provider converts this to a per square foot price using the developer's reported "carpet" area per apartment, which is the area of usable space within the apartment including interior walls, but excluding exterior walls, outdoor spaces such as balconies, and any public spaces within the building. The provider uses these data to conduct valuations for banks that make mortgages, and note that in this context they are asked to provide an estimate of the total value of the apartment, including both the reported value and any under-reporting to avoid the transaction tax. For what it is worth, the provider reports that their prices are likely to be over-estimates of transaction prices, by roughly 2-4% given developers' incentives to obtain the highest possible prices for their apartments (which may go lower following negotiations with buyers).[22]

These price data serve as our proxy p for m, the true market value of transactions in the administrative data. We match Propstack transaction level data to Propequity price data, using the name of the project in which the transaction occurs, and the location (by latitude and longitude) every quarter. Of the total of 260,614 transactions recorded in Mumbai and Mumbai suburban regions between 2013 and 2022, 60.01% or 156,645 Propstack transactions are exact project-level matches to Propequity. We match the remaining 40% of transactions to the most geographically proximate Propequity project.

[22]Propequity also reports an estimated number of units sold within the building in a given quarter. As we have administrative data on sales from the registration documents, we do not use this information in our main analysis.

Figure B.3 documents the match quality in the data. Overall, 95% of all transactions are matched to Propequity transactions within 500 meters of the latitude-longitude of the purchase transaction, and we eliminate from our analysis successful matches that are more than 1 kilometer away from the location of the associated Propstack transaction. Figure B.4 shows the spatial distribution of the transactions in our final sample for study.[23] While the overall match quality is high, we carefully investigate effects of measurement error in p in greater detail later in the paper.

Table 2.1 presents means, medians, and counts by year for our primary analysis variables. 71% of transactions are sales made by a corporate entity (typically a real estate developer), with the rest made by individuals. The average property is 76 square meters (818 square feet) in size. The average reported value r over the sample is US$ 321,770. The average government guidance value c is 19% lower, at US$ 261,860, but the average p value is higher, at US$ 367,290. Based on these simple raw averages we would estimate an under-reporting rate of 12.4%, though we approach this more rigorously below.

Appendix Figure B.5 shows a binned scatter plot of p, r and guidance values c from our main sample of 260,614 transactions. The p and r values are highly correlated with c, suggesting that guidance values c are set to match geographic variation in market prices. Reported values r are lower on average than estimated market values p, but also strongly correlated with them. As mentioned earlier, the positive relationship between p and c lends credence to the idea that government assessed values can be useful proxies for market values in developing country cities where high quality individual transaction market price data is not available (e.g., see Anagol, Ferreira, and Rexer, 2021; Tsivanidis, 2019; Gechter and Tsivanidis, 2020; Harari and Wong, 2018).

[23] As we describe later, for our mortgage analysis, we use IGR data between 2013 and 2018 augmented with registered mortgage transactions. Appendix Section B.2.1 describes sample construction in detail.

We next turn to documenting bunching and computing a measure of property value under-reporting using the data.

2.5 Baseline Results

Figure 2.2a plots the empirical distribution of $\frac{r-c}{c}$ around zero; to our knowledge, the first analysis of reporting behavior around government assessed values in this or other markets. The figure reports the number of transactions within 2% bins, with the bin around zero ranging from -1% to +1%.

The plot reveals that 14.7% of households report below the government assessed value, with an average (median) of 24.9% (14.7%) lower than the guidance value. This is further evidence of the imperfect tracking of m by c, and suggests that a non-negligible fraction of households is willing to pay verification costs to certify that $m < c$. Second, there is a clear spike in r at government assessed c.[24] The figure also shows how our proxy for market value p varies around the circle rate. The plot is consistent with under-reporting as predicted by the model, though a caveat is that we must assess the extent to which p is a noisy proxy can affect this inference.

Keeping this caveat in mind, we analyze the distribution of reported values relative to government assessed values in light of our simple model. We assume that households act as if detection/audits are random, and that all households have the same beliefs about the probability of detection. Under these assumptions, a household will report r equal to c (i.e., bunch) if reporting one more rupee in value costs them more than the expected penalty: $\tau > \tau\pi_1 n$. Given that the statutory penalty for under-reporting is four times the avoided stamp tax ($n = 4$), we have that non-bunching households

[24] Appendix Figure B.6 presents a version of Figure 2.2a with one-tenth of 1% bins, showing that the bunching of transactions at the guidance value is distinct and sharply identified.

must believe that $\pi_1 > \frac{1}{4}$. Conservatively classifying only households that report 10% above the government assessed value as non-bunchers, this reporting behavior suggests through the lens of the model that 63% of households believe the audit probability is over $\frac{1}{4}$. While there is no official data on the audit rate of transactions, anecdotal evidence suggests that the rate of detection/audit is much lower than 25%, suggesting there are likely other non-penalty incentives driving behavior.

The green line with triangles in Figure 2.2a shows how p, the Propequity proxy for market values is distributed around c. The $\frac{p-c}{c}$ distribution shows no bunching at zero, exhibiting a smooth distribution centered at roughly 30%. To the right of the bunching region, the market value distribution resembles a right-shifted version of the $\frac{r-c}{c}$ distribution. To the left of zero, the distribution of $\frac{p-c}{c}$ is close to the distribution of $\frac{r-c}{c}$, which is consistent with buyers of properties with $m < c$ truthfully reporting market value, though a small fraction of transactions has significantly lower r than c. One difference is there is more mass of p values between -.25 and 0 away from the guidance value, relative to the reported values. One explanation for this, consistent with the model, is that there are incentives to bunch at zero even for those transactions whose $m < c$. A second explanation is that measurement error smooths the p distribution around c.

2.5.1 Under-reporting Estimation in the Presence of Measurement Error in the Market Value Proxy

As discussed earlier, the patterns documented in Figure 2.2a are consistent with under-reporting, as shown in Figure 2.1a. These patterns are also potentially consistent with *no* under-reporting in the special case when p, the proxy for m, is measured with substantial error. To develop intuition, Figure 2.1 shows two (extreme) simulations from

the model. Both exhibit bunching of r at c, but generated from very different levels of under-reporting. In the "high under-reporting, without measurement error" case, $\theta = .4$ (i.e., 40% of households under-report) under the assumption that p is an error-free proxy (i.e., $p = m$); this is the baseline case discussed in Section 2.3. The excess bunching mass in this case is only due to under-reporting.

The Figure shows that this bunching pattern can also be generated in a very different "no under-reporting, with measurement error" case (i.e., $\theta = 0$). This is possible if government guidance values c *exactly equal* market values m. This could happen for two possible reasons: 1) if the tax authority sets c values to perfectly match m; and 2) if buyers and sellers perfectly anchor on c when negotiating m. Separating the two cases is difficult since we do not directly observe m; we only observe p, the proxy of transaction prices, which can be noisy. This leads to a confound since the excess bunching mass could either be generated by under-reporting, or by perfect coincidence of c and m plus an incorrectly measured counterfactual p which differs from m. This simulation is a stark presentation of this confound to emphasize the point that bunching at c alone does not perfectly identify under-reporting behavior—stark because we assume that the distribution of reported values around government assessed values ($\frac{r-c}{c}$) takes exactly the same form in both cases. Appendix B.3 describes the inputs to the simulation in greater detail.

This confound is difficult to surmount since true transaction prices are generally unobservable given incentives to under-report. This forces reliance on proxies such as the Propequity measure that we use. Hedonic models are not a solution here either, since they are generally estimated using r or some other price contaminated by reporting incentives. A new solution is therefore required, and in what follows, we propose two novel ways to correct for this potential confound.

Dealing with Classical Measurement Error: Aggregation

The first approach to distinguish the case with measurement error from genuine under-reporting involves inspecting the behavior of *aggregated* reported and market values by $\frac{r-c}{c}$ bins. Figures 2.1b and 2.1d plot simulated aggregated reported and market values for the two cases considered above (in Figures 2.1a and 2.1c respectively). In these figures, we aggregate the simulated transactions, summing all measures of their value (p, r, c, m) within $\frac{r-c}{c}$ bins of 0.02 width. In both figures, the x-axis indicates the total reported value of transactions; for example, in Figure 2.1b at zero, the blue bar is the sum of all r with for all transactions with $-.02 \leq \frac{r-c}{c} < .02$, and the green bar is the total market value of transactions in each $\frac{r-c}{c}$ bin. In Figure 2.1d, we separately plot red bars, which show the sum of all p within each bin.

In Figure 2.1b, with simulated high under-reporting, but no measurement error in p, we can clearly see under-reporting in the zero bin, where aggregated r is substantially lower than aggregated m (since there is no measurement error assumed, the sum of all m also equals aggregate p). In contrast, in the simulation in which there is no under-reporting, but substantial measurement error in p, Figure 2.1d shows that aggregated m and aggregated noisy p are (approximately) the same within each bin. The insight here is that aggregation within bins smooths out (symmetric) measurement error in p, and allows us to observe differences between r and m. This allows us to distinguish between the two cases, which are identical in the simulated count/bunching distributions plotted in Figures 2.1a and 2.1c. Put differently, truthful reporting with measurement error will exhibit no mass difference between aggregated p and r, and under-reporting will result in a mass difference between aggregated p and r in the central bin.[25] Of course, the data is likely to feature a mix of under-reporting and

[25]Wider bins can, to a point, help us to smooth measurement error even further and better identify under-reporting at the expense of moving away from the sharp point at $r = c$. Panels A and B

measurement error in the proxy p for m, and Figure 2.1f shows that in the presence of both under-reporting and measurement error, aggregation within bins smooths out measurement error and still facilitates a direct comparison of the average differences between r and m.

This approach also allows us to answer the extent of measurement error needed in our p measure to produce Figure 2.2a, i.e., the plots *without aggregation*, even with no under-reporting, but assuming that c, the government guidance value, perfectly tracks market value m. To do so, we assume that $m_i = r_i$, and simulate Propequity prices $p_i = m_i + \epsilon_i$. We find that p that are on average 15% above m, with a standard deviation of 15% can replicate Figure 2.2a even with no under-reporting. This result illustrates the challenge in estimating under-reporting based purely on inspecting bunching in r around c versus that in p, though it relies on the assumption that $c = m$.

Figure 2.2b applies the aggregation procedure to the actual data, plotting aggregated reported and Propequity values in the data by $\frac{r-c}{c}$ bins. The green triangles indicate the total amount of Propequity estimated value (p) transacted amongst transactions within a .02 $\frac{r-c}{c}$ bin, and the blue dots indicate the total value reported (r) for the same set of transactions within the $\frac{r-c}{c}$ bin. The figure is consistent with Figure 2.1f, i.e., we confirm the presence of under-reporting using the aggregation approach, and the figure reveals that the largest amount of unreported value (the differential between aggregated r and p) comes from transactions that report r exactly equal to c. It is also worth noting that in Figure 2.2b, the overall overlap of the p and r distributions is tighter than in the pure bunching plot in Figure 2.2a, with the sharpest deviation evident between the distributions at exactly c. This pattern is more closely consistent with the "two-type" model presented in Section 2.3, which predicts an atom of mass

of Appendix Figure B.7 present aggregated estimates with bin-width set to 4% and 8% respectively, confirming this in the data.

in the distribution of r at c and a smooth distribution that is closer to p otherwise (with mass shifts from both the right and left of c, but for different reasons, i.e., minimizing appeal costs from the left, and under-reporting from the right).[26]

The figure also reveals that there is a monotonic decline in the gap between m and r as r increases to the right of c. For buyers who report more than 50% above the government assessed value, there is no visible gap between aggregate reported value r and estimated market values p. Now, the simple model that we present assumes a fixed perceived probability of detection π_1. This finding suggests that π_1 may be better modelled as an increasing function of the gap between r and c, leading to more reporting closer to the market value as the the gap $r - c$ increases. Overall, when we aggregate the rupee value of under-reporting across all bins in Figure 2.2b, we estimate ₹455.1 billion (US$9.1 billion) in under-reported real estate value in our sample over the period 2013-2022.

Figure 2.2c directly shows the fraction of estimated market value under-reported by $\frac{r-c}{c}$ bin. Transactions with r up to 50% lower than c have an estimated under-reporting rate of roughly 20% on average.[27] There is a sharp discontinuity at zero and a marked change in slope as we move to transactions that report $r > c$, with estimated under-reporting dropping to zero for transactions that with r 50% above c. For properties that report 75% above c, we estimate negative under-reporting (i.e., over-reporting)

[26]For bins with $\frac{r-c}{c} < 0$ (i.e. transactions with reported values less than government assessed values) the blue circles in Figure 2.2b aggregate c. This is because if $r < c$ the tax base is effectively c, i.e., the government assessed value, since the government assesses taxes at c pending a successful appeal. Appendix Figure B.8 replaces the blue circles in Figure 2.2b with the total r within each bin, which means that to the right of zero, these figures are identical, but will differ to the left of zero. The figure shows that the total aggregated r in the bins immediately to the left of zero are lower than that implied by p, consistent with a low probability of successful appeals (i.e., π_2 is low), leading buyers in this range to bunch at c.

[27]Note that these transactions are measured as reporting the government assessed value (because when $r < c$ the tax base is assumed to be c). A caveat is that it may be that for some of these transactions, $m < c$ and buyers disputed the accuracy of the government-issued circle rate c, so the under-reporting rate at zero (i.e., $r = c$) is potentially more accurate as it is less subject to this confound.

rates, although these estimates are based on a relatively small number of transactions (the counts per bin in this region are shown in Figure 2.2a).

Dealing with Biased Measurement Error

We have this far discussed p as a potentially noisy but unbiased estimator of m. A more general measurement error model is $p = m + \nu(r, c, m) + \epsilon$. Here the $\nu(r, c, m)$ is a bias term that can itself be a function of the transaction's reported, guidance or true market values. In reality any form of this function is possible; indeed a form of this measurement error function can always be found that fits the data and rationalizes perfectly truthful reporting under the assumption that c exactly equals m. For example, the pattern of under-reporting across bins in Figure 2.2c could be explained by truthful reporting at m plus a $\nu(r, c, m)$ function with the largest measurement error for bunching transactions reporting $r = c$, moderate measurement error for transactions with $r < c$, and declining measurement error as r increases relative to c. Ultimately, the question boils down to whether a specific form of measurement error (plus the assumption that m equals c for a large number of transactions) is more or less plausible than the hypothesis that agents under-report property values.

To illustrate this issue, we assume perfectly truthful reporting $r = m$, solve for ν and then present the form that measurement error would be required to take to explain the empirical patterns we observe.[28] We ignore the classical measurement error term ϵ as this averages out when we analyze bin-aggregated data. If $r = m$ then we can solve (post-averaging) for $\nu(r, c, m)$ as $\nu = p - r$. We can also decompose the total bias in measurement error $\bar{\nu}$ by reporting bins. Let $\bar{\nu}$ be the aggregate measurement error bias, equal to the sum of ν across all transactions divided by the sum of true transaction

[28] To be more precise, the tax-base value is equal to $\max[r, c]$; under completely truthful reporting we assume that the tax-base value is the true market value, but we simply write $r = m$ here for parsimony.

values (i.e., the denominator is the sum of reported tax base values $\max[r, c]$ under the assumption of perfectly truthful reporting). Let $j \in 1, ..., J$ index $\frac{r-c}{c}$ bins (2% bins in our case), and let N_j be the number of transactions in bin j. Then:

$$\begin{aligned} \bar{v} &= \frac{\sum_{j=1}^{J} \sum_{i=1}^{N_j} (p_{ij} - r_{ij})}{\sum_{j=1}^{J} \sum_{i=1}^{N_j} r_{ij}} \\ &= \frac{1}{\sum_{j=1}^{J} \sum_{i=1}^{N_j} r_{ij}} \left(\sum_{i=1}^{N_1} (p_{i1} - r_{i1}) + \sum_{i=1}^{N_2} (p_{i2} - r_{i2}) + ... + \sum_{i=1}^{N_J} (p_{iJ} - r_{iJ}) \right) \end{aligned} \tag{2.3}$$

where again, r stands in for the tax base value $\max[r, c]$.

Each term in the sum in the numerator of equation (2.3), under the assumptions made about truthful reporting, is the contribution of biased measurement error in a given bin to the aggregate (truthfully) reported value. Appendix Figure B.9 plots these terms estimated in the data, and shows that under the hypothesis of truthful reporting, measurement error in p must be substantially positively biased for bunching transactions at $r = c$, and decline monotonically with the gap between r and c. Put differently, for measurement error to drive our results, one would need to attribute the bunching patterns that we observe in r around c to truthful reporting, and attribute a very specific form of measurement error in the p proxy that essentially mirrors the "excess bunching mass" identified in Figure 2.2b.[29]

We highlight three results from this exercise. First, the form of measurement error required for truthful reporting to explain the patterns in the data does not fit with many plausible sources of measurement error, such as one in which p is some fixed

[29] Figure B.10 plots the bin-level measurement error rate (i.e., each term on the right hand side of equation (2.3)) under the hypothesis that measurement error explains our data. Again, we see that the measurement error rate would specifically have to be highest for bunching transactions and then decline linearly for transactions where buyers self-select in reporting $r > c$.

percent higher than true m, regardless of reporting behavior (as would be the case if Propequity prices were inflated versions of true market values). Second, under the hypothesis of truthful reporting plus biased measurement error we would not see bunching of r at c unless guidance values are set extremely carefully to match market prices. (We investigate this issue more deeply below, and find substantial evidence to the contrary.) Third, if we assume all deviations between p and tax base values (i.e., $\max[r, c]$) arise from measurement error, this sets a floor of zero on the total amount of under-reporting, and if we instead assume that p is a perfect proxy and all deviations come from under-reporting, we get an upper bound on under-reporting of 10.94% (95% bootstrapped C.I. = [10.8%, 11.31%] in our data. A conservative interpretation of our results is therefore that under-reporting lies between zero and 11% in aggregate, which is substantially lower than the anecodotal and small-scale estimates in this market, and also low relative to tax-evasion estimates for self-employed workers in developed countries.

Exact Matching of Propequity Projects to Transactions

A different way to assess the importance of measurement error in p is to focus on a subset of the data that has plausibly less measurement error. Appendix Figures B.11, B.12 and B.13 present counterpart results to our analyses in Figures 2.2a, 2.2b and 2.2c estimated on a restricted dataset of the 60% of transactions where there is an exact match between building names in the Propequity data and the Registrar transactions data. To the extent that measurement error arises from imperfect matching across buildings (e.g., transactions in older buildings being matched to new luxury building launches in Propequity) we expect this sample to be less affected by measurement error. Appendix Figure B.11, the replication of Figure 2.2a, exhibits a very similar shape to the full sample version, indeed the bunching patterns appear slightly sharper

(bins just to the right of zero do not show as much increase in mass as the bins just to the right in the full sample version). The shape of estimated levels of under-reporting by bin is essentially the same (Appendix Figure B.12 versus 2.2b), but the exact match sample levels are lower in some bins.[30] Overall, the aggregate under-reporting rate in the exact match sample is 9.48% (95% bootstrapped C.I. = [9.37%,9.48%]), which is similar to the full sample estimate of 10.94% (95% bootstrapped C.I. = [10.8%, 11.31%]).

Stale Government Assessed Values

For measurement error to explain our results rather than under-reporting, c must perfectly track m across both time and space to explain visible bunching of r at c. However c is set in a geographically coarse manner, and infrequently updated. Over the sample period, real estate prices have grown substantially, with growth rates that vary considerably in different Mumbai sub-regions, making it implausible that c perfectly tracks m over time given the government's process for setting c. Moreover c is set at a relatively broad geographical level. Given the considerable spatial variation of property values, it is also implausible that c perfectly tracks m spatially. In the data, c per sq. meter values are very close or the same for all properties within each sub-zone (the average (median) subzone in the data is 686,818 (264,136) sqm).[31] This means that a single guidance value for a large region is unlikely to be an accurate reflection of the full distribution of the true value of the assets in that region at any given point in time.

Despite these issues, it might still be the case that c values are set carefully to match the first spatial moment of m. However, if this is true, the mass of transactions happening

[30] For example, in the exact sample match the estimated under-reporting rate when $\frac{r-c}{c} = 0$ is approximately 0.2, whereas it is 0.25 in the full sample.

[31] The guidance values do incorporate some adjustments for whether a building is categorized as luxury, the floor the apartment is located in, and whether a parking space is included; but these are all categorical adjustments that are essentially swamped by the price variation within locations across buildings.

at prices above (infrequently updated) c will rise or fall over time as house prices grow or shrink on average given regional and aggregate price variation. This generates a concrete prediction. If reporting is truthful, with such time-variation we would expect to find the greatest bunching of r at c immediately *after* c values are updated (i.e., when c is closest to m), and a gradual decline in bunching as m drifts away from c before c is updated again. If counterparties anchor m at government-determined c, a similar prediction obtains, as the accuracy and relevance of c might be expected to be highest immediately after it is updated. Moreover, infrequent updates in the presence of anchoring can create incentives for sellers to wait for c values to increase, as it could allow them to negotiate for substantially higher prices (in the three years we observe guidance value changes the average increase were 14.4 %, 10.5 %, and 6.98 %, see Table B.3).[32]

In contrast, if buyers and sellers under-report to evade taxes, and with the dates of c changes publicly announced in advance, we would expect that bunching would be greatest immediately *before* any increases in the government-assessed value. Put differently, in our model, if c predictably increases, there is a predictable jump in the tax burden incurred by under-reporting the transaction value $r = c$ immediately after the rise in c as opposed to immediately before the rise in c, delivering a strong incentive to under-report prior to the change in c.

In the data, Figure 2.3b documents behavior that is consistent with the under-reporting explanation, and inconsistent with either the measurement error or anchoring explanations. The figure shows that bunching mass at $r = c$ spikes immediately *prior* to

[32]Even if sellers believe demand will be lower after c increases, they should be able to obtain some of the surplus generated by transacting at lower c values in the present by waiting and transacting at higher future c values tomorrow. Such arguments depend on discount rates and demand and supply elasticities and there are possibly constellations of parameter values that can deliver greater bunching prior to c value changes under truthful reporting. Ultimately, Occam's razor suggests that such arguments are potentially less plausible than under-reporting being higher amongst buyers who backdate transaction times to take advantage of infrequent updates.

scheduled guidance value changes.[33] While this is clearly evident in the plot, it is difficult to draw strong conclusions relative to the month-by-month variation overall from pure visual inspection. Table B.2 therefore estimates a regression model in which we explain the under-reporting rate using a time-trend and month of year fixed effects, and check whether the under-reporting rate varies in months prior to assessed value changes. The table shows that under-reporting rates are indeed approximately 6 % higher (statistically significant at the 5% level) in months prior to changes in assessed values c. This is a large increase in under-reporting relative to the sample average under-reporting rate of 6%.[34] Finally, more support for the under-reporting explanation is provided by Figure B.14, which uses the observed agreement date and the registration date for each transaction in the data to check for backdating behavior. The figure shows that there is a clear pattern of backdating agreement dates just prior to guidance value changes to take advantage of lower guidance values.

High-Frequency (Daily) Reporting Behavior

Figure 2.4 studies *daily* reporting behavior around scheduled guidance value changes (as indicated by the green vertical lines in 2.3a). The dates of scheduled guidance value changes have moved around over time (i.e., they are not always on January 1st each year), which reduces concerns that these dates are scheduled for days with other policy announcements. Figure 2.4a shows a large spike in registered transactions on the day directly before the scheduled guidance value change, Figure 2.4b shows that the bunching rate for the large number of transactions registered right before the guidance value increase is approximately 10% higher, and Figure 2.4c shows the fraction

[33] The circle rates were increased in 2014, 2015 and 2016. The Maharashtra government chose not to increase the guidance values in 2017 and 2018. See Appendix Table B.3.

[34] This sample average is the simple average under-reporting rates across transactions; this differs from our aggregate under-estimating estimate of 10.94% because it does not weight transactions by their size.

of transactions registered below the guidance value which also shows a moderate increase in the days prior to the guidance value change. These results suggest that even non-bunching transactors prefer to avoid the new guidance values which are expected to increase their tax burden (this could happen, for example, if the new guidance values are above what buyers planned to report).

2.5.2 Formal Elasticity Estimation

In addition to estimates of under-reporting rates, we formally estimate the elasticity of reported value r to the transaction tax rate τ (see, e.g., Kleven (2016)). To do so, we split the sample into ten deciles based on guidance value c, and then separately estimate the elasticity for each group from bunching behavior, thus capturing heterogeneity in the elasticity across the distribution of property prices. Appendix Section B.4 describes the underlying method in detail.

Table 2.2 presents estimated elasticities and associated standard errors, and Appendix Figure B.15 shows the corresponding bunching plots. For the lowest decile of c (median c in this decile is about US$75,000), observed bunching translates to an elasticity of approximately 2.1, suggesting a strong sensitivity to transaction tax rates for low value transactions. This elasticity means that an increase in τ from 0% to 5% would decrease reported values by a little over 10%. The second through fourth deciles (properties with US$ 99$< c <$US$ 147,000) have estimated elasticities between 1.2 and 1.5; transactions with $c >$US$ 147,000 have estimated elasticities less than 1.2.

These estimated elasticities are substantially larger than the structural elasticities of taxable income for self-employed and wage earners using tax notches in Pakistan, which range from 0 to 0.28 (Kleven and Waseem, 2013).[35] Our estimates are also large

[35]We compare to Kleven and Waseem (2013)'s preferred estimates that exploit the fraction of taxpay-

relative to those estimated for medium and large sized South African small businesses, which range between 0.23–0.27 (Anagol, Lockwood, Davids, and Ramadorai, 2022); the smallest businesses in the Anagol, Lockwood, Davids, and Ramadorai (2022) sample have an estimated elasticity of 1.75, which is in the range that we find for real estate transactions. These larger estimated elasticities may reflect the relative ease of under-reporting in our context (obtaining cash for the transfer and simply writing down a lower reported value) as opposed to corporate or personal income contexts where many transactions have to be adjusted to under-report overall income (and wage earner contexts where a worker must change jobs or adjust hours worked). They may also reflect households' different views of one-off transactions taxes paid to the state as opposed to ongoing federal income taxes for which detection/audit probabilities are likely higher.

We note the following caveats with this analysis. First, we assume that the transaction tax rate is the main driver of behavioral responses; if capital gains are also important drivers of r, then our elasticity estimates will be upward biased (as some of the bunching response would come from capital gains incentives as opposed to the transaction tax rate). Second, if there are non-tax benefits of under-reporting property values (i.e., money laundering incentives) then we will over-estimate the elasticity to transaction tax rate changes. Third, we do not consider extensive margin responses here (i.e. changes in the number of transactions occurring), which could also lead to a higher total elasticity of reported values to tax rate changes.

ers that report incomes in the "dead zone" above notches to correct elasticity estimates for frictions.

2.5.3 Model Based Estimates of Audit Beliefs

The model we consider here is one where π_1, the perceived probability that the reported value r is verified by the tax authorities is such that:

$$\pi_1 = \pi_{1,0} + \hat{\pi}(\frac{m}{r} - 1) \tag{2.4}$$

To find the best parameter values for $\hat{\pi}$ that most closely matches the observed reported value distribution in the data, we undertake a grid-search for various values of $\hat{\pi} \in [0.01, 4]$ and minimize following loss function:

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}(r_k - m_k)^2} \tag{2.5}$$

where k are 2% $\frac{(r-c)}{c}$ bins from -1 to 1, and r_k is the frequency count in the observed reported value distribution and m_k the model implied reported value distribution within each bin for various parameter values of $\hat{\pi}$. As discussed in Section 2.3, for a given $\hat{\pi}$, when $r^* < c$ when $m > c$, we set $r^* = c$ given the incentive compatible lower bound at c.

Appendix Figure B.16 presents the estimated loss function for various parameter values for $\hat{\pi}$. We find that $\hat{\pi} = 1.137$ deliver the closest match to the observed reported value distribution from our model. Appendix Figure B.17 plots the best model-implied reported values for each transaction on the x-axis, and the observed reported values for each transaction on the y-axis, both on log-scale. We observe that the model captures a large amount of variation in the observed reported values in the data. A regression of $log(r) = \alpha + \eta log(r^*) + e$, delivers $\hat{\alpha} = 0.29(s.e = 0.006)$ and $\hat{\eta} = 0.92(s.e = 0.001)$ and an adjusted-R^2 of 0.72. Using these parameters, the best es-

timated model to most closely match the observed reported value distribution is presented in Appendix Figure B.18. The implied average under-reporting rate by these parameters for the $m \geq c$ case is

$$1 - \frac{r^*}{m} = 1 - \sqrt{\frac{n\hat{\pi}}{1 + n\hat{\pi}}} \tag{2.6}$$

Using $n = 4$, $\hat{\pi} = 1.137$ we find an average under-reporting rate of about 9%. Even though the objective probability of verification $\pi_{1,0} = 0$, the perceived probability of audit in our setting $\hat{\pi}(\frac{m}{r} - 1)$ is 10%[36].

Penalty-Audit Trade-off: Anecdotal evidence suggests that $\pi_{1,0} = 0$ while the law specifies $n = 4$. In reality verification by tax authorities come at a cost. In particular, large audit probabilities are not possible in economies with limited state capacity given that non-pecuniary costs associated with audits are high. As discussed in 2.3, our economic framework suggests that at an $n = 4$, tax authorities need this audit probability to be 0.25 to induce truthful reporting. However, our model suggests a way to assess where this audit probability can be set. Appendix Figure B.19 presents the efficiency frontier, i.e., the combination of values for $\pi_{1,0}$ and n where there is no under-reporting and hence the maximum tax revenue generated by authorities. Doubling n to 8 reduces the required audit probability to induce truthful reporting dramatically, although even then, this remains high. Other ways of achieving this audit probability without physical verification requires the use of third-party information systems similar to the method developed in this paper.

[36]The first order condition of Equation 2.1 with π_1 as in Equation 2.4 yields $r^* = m\sqrt{\frac{n\hat{\pi}}{1+n(\hat{\pi}-\pi_{1,0})}}$. Given that the objective probability of a verification by the tax authorities is 0, we have $\pi_{1,0} = 0$

2.5.4 Estimated Aggregate Revenue Losses Over Time

Finally, we present estimates of total lost transaction tax revenues from under-reporting. Government stamp tax revenues are directly related to the under-reporting rate: if the true value of a property is v, with an 11% under-reporting rate the reported value is $0.89v$ meaning that a stamp tax of 5% on the true transactions price effectively translates into 4.45% of this true price being received by the exchequer, i.e., a reduction of 11% in stamp tax revenues.

Using data on the aggregate tax revenues obtained from property registrations, Figure B.31 estimates the loss of revenues attributable to under-reporting. If we take our estimated under-reporting rates from Mumbai as representative for the entire state of Maharashtra, this translates into a reduction of about US$450 million in tax revenues from property transactions (1% of total revenue) for the government for the year 2018. While this back-of-the-envelope calculation imposes several assumptions and is not precise, it provides a sense of the broader economic magnitude of tax losses arising from our under-reporting estimates.

2.6 Incentives to Under-report: Demonetization, Seller Type, and Mortgage Constraints

Having developed our novel approach to detecting under-reporting, in this section, we explore the effect of economic incentives on under-reporting behavior. We begin by analyzing the effect of demonetization (which reduced cash availability in the economy) on under-reporting rates. We then consider the effect of heterogeneity in seller type on under-reporting rates, and finally, analyze the interaction between financial

constraints and under-reporting behavior.

2.6.1 Impact of India's Demonetization on Under-Reporting

We first apply our method of measuring under-reporting to evaluate a major program, with an important stated motive to eliminate unaccounted "black money" cash hoards, believed to be frequently employed in under-reported real estate transactions. On November 7, 2016 the Indian Prime Minister declared the ₹500 and ₹1,000 currency notes as no longer legal tender; these notes together comprised 86% of the nation's currency notes. Citizens would have approximately three months to deposit any of these currency units in banks; outside this window, these notes would be worthless. Banks were required to conduct audits on any deposits over ₹250,000, where the depositor was required to report the source of such cash holdings. Nearly 100% of the outstanding ₹500 and ₹1,000 were ultimately deposited in banks (see, e.g., Lahiri (2020)), suggesting that the policy did not expropriate wealth from cash hoarders. Chodorow-Reich, Gopinath, Mishra, and Narayanan (2020) estimate the policy caused a 2% decline in GDP in the fourth quarter after implementation, with dissipated impacts after that; Karmakar and Narayanan (2020) finds the largest consumption impacts for households without bank accounts.[37]

As we have described, a potentially significant illicit use of cash is to facilitate property-value under-reporting and tax evasion via the unaccounted payment of cash from buyer to seller. Indeed, eliminating this form of evasion was a primary stated goal of the policy.[38] Post-demonetization, with less cash in the economy, we might therefore

[37] Nigeria engaged in a similar demonetization exercise in October, 2022 (AFP, 2023).

[38] In February 2017 (three months after the demonetization policy was implemented) Union Finance Minister Arun Jaitley stated: "Common people have supported us graciously without asking for anything in return. I believe that demonetization was ideal. Was there even one major demonstration or unrest against demonetisation? Isn't it true that 75% of the people supported it? It is [Isn't it] true that cash as a way of life has changed today? Nobody will tell you to pay 40% cash if you buy property.

expect a drop in under-reporting behavior. In the short-run, this could occur because transactions are delayed or abandoned completely as the cost of obtaining cash rises sharply. In the longer-run, agents may worry that a demonetization-like policy could be implemented again, or that demonetization sends a strong signal of the government's intention to crack down on cash transactions. We might also therefore expect the trend of under-reporting behavior to change. On the other hand, real estate developers may also enable non-cash mechanisms to undervalue properties. For instance, real estate developers may price new apartment purchases at a lower value and simultaneously invoice buyers for services (operations, maintenance, infrastructure charges, etc.). While the purchase contract is completely under the gaze of law, service contracts do not have similar transparency, thus allowing buyers to pay for such contracts using electronic payment. This dual-invoicing mechanism, reminiscent of Fisman and Wei (2004), could allow developers to make up for any reductions in the reported value on a transaction deed through a separate electronic payment—and could explain why under-reporting continues even when cash became scarce due to demonetization.[39]

Demonetization Aggregate Effects on Under-Reporting

Figure 2.5a shows weekly transaction counts around the demonetization event; we do not observe a sharp reduction in the total number of transactions in the weeks surrounding demonetization, despite the large macroeconomic shock it occasioned. Appendix Figures B.21, B.22, and B.23 show no sharp changes in the total estimated value of transactions, the total amounts under-reported, or the estimated under-reporting rates respectively. Overall, these plots suggest that demonetization did not have a

People have started changing themselves and the long-term impact will be seen soon" (Das, 2020).

[39] Our measure p of market value at attempts to include costing for such services, so we do not expect this to be a major form of biased measurement error in our estimates of market prices. Appendix Table B.20 presents an example of a price sheet provided by our data provider showing that such service costs are included in property valuation.

major impact on aggregate real estate transactions or aggregate measures of under-reporting in Mumbai.[40] We next look deeper to see if demonetization affected the composition of real estate transactions.

Decomposing Demonetization Effects

We decompose aggregate changes around demonetization into changes within different parts of the $\frac{r-c}{c}$ distribution. We expect the greatest changes in "buncher" transactions (with $r = c$), as these had the highest estimated under-reporting rates and therefore should be most affected by the cash crunch induced by demonetization. To this end we analyze transactions falling in three sets: 1) $\frac{r-c}{c} < -.05$, 2) $-.05 \leq \frac{r-c}{c} \leq .05$ 3) $\frac{r-c}{c} > .05$.

Figure 2.5b plots the fraction of transactions whose reported value is within 5% of the transaction's guidance value. There is a clear drop evident in the fraction of these "bunching" transactions immediately following the week of demonetization; an event study model estimated on the weekly data translates this drop into a 5.2% decline in the fraction of transactions reporting just around the guidance value, which is a roughly 20% decrease given that approximately 25% of transactions bunched just prior to demonetization. Figure 2.5c shows a corresponding increase in the fraction of transactions reporting over five percent more than their guidance value. (Appendix Figure B.24 shows the weekly fraction of transactions with reported value less than 5% less than the guidance value; here there is no discrete jump, but this fraction drifts up slightly over the six months after demonetization.) Two other major economic reforms affected the housing market shortly following demonetization, namely the passing of the Real Estate Regulation Act or RERA, and introduction of the Goods and Services

[40] This evidence is in interesting contrast to Alvarez and Argente (2022), who find large negative quantity effects of government restrictions on cash usage for Uber rides.

Tax or GST. There is also a decline in the fraction of transactions bunching after the introduction of these reforms, and a corresponding increase in the fraction of transactions reporting 5% or more above the guidance value.[41] However, despite this visible redistribution of transactions from bunching to non-bunching, Appendix Figures B.22 and B.23 confirm the aggregate results that there is no major change in the weekly amounts under-reported or the overall under-reporting rate.

Using a different approach to verify these results, Figure 2.5d compares the bunching behavior of transactions in the in the 180 days pre- and post- demonetization. Consistent with the weekly aggregate plots, approximately 5% fewer transactions bunch within 1% of the guidance post-demonetization relative to pre-demonetization. Figure 2.5e shows the estimated under-reporting rates pre- and post- when we aggregate all transactions within each bin. While both pre- and post-demonetization series show that bunching transactions have the highest estimated under-reporting rates, there is very little change in these estimated under-reporting rates by bin, which translates into very little change in aggregate under-reporting rates.[42] Overall, it seems that while demonetization seemingly deterred bunching transactions, these account for only a fraction of overall transactions; moreover, the bunching transactions that continued after demonetization had similar under-reporting rates to the bunching transactions prior to demonetization. This explains why we do not observe a significant change in overall under-reporting rates after demonetization. Figure 2.5f reconfirms this inference of small changes pre-and post-demonetization by showing the aggregated under-reporting amount in millions of USD within each bin pre- and post-demonetization periods. Indeed, this plot shows a slightly larger amount of under-reporting coming

[41] Appendix B.5 discusses these two major reforms during our sample period.

[42] Aggregating over all transactions, we estimate 254 million dollars of under-reporting in the 180 days prior to demonetization and 355 million dollars in under-reporting in the 180 days after demonetization; naturally this time-series difference can be explained by many factors, but overall it seems unlikely that demonetization itself reduced under-reporting amounts in a major way.

from buncher transactions after demonetization versus before, despite the fact that demonetization did reduce the number of buncher transactions overall. This results because while the number of bunching transactions post-demonetization is smaller, the average estimated market value of these transactions and their associated under-reporting rates are slightly larger, so the total amount under-reporting ends up similar.[43] Furthermore, the under-reporting rate in non-buncher bins is similar across the pre-demonetization and post-demonetization periods, which explains why we do not see an aggregate change in under-reporting in the weekly time-series figure (Appendix Figure B.23).

To further quantify how the change in bunching changes the overall under-reporting rate, we can use a similar procedure to that outlined in equation (2.3), but now under the assumption that the proxy p is measured without error. This allows us to decompose aggregate under-reporting into bin-specific contributions both pre- and post-demonetization. Appendix Figure B.25 plots each of these terms comparing pre-demonetization and post-demonetization amounts. The aggregate pre-demonetization under-reporting rate is 0.098 (95% CI = [0.090,0.120]), while the aggregate post demonetization under-reporting rate is 0.110 (95% CI = [0.107,0.128]). The figure shows that the contribution to the total under-reporting rate from the $r = c$ bin is larger in the post-demonetization period (consistent with the fact that there were fewer bunching transactions), however this difference is swamped by the bin by bin variation across the rest of the distribution. Overall, the results suggest that demonetization reduced the number of high under-reporting bunching transactions, but the effect on aggre-

[43]The total amount of under-reporting in the 180 days before demonetization amongst buncher transactions is 134 million USD, which equals 3,321 transactions times an average estimated market value of 0.191 million USD times a bin-level under-reporting rate of 0.211. The total amount of under-reporting in the 180 days after demonetization amongst buncher transactions is 140.7 million USD, which equals 2,593 transactions times an average estimated market value of 0.223 million USD times a bin-level under-reporting rate of 0.243.

gate under-reporting in this market was likely small. We caveat these results noting that our results are specific to the Mumbai and Mumbai suburban districts, and that demonetization results in other locations may have been different.

2.6.2 Seller Type and Under-reporting Behavior

Developer vs. Resales

The data allow us to classify transactions into primary sales made by developers, and those that occur in the secondary market, i.e., resale transactions. Given that firms interact with many buyers, there is greater risk that one buyer may whistle-blow under-reporting behavior leading to an audit or detection (Kleven, Kreiner, and Saez (2016) argues that under-reporting of wages is lower amongst large firms because of the greater risk of a disgruntled employee revealing the under-reporting collusion).

Figures 2.6a and 2.6b show the count of developer sales and resales by $\frac{r-c}{c}$ bin respectively. The bunching amongst developer sales is "sharper," in the sense that the increase in transactions exactly in the bunching bin is high relative to nearby bins. Bunching amongst resale transactions is also apparent, but counts decline more smoothly as we move away from the bunching bin. Figures 2.6c and 2.6d show the bin aggregated reported and market values for developer sales and resales. Figure B.27e shows bunching behavior as a fraction of total transactions, so the developer and resale curves are comparable. There is more bunching for resale than for developer driven transactions, consistent with Kleven, Kreiner, and Saez (2016). Figure 2.6f shows under-reporting rates by $\frac{r-c}{c}$ bin for developer sales and resellers. Both types of sales show substantial bunching at $r = c$, and a similar pattern of under-reporting by $\frac{r-c}{c}$ bin. While our overall under-reporting rate is not solely driven by resale trans-

actions, we do see more bunching and under-reporting for resale transactions.

Heterogeneity by Transaction Amount

Are under-reporting rates higher for larger value transactions that may be more visible to the tax authorities? Figure B.27 presents corresponding figures on densities, bin-specific aggregates of reported versus market values, and under-reporting rates. We find that smaller sized transactions show a more substantial jump in the under-reporting rate when $r = c$, consistent with the idea that larger value transactions are harder to under-report. These results are consistent with our findings in Appendix Figure B.28, where we see higher reporting elasticities for low versus high valued transactions.

2.6.3 Financial Constraints, Mortgages, and Under-reporting

Nearly 60% of all transactions in the Mumbai sample report values extremely close to our proxy p for market value. In the simple model developed in Section 2.3, we interpret this as non-bunching households believing that the probability of audit π_1 is larger than 25% (given that $n = 4$). This number appears high relative to anecdotal accounts of the rate of audit by the registrar, or indeed, the income tax authority. Of course, there are likely other important incentives that drive truthful reporting; one particularly important non-penalty driven incentive is the desire to alleviate financial constraints when borrowing to fund a house purchase.

Mortgage lending policies in India (and indeed in many other jurisdictions) often take the official reported value of a property into account when undertaking credit-screening, thus linking the decision to under-report with the extent of financial constraints. More specifically, in our context, many banks in India, including the largest

national lender, the State Bank of India, are only prepared to lend up to the reported value r. One important incentive to truthfully report, therefore, is generated by the desire to unlock greater mortgage financing. The resulting prediction is that we are likely to observe less under-reporting for borrowers that are financially constrained and require mortgage financing, and thus a reduction in the extent of observed bunching if our proxy is an accurate reflection of under-reporting behavior. In support of this idea, Appendix Figure B.29 presents the relationship between the percentage of mortgages in each $(r-c)/c$ bin in Figure 2.2 and the under-reporting rate within each of those bins. The figure shows a strong negative relationship between the incidence of mortgage-based transactions and the under-reporting rate, with a slope of $-0.018(s.e.=0.005)$ and an r-squared of 22%.

To pursue this intuition more formally, we extend our basic model framework to incorporate a penalty that increases with the tightness of the mortgage constraint, as in Andersen, Badarinza, Liu, Marx, and Ramadorai (2022). Consider a bank that is willing to lend $(1-\gamma)r$, where γ is the down-payment constraint. Now consider a potential buyer of a property, who requires funding of $(m-d)$ where d is their available liquidity and m is the market value of the property. The shortfall $[(m-d)-(1-\gamma)r]$ can be overcome at a cost, which for simplicity, we model as quadratic in the extent of the shortfall.[44] In this simple setup, borrowers must report more (or scale down the desired house size) in order to alleviate the financial constraint, which scales with the size of the financial shortfall.

In the case where $m > c$, the borrower (potential buyer) chooses r to maximize the

[44]The down-payment constraint affects transactions irrespective of whether true $m > c$ or true $m \leq c$. Andersen, Badarinza, Liu, Marx, and Ramadorai (2022) find that the quadratic functional form offers a good approximation to the penalty function characterizing the population of house-sellers in Denmark.

following problem:

$$U = -\tau \left[(1-\pi_1)r + \pi_1 \left(r + n(m-r) \right) \right] - \frac{\mu}{2} \left[(m-d) - (1-\gamma)r \right]^2 \tag{2.7}$$

Here, the parameter μ determines the tightness of the financial constraint. Setting $\frac{dU}{dr} = 0$, and solving for r^*:

$$r^* = \frac{(m-d)}{(1-\gamma)} - \frac{\tau(1-n\pi_1)}{\mu(1-\gamma)^2} \tag{2.8}$$

Scaling both sides by m, we obtain:

$$\frac{r^*}{m} = \left(\frac{1}{1-\gamma} \right) \left(1 - \frac{d}{m} \right) - \frac{\tau(1-n\pi_1)}{m\mu(1-\gamma)^2} \tag{2.9}$$

In equation (2.9), $1 - \frac{d}{m}$ is the True LTV ($TLTV$) on the mortgage. This relationship suggests that the incentives to alleviate the down-payment constraint can enable potential buyers to report r more truthfully, thus presenting a countervailing force to the tax incentives at play.

To test the predictions of this simple model, we require data on mortgages. In our setting, mortgages also have to be reported to the registrar, but to make progress, we must match transactions and mortgages, which are separately reported to the registrar. Using the administrative data, we undertake this matching exercise, and we are able to match roughly 31,000 reported transactions to a mortgage (we describe this matching process in detail in Appendix Section B.6, but note here that there are 78,414 mortgages which are not matched to transactions, which means there is likely substantial measurement error and potentially selection in the mortgage-matched sample). We assume that any transaction that is unmatched to a mortgage transaction has

a loan to value ratio of zero, since this biases us against finding any differences between zero LTV and positive LTV transaction samples (this mechanically makes the zero LTV sample more similar to the positive LTV sample, because some of the zero LTV sample likely have a positive but unmeasured LTV).

Ideally, we can test the relationship in equation (2.9), empirically estimating $\frac{d}{m}$ can be estimated using the proxy p for m. However, we are only able to match 8,913 of the 31,000 mortgage-matched transactions to a Propequity value p; in this small subsample of transactions, we do find an upward sloping relationship between reporting ($\frac{R}{P}$) and $TLTV$ in Appendix Figure B.30 as predicted by the model (the estimated slope is 0.65, and the regression has an r-squared of 42%). To more fully utilize the matched data, we take a different approach, plotting how bunching behavior varies with loan-to-value ratios for the full set of ≈31,000 matched mortgage transactions.

Figure 2.7a shows that transactions with progressively higher loan-to-value ratios tend to exhibit less bunching and that this relationship is monotonic. This finding is consistent with the model's prediction that incentives to relax credit constraints (since higher reported values lead to the possibility of greater mortgage loans) cut against incentives for tax evasion. The magnitudes are sizeable—low loan-to-value loans are approximately 10 percentage points more likely to bunch than transactions associated with a high loan-to-value mortgages. We also note that this result could also be driven by a negative correlation between preferences for tax evasion and credit constraints.

In Figure 2.7b, we split the matched mortgage-registered transaction sample based on the organizational structure of the lending bank (we observe the identity of the bank in the administrative data). We find that transactions associated with mortgages from cooperative banks demonstrate the greatest bunching, followed by banks which we were unable to perfectly classify, followed by public sector banks, and finally, the

lowest levels of bunching are observed in private and foreign banks. This heterogeneity can be attributed to both self-sorting of different types of borrowers to different types of banks, as well as by borrowers under-reporting more or less depending on the lending bank's credit-screening policies. Sorting of borrowers across banks could be driven by borrowers choosing different types of banks, or by banks having different lending rules which ex-post lead to selection in the type of borrowers at different banks. Conditional on a borrower matching with a bank, the bank may also have different rules which encourage different reporting amounts for the same borrower. For example, some banks will only lend up to the reported value on the sales deed, while others will lend based on their own assessment. Overall, these results are consistent with bank culture being correlated with borrower type. For example, Mishra, Prabhala, and Rajan (2022) show that public-sector Indian banks appear to have laxer credit screening standards and slower technology adoption than private sector Indian banks, and link this finding to differences in organization culture.

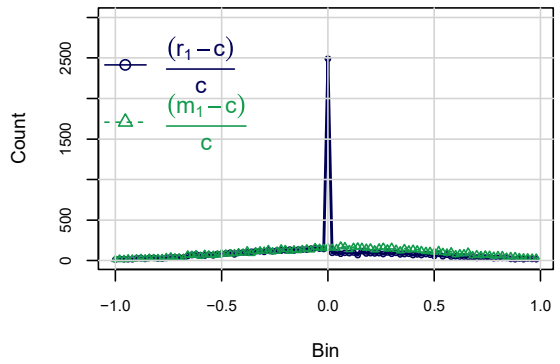
Finally, Figure 2.7c presents bunching behavior for the same group of 31,000-odd transactions with mortgages, but splits the sample by the average non-performing loan (NPL) rate (over the period 2013-2018) of the lending bank to focus more closely on a possible credit-screening channel. The figure shows that loans issued by banks with the highest NPL rates also exhibit the greatest amount of bunching. This correlation could be generated by borrowers that are more likely to default also being the types who under-report, other types of selection correlated with banks' differential lending process, or borrower selection into banks that have more lax screening of borrowers and collateral.

2.7 Conclusion

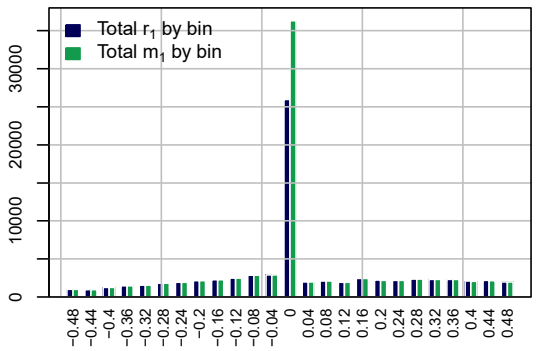
We develop a new method to detect under-reporting of asset values for tax purposes, and apply it to a large administrative dataset of reported property values for transactions tax purposes in Mumbai, India between 2013 and 2022. Our method reveals an overall 11% under-reporting rate over the period. Using monthly under-reporting rates, our back-of-the envelope calculation in 2021 is that annual losses to the state government are on the order of 1% of total revenues from lost stamp duties on property value under-reporting. These estimates are meaningfully large for the public exchequer, but are nonetheless substantially lower than prior estimates using smaller data samples, and lower than anecdotal priors related to the purported prevalence of tax evasion and "black money" in emerging market real-estate transactions. Over and above these empirical results, we view the methods that we develop as more widely applicable to a range of contexts in which the tax basis is set on the basis of government-assessed asset values.

We find that under-reporting rates are significantly higher immediately before pre-announced changes in government-assessed values, which add credence to calls to index such policy instruments, frequently-used in emerging economy contexts, to more frequently update measures of market value (Campbell, Ramadorai, and Ranish, 2015). We also find immediate effects of India's massive anti-corruption demonetization campaign on bunching behavior in reported real estate transaction values, but little effect on overall under-reporting. Finally, we find evidence of a strong correlation between the degree of bunching of reported values at guidance values and features of the mortgage contract such as the LTV ratio on the loan, and the identity and financial health of the bank issuing the loan. This relationship is intriguing, and suggests a link between the quality and extent of financial screening and household incentives for tax

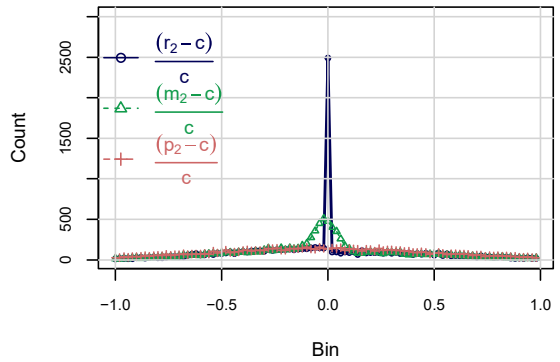
evasion, a link we believe should be explored more carefully and fully going forward.



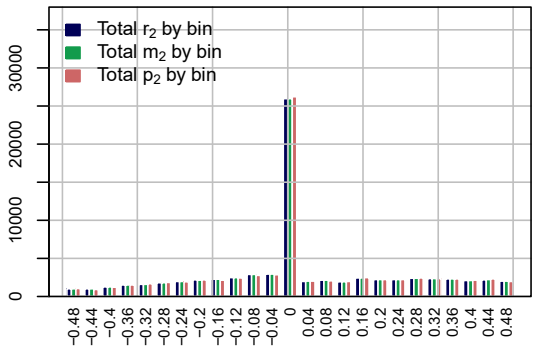
(a) High under-reporting, without measurement error



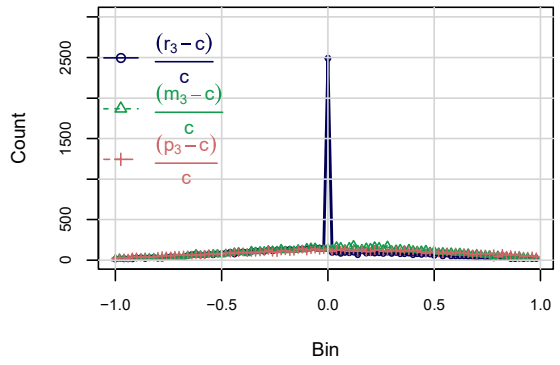
(b) High under-reporting, without measurement error



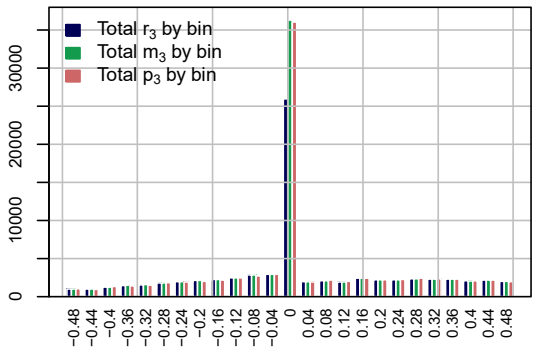
(c) No under-reporting, with measurement error



(d) No under-reporting, with measurement error



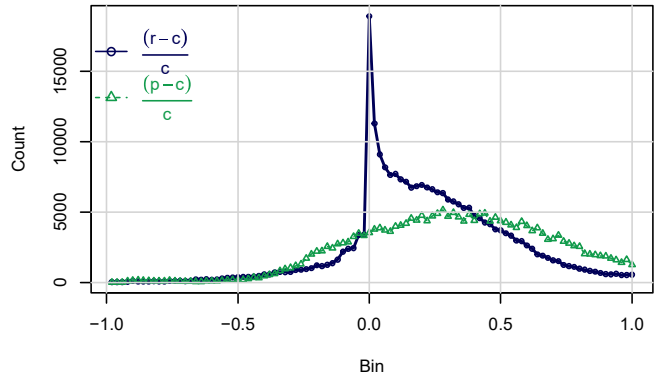
(e) High under-reporting, with measurement error



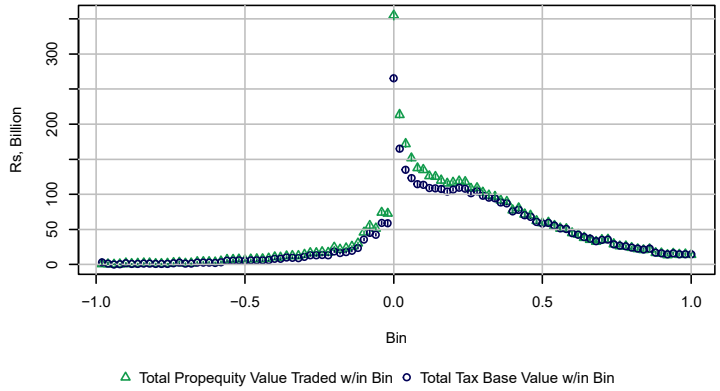
(f) High under-reporting, with measurement error

Figure 2.1. Simulation Results

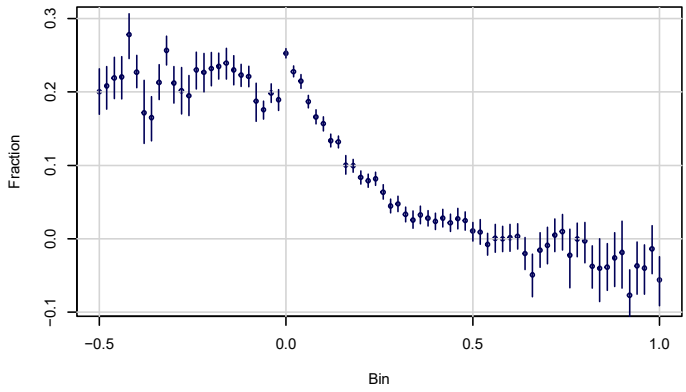
r_1 and m_1 are reported and transaction values in the high under-reporting case. r_2, m_2 and p_2 are reported, true transaction, and noisily measured transaction price variables for the no under-reporting case. r_3, m_3 and p_3 are reported, true transaction, and noisily measured transaction price variables for the high under-reporting case with measurement error. c is the government assessed value. Measurement error refers to noise in our estimates of market prices relative to the true unobserved market price.



(a) Bunching of Reported and Propequity Values Around Circle Values



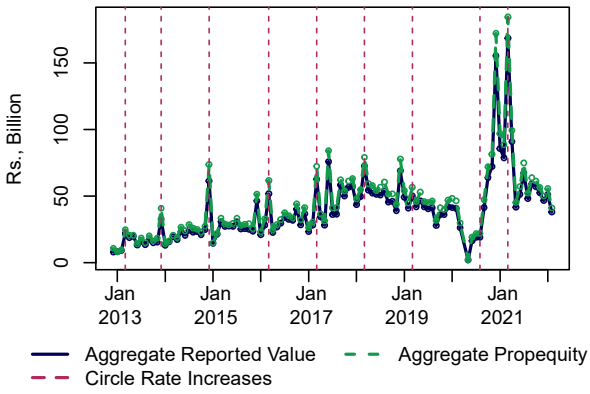
(b) Aggregate Taxbase and Propequity Values by Reporting Behavior Bins



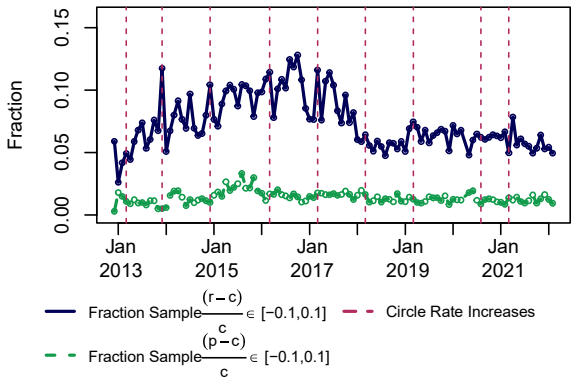
(c) Under-Reporting Rate by Reporting Behavior Bins

Figure 2.2. Main Results

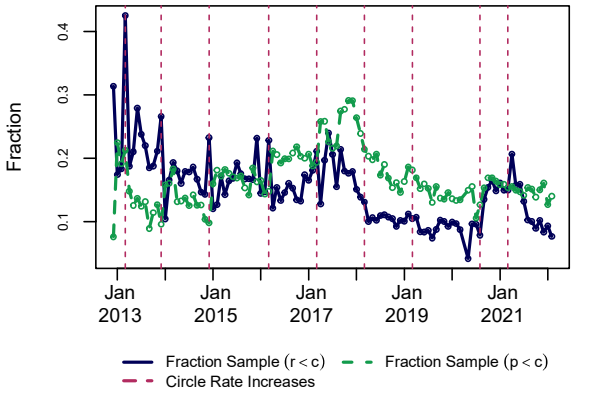
The blue line/circle shows the distribution of reported values across 2% reported value bins, where a reported value bin is measured as a deviation from the government assessed value. The green line/triangle shows the distribution of our noisily measured estimate of the market price (the Propequity values) for the same underlying set of transactions reported in the blue line/circle. The blue circles in (c) present the under-reporting rate within 2% reported value bins. Vertical bars represent 95% bootstrapped confidence intervals.



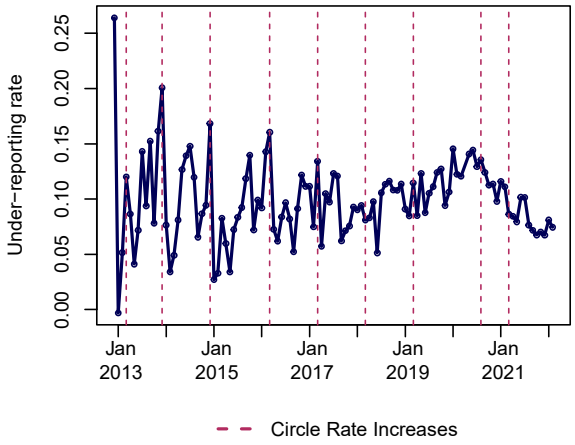
(a) Monthly Aggregates



(b) Monthly Bunching



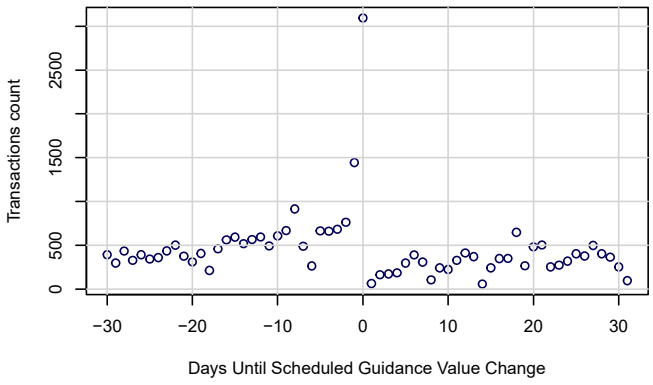
(c) Reporting/Market Values $<$ Guidance Values



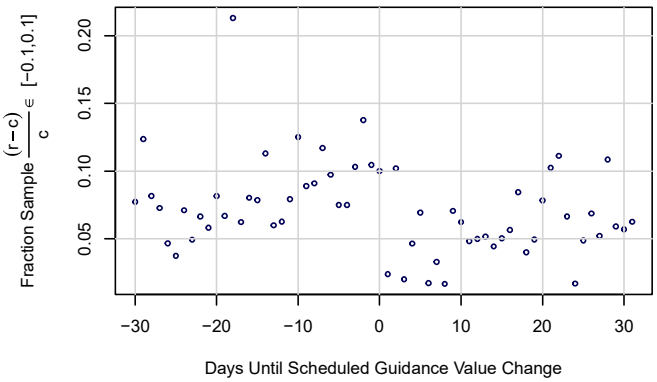
(d) Monthly Under-Reporting Rate

Figure 2.3. Heterogeneity Over Time

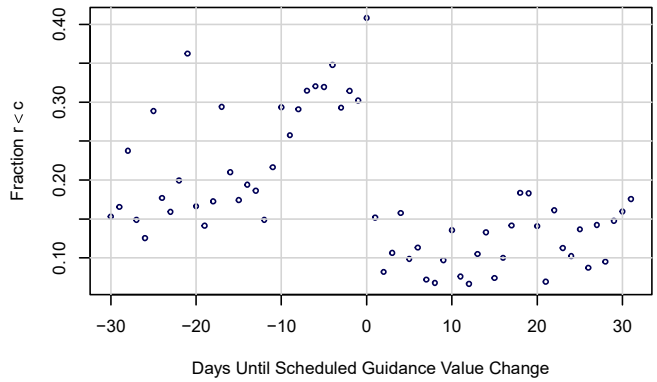
Red vertical dashed lines refer to scheduled circle rate (guidance value) increases. The circle rates were increased in 2013, 2014 and 2015; The government kept circle rates the same in 2016, 2017 and 2018, but the market likely still expected a possible increase in circle rates in those years. Circle rates were reduced in 2020 in response to the Covid-19 pandemic, but they increased again in 2021. The large increase in transactions in 2021 is likely due to a transaction tax rate reduction during that time period.



(a) Transaction Counts



(b) Fraction Bunching



(c) Fraction $r < c$

Figure 2.4. Reporting Behavior in Days Around Scheduled Guidance Value Increases

Panel (a) shows counts of transactions made per day within a 60 day window of all the scheduled guidance value changes that occurred over our sample period. Panel (b) shows the fraction of the transactions on the given event-day that "bunchers", i.e. had a reported value within 1% of the guidance assessed value. Panel (c) shows the fraction of transactions that had a reported value less than the guidance assessed value.

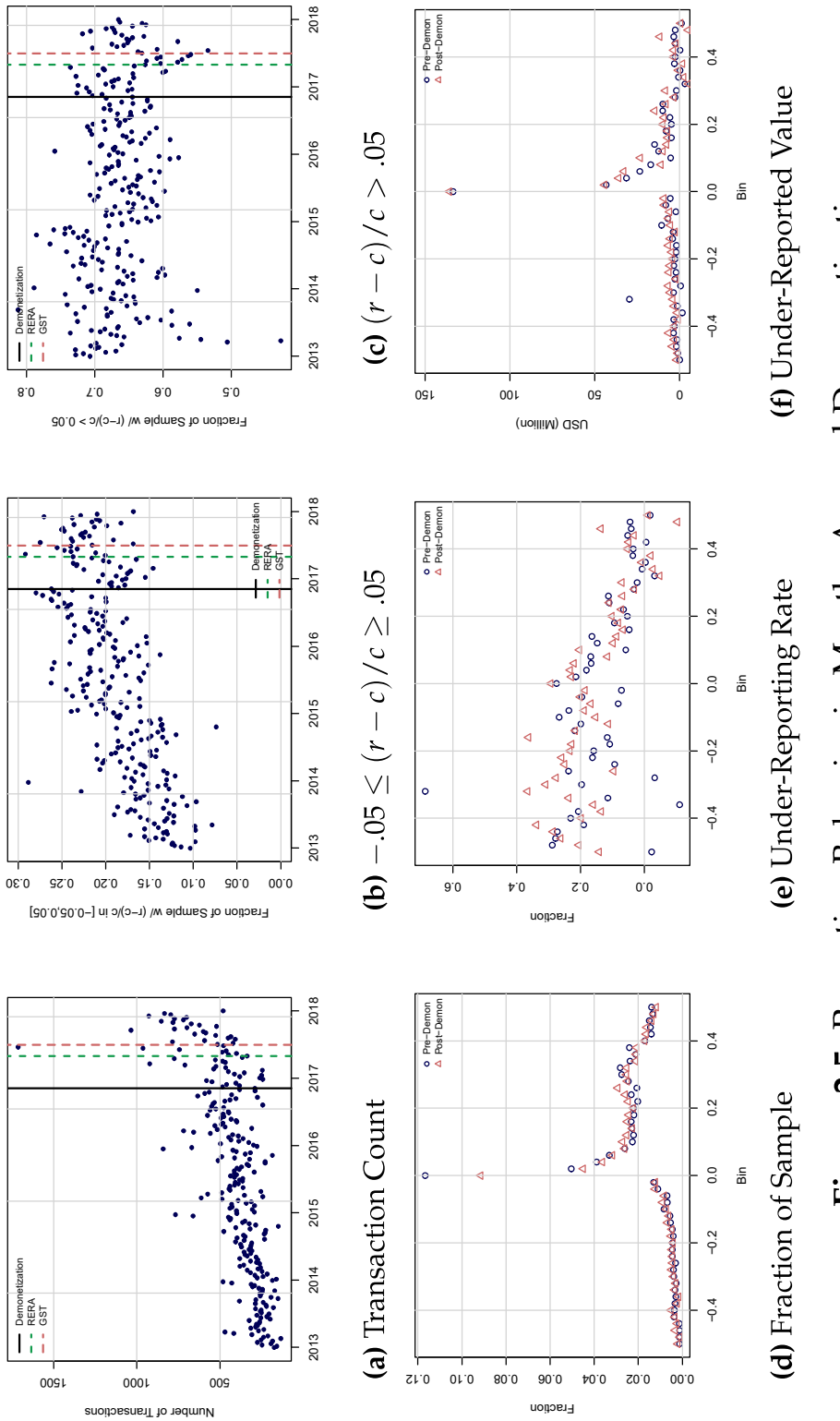
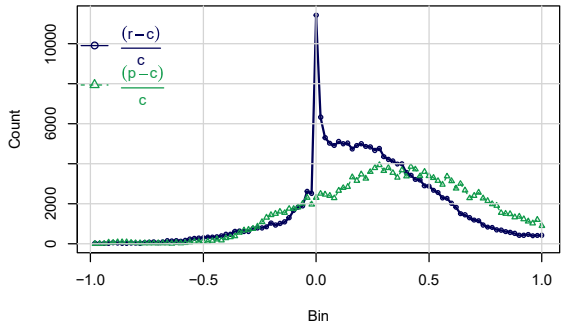
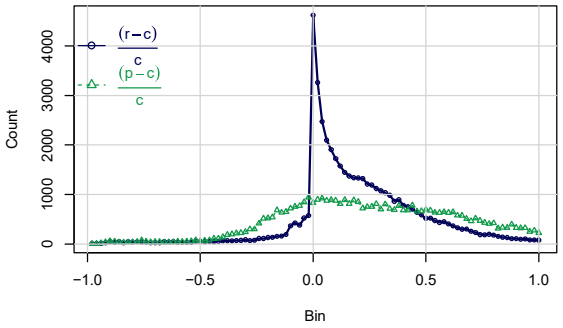


Figure 2.5. Reporting Behavior in Months Around Demonetization

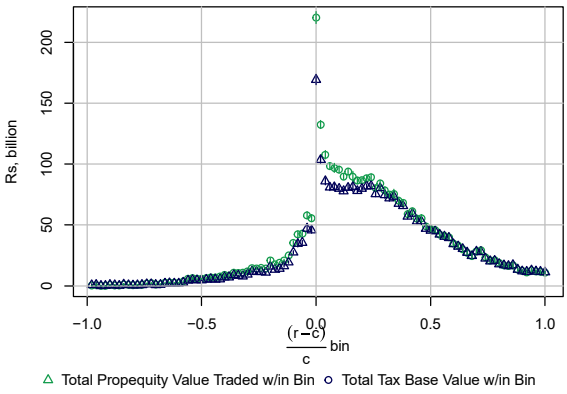
Figures 2.5b and 2.5c present the fraction of sample in each $(r-c)/c$ bin. Blue circles and red triangles in Figures 2.5d, 2.5e and 2.5f present estimates for 180-days before and after demonetization, respectively. The under-reporting rate in each bin in Figure 2.5e is calculated as $(M-T)/M$ where M is the total estimated market value of transactions within the bin, and T is the total tax base value reported to the government.



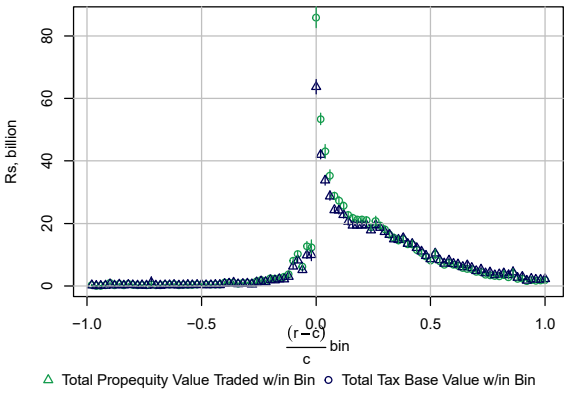
(a) Developer Sales



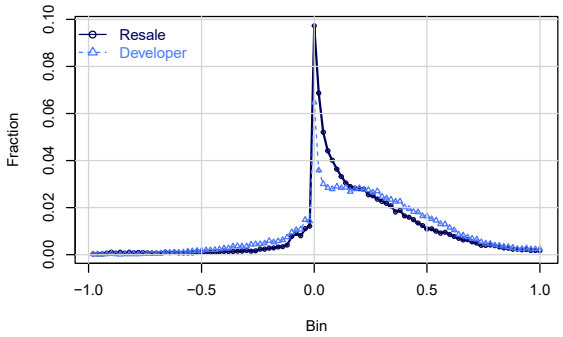
(b) Resale



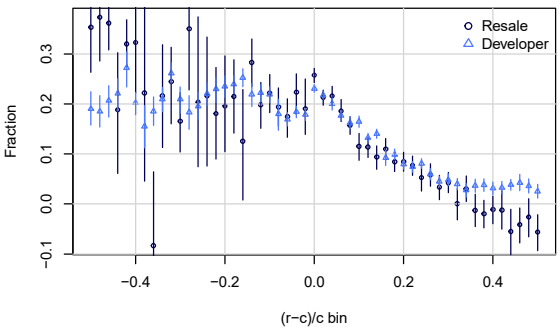
(c) Developer Sales



(d) Resale



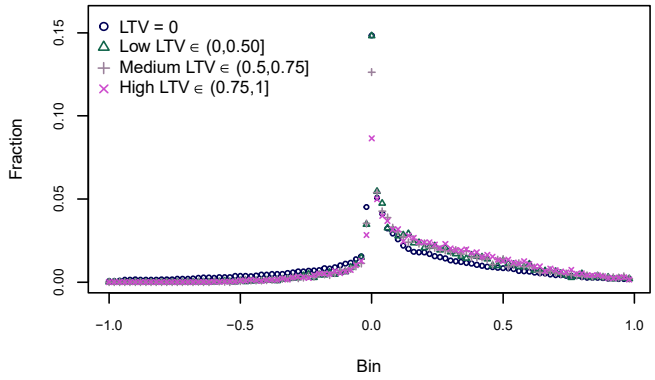
(e) Reported Counts (Density) Comparison



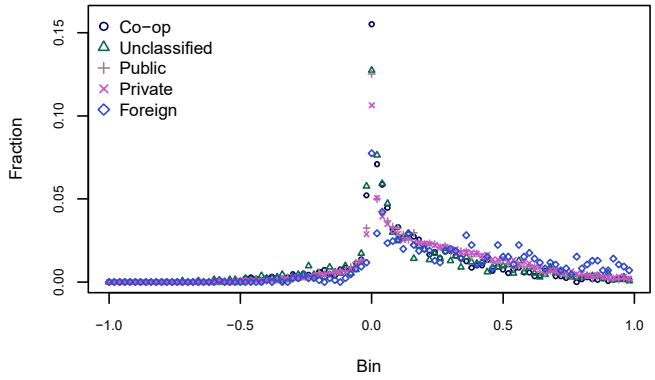
(f) Under-Reporting Rates

Figure 2.6. Developer vs. Resale Heterogeneity

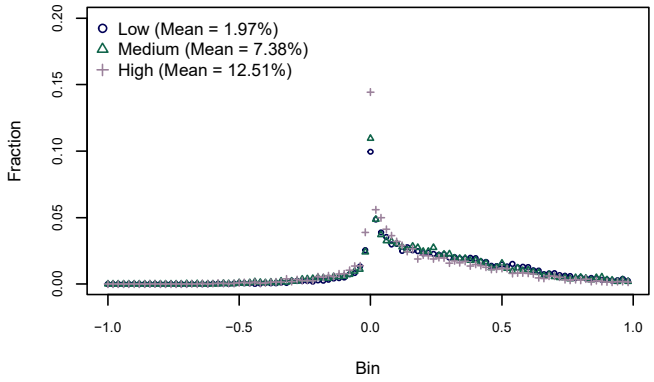
See Figures 2.2a, 2.2b and 2.2c for detailed descriptions.



(a) Loan-to-Value Bunching Heterogeneity



(b) Bank Type Bunching Heterogeneity



(c) Non-Performing Loan Heterogeneity

Figure 2.7. Heterogeneity by Mortgage Status

Panel (a) shows the distribution of transactions across 2% reported value bins, where reported value is measured as the percentage deviation from the guidance assessed value. The sample in (a) includes 187,999 transactions in Mumbai and Mumbai suburban districts from the IGR data until 2019. Panel (b) shows the distribution of transactions by the ownership structure of a bank making an associated mortgage - the sample here is 32,166 transactions where we were able to successfully match a mortgage. Panel (c) uses the same sample as panel (b) but presents distributions based on the terciles of the average non-performing loan rate of the associated bank.

	Reported Value '000s USD		Guidance Value '000s USD		Propequity Value '000s USD		Primary Transaction = 1		Area (sq M)		No. Obs.
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median	
2013	275.66	173.01	208.71	146.04	326.32	205.23	0.66	1	85.55	76.58	13,648
2014	313.27	195.57	234.15	158.61	362.44	225.50	0.69	1	88.75	75.14	17,213
2015	319.13	203.41	251.66	175.52	362.99	243.08	0.70	1	83.16	72.46	20,615
2016	315.70	205.50	258.22	175.88	363.41	241.19	0.70	1	79.13	69.90	23,803
2017	337.22	220.00	289.13	194.27	386.11	258.64	0.74	1	78.00	68.82	31,104
2018	324.03	212.86	264.82	175.48	367.51	243.07	0.76	1	73.53	65.04	38,228
2019	315.59	215.04	254.19	171.50	360.73	245.33	0.73	1	71.02	62.97	30,602
2020	320.36	210.00	266.78	172.50	371.71	241.19	0.72	1	71.35	62.34	30,289
2021	334.36	212.52	274.85	177.39	374.59	246.06	0.68	1	72.21	62.15	49,663
2022	324.45	220.00	257.77	176.44	355.16	240.91	0.63	1	68.14	60.57	5,449
Total	321.77	210.00	261.86	174.42	367.29	242.02	0.71	1	76.06	66.28	260,614

Table 2.1. Summary Statistics on Transactions

The table reports summary statistics for the set of transactions that is either matched to the same project from Propequity or to the nearest Propequity project. A primary transaction is one where the housing unit is sold by a real estate developer.

Percentile of Guidance Value Distribution	Lower End of Guidance Region ('000s US$)	Kink ('000s US$)	Upper End of Guidance Region	Conventional Elasticity	Standard Error of Conventional Elasticity
5–15	58	75	88	2.10	0.124
15–25	88	99	111	1.20	0.034
25–35	111	123	135	1.53	0.060
35–45	135	147	160	1.53	0.088
45–55	160	174	190	1.23	0.043
55–65	190	208	230	1.24	0.041
65–75	230	254	283	0.94	0.033
75–85	283	324	387	1.05	0.032
85–95	387	494	715	0.75	0.028

Table 2.2. Elasticity of Reported Value to Transaction Tax Rate

The table reports the formal estimates of the elasticity of reported value to the transaction tax rate.

Chapter 3

Privacy Policies and Consumer Data Extraction: Evidence From U.S.

With Tarun Ramadorai and Ansgar Walther

3.1 Introduction

There is growing public interest in consumers' data privacy, and in firms' ability to track consumers' behavior and personal details.[1] Academic research on data privacy has illuminated many facets of consumers' *demand* and willingness to pay for privacy (e.g., Goldfarb and Tucker, 2012; Acquisti, Brandimarte, and Loewenstein, 2015; Tang, 2019; Chen, Huang, Ouyang, and Xiong, 2021). The literature increasingly emphasizes potential behavioral biases that may affect consumers' privacy-related choices (e.g., Acquisti, John, and Loewenstein, 2013; Liu, Sockin, and Xiong, 2020). This is

[1]Faced with recent events, such as the Cambridge Analytica scandal, Congress is considering whether the US should adopt stricter, European-style regulation; see, for example: "Ad world flocks to Congress urging federal data privacy legislation", *The Drum*, 26 February 2019, and "Should Congress override state privacy rules? Not so fast.", *The Washington Post*, February 26, 2019.

supported by empirical evidence on the demand side. Indeed, many consumers can be reassured by the mere presence of legal text such as privacy policies or other generic information, even if such information is unrelated to firms' actual data extraction practices (Adjerid, Acquisti, Brandimarte, and Loewenstein, 2013; Hoofnagle and Urban, 2014; Athey, Catalini, and Tucker, 2017). In this context, we argue that it is especially important to understand the *supply* of privacy: What determines firms' data extraction practices? How do firms communicate these practices to consumers via their privacy policies? Only a small body of empirical work sheds light on these questions.[2]

In this paper, we provide new evidence on the supply of privacy by large US firms. We acquire, process, and analyze the privacy policies of a comprehensive set of large US firms. To acquire the data, we first search for the privacy policies of all 5377 US firms in Compustat, using a combination of automated Google searches, web crawling techniques, and manual searches using each firm's main web domain. In total, we are able to obtain policies for 4078 firms (75.4% of the entire Compustat sample). We also leverage recent web measurement tools (Englehardt and Narayanan, 2016) to analyze the code of each firm's website, and to detect the presence of third-party cookies and other data extraction devices.[3] We relate the attributes of firms' privacy policies, as well as firms' actual web tracking behavior, to a number of firm characteristics. Our contribution is to demonstrate three sets of stylized facts.

First, we find that most of the variation in privacy policies is within, rather than between, industries. Concretely, the median cosine similarity between individual poli-

[2]While there is a large theoretical literature in economics on the supply of privacy (for surveys, see Acquisti, Taylor, and Wagman (2016) and Bergemann and Bonatti (2018)), the empirical literature is smaller and focuses mostly on online markets (e.g., Bonneau and Preibusch, 2010; Kummer and Schulte, 2019). There is a broader literature in computer science that studies the properties of firms' privacy policies, but which does not typically link privacy policies to firm and industry characteristics (e.g., Amos, Acar, Lucherini, Kshirsagar, Narayanan, and Mayer, 2020, and references therein).

[3]We have made our data and code publicly available for other researchers at: https://github.com/ansgarw/privacy.

cies and the sample centroid is 0.57, which translates to a 55-degree median angle between policy word-frequency vectors and that of the average policy across all firms in our sample. This measure of similarity does not change substantially if, instead of comparing policies to the grand sample average, we compare them to averages within increasingly fine industry buckets.[4] Therefore, we conclude that firms do not use simple "industry boilerplate" policies. Rather, firms in the same industry appear to make conscious choices to use meaningfully different privacy policies, which motivates a deeper analysis of what drives these choices.

Second, we analyze the association between what firms say about privacy, as measured by their communication in privacy policies, and what they do in practice, as measured by their online data extraction routines. We relate the number of cookies on a firms' website, which closely correlates with other measures of data extraction, to several attributes of firms' privacy policy text. We show that firms who extract more data via cookies are more likely to have a privacy policy, display these policies more prominently on their website, and write longer policies. Moreover, we measure the legal sophistication of firms' policies, as judged by a human legal expert who read through a 10% sample of the privacy policies in the dataset. We show that firms who extract data using cookies write more legally sophisticated policies on average. These patterns prevail after removing industry fixed effects, and become stronger when we restrict our sample to firms in the service industry. It is therefore unlikely that they are driven by firms (e.g., traditional manufacturers) whose business model is orthogonal to consumer data.

This set of stylized facts has an interesting economic interpretation. In particular, the

[4]The median cosine similarity with the 3-digit SIC-level centroid is around 0.62, corresponding to a 51-degree median angle between firms' policy word frequency vectors and that of the industry-average vector. We find similar results when using topic frequencies from a simple language model. This suggests that the variation within industries is not driven by spurious semantic differences, such as two firms choosing different words to describe the same practice.

association of data extraction with visible yet sophisticated policies is possibly worrying if consumers are wrongly reassured by the mere presence of legal text. There is growing evidence in favor of such a hypothesis. In the 2009 *Berkeley Privacy Survey*, 62% of respondents believed that "If a website has a privacy policy, it means that the site cannot share information about you with other companies, unless you give the website your permission." (Hoofnagle and Urban, 2014, p.309). Athey, Catalini, and Tucker (2017) show that, in an experiment, the introduction of generic information about privacy protection, which was irrelevant to the actual extraction policy, made participants less likely to avoid surveillance.[5] Our evidence is consistent with these studies: If firms are aware of consumers' tendencies to be easily reassured, then those who intend to extract data have an incentive to prominently display their privacy policy, while simultaneously making it legally sophisticated.

Third, we relate firms' privacy policies and practices to their economic characteristics. We find pervasive size effects, with the largest firms conducting the most data extraction. We also find an interesting non-monotonic relationship between privacy policy attributes and firms' technical sophistication, which we measure using the Peters and Taylor (2017) measure of knowledge capital, expressed as a share of firms' total capital. Firms with intermediate technical sophistication are most likely to extract consumer data. Once again, we confirm that the groups of firms who extract consumer data (i.e., large firms or firms with intermediate knowledge capital shares) have more visible, longer and more legally sophisticated privacy policies.

The non-monotonic relationship between technical sophistication, privacy policies and data extraction is intriguing suggests that our standard measure of technical sophistication masks multiple dimensions of data-related technology, which lead to different

[5]Relatedly, Adjerid, Acquisti, Brandimarte, and Loewenstein (2013) show experimentally that small, irrelevant distractions reduce consumers' sensitivity to the policy

behavior by the most sophisticated firms. For instance, our findings are consistent with a scenario where the most technically sophisticated firms deciding to process data "in-house" rather than share it, or where these firms (e.g., big data intermediaries such as Google) have diverse alternative sources of consumer data without directly tracking visitors using cookies on their website. These results are also aligned with recent findings by Rock (2019), who argues that the launch of new data processing technology most benefited firms with intermediate initial sophistication.

Motivated by these empirical findings, we propose a simple theoretical model of data extraction and privacy policies, which reflects our preferred interpretation of the stylized facts. We model interactions between consumers and firms. Firms privately observe their type, which determines their incentives to extract consumer data, and decide on the contents and visibility of their privacy policy. Consumers do not observe the firm's type or their actual data extraction behavior, but can make inferences about these quantities from firms' privacy policies. A key ingredient in our environment is that privacy is, to some extent, a "shrouded attribute" in the sense of Gabaix and Laibson (2006). We assume that a fraction of consumers is myopic. Myopic consumers take the mere presence of a privacy policy as evidence of benign data extraction behavior, without perusing the policy's actual contents. In equilibrium, "active" type firms, who want to engage in data extraction, have an incentive to use visible policies as a tool to reassure potentially myopic consumers, thus shrouding their actual data extraction behavior. By contrast, "passive" type firms, who do not wish to extract consumer data, can use low visibility of their privacy policy as a *signal*, which will be interpreted as evidence of benign behavior by sophisticated consumers. We show that, in a separating equilibrium between active and passive firms, our model generates a positive correlation between visibility and data extraction, which a purely rational model cannot generate. Thus, our interpretation of the evidence is that a model with some degree

of shrouding is more consistent with the data than a purely rational one. We further demonstrate that this model can be made consistent with our remaining empirical findings, including the non-linear relationship between technical sophistication and privacy/extraction policies.

The remainder of the paper is organized as follows. Section 3.2 describes our data on firm characteristics, privacy policies, and web tracking behavior, and provides basic descriptive statistics. In Sections 3.3, 3.4 and 3.5 we present the three sets of stylized facts about variation in privacy policies, the association between privacy policies and data extraction, and the role of firm characteristics. Section 3.6 presents our simple model of data extraction and privacy policies. Section 3.7 concludes.

3.2 Data

3.2.1 Firm data

For all firms in the US Compustat database, we obtain data on market capitalization, book values of assets and equity, sales, intangible assets, R&D, SG&A, and marketing expenditures. For a more precise measure of intangibles, we construct intangible capital following Peters and Taylor (2017). This is the sum of Compustat-recorded on-balance-sheet intangible capital, and the replacement values of knowledge and organizational capital. Peters and Taylor (2017) estimate the replacement value of knowledge capital by accumulating past R&D expenditures for firms assuming an industry-specific depreciation rate, and the replacement value of organizational capital almost identically to Eisfeldt and Papanikolau (2014), by accumulating a fraction of past SG&A expenditures.

For stock variables (market values, book values, assets, and intangibles) we take the latest available quarterly observation (2018Q1). For flow variables (sales, R&D, and marketing expenditures) we take the average over the last three years. As in Crouzet and Eberly (2018), we drop firms with missing sales data, or values of sales that are less than or equal to zero, firms with missing data on asset book-value, or those with asset book-value less than or equal to US$ 1MM, firms with missing market value,[6] and firms that do not list their web domain on Compustat. Following these filters, the Compustat sample comprises 5,377 firms.

For each firm, we calculate the market-to-book ratio of assets, the firm's market share of sales in its (2-digit SIC code) industry, the share of its capital accounted for by intangible and knowledge capital (i.e., the fraction of knowledge capital and intangible capital in the firm to total capital, where total capital is the sum of (the replacement values of) knowledge capital and organizational capital, and total assets[7]), and the ratio of marketing and R&D expenditures to assets.

Table 3.1 shows descriptive statistics of these firm characteristics, following winsorization at the 1 and 99 percentile points to reduce the influence of outliers. On average, knowledge capital accounts for roughly 8% of total capital, with total intangible capital roughly double this amount. The median knowledge share is zero, meaning that this is a skewed distribution, and some firms in the data exhibit very high fractions of knowledge capital. There is also a fairly high standard deviation across firms in knowledge share.

[6] Where possible, we replace missing market values with the product of number of shares and price per share.

[7] Total assets in Compustat are the sum of Current Assets - Total (ACT), Property, Plant and Equipment (Net) - Total (PPENT), Investment & Advances - Equity (IVAEQ), Investment & Advances - Other (IVAO), Intangible Assets - Total (INTAN), and Assets - Other - Total (AO).

3.2.2 Privacy Policies

We search for firms' privacy policies using automated Google searches and web crawling techniques. We restrict this search to each firm's main web domain, as listed in Compustat. We supplement this method by a web crawl of the firm's domain, and finally, by manual checking, in cases where a policy was not found automatically. We scrape the text of each policy, and discard it if it does not contain the word "privacy". We also discard all paragraphs that have fewer than 100 characters (usually consisting of headings or snippets of HTML code). In total, we are able to obtain policies for 4,078 firms (75.4% of the sample).[8] To prepare the policies for textual analysis, we remove all non-English words, and words associated with named entities such as organizations, persons, and locations. We also remove very common English words from a standard list of "stop words" that convey little semantic meaning (e.g., "is","in", "and", "or").[9]

Figure 3.1 visualizes the textual content of our sample in terms of the most important bigrams (pairs of consecutive words). We measure the importance of bigrams in each policy with a standard TF-IDF (term frequency-inverse document frequency) metric, which attaches high importance to bigrams that are frequent within a policy relative to its overall length, and penalizes generic bigrams that occur in a large fraction of documents.[10] As might be expected, the policies prominently feature the word "privacy."

[8] These policies are scraped from 4,062 unique web domains. One web domain is shared by three firms in our sample, and 14 domains are shared by two firms each.

[9] We detect non-English with the pyenchant spellchecker (see), and named entities with the Stanford NER tool (see Finkel, Grenager, and Manning (2005) and). We use the NLTK list of English stopwords (see).

[10] We divide each bigram's number of occurrences in each policy by the total number of bigrams in the policy, for the bigram's "term frequency" (TF). We then multiply the TF by the log of the inverse fraction of documents containing the bigram, known as the "inverse document frequency" (IDF). Let P_{ij} be the number of times that bigram j appears in document (in our case, policy) i. The TF-IDF metric is:

$$\hat{P}_{ij} = \underbrace{\left(P_{ij} / \sum_k P_{ik} \right)}_{TF} \cdot \underbrace{\log \left(\frac{N}{\sum_i 1\{P_{ij} > 0\}} \right)}_{IDF}$$

They also prominently feature the terms "personal information," "personal data," and "personally identifiable information." Finally, an important and frequently used term that is clearly evident is "third party," which we will return to discussing later in the paper.

Table 3.2 shows descriptive statistics for four sets of attributes of privacy policies. The first two rows show how easy it is to find privacy policies. We found privacy policies for about 75% of firms in our sample of 5,377 firms in Compustat. A link containing the word "privacy" is visible on the homepages of only 65% of the sample.[11] The third and fourth rows show the length of policies in terms of paragraphs and words. When found, the average privacy policy contains around 31 paragraphs, comprising roughly 1,900 words, with considerable variance across policies in both length and paragraph structure. The correlation between the word count and paragraph count is 91%—we use word count to capture policy length in the remainder of the paper. The final row shows the descriptive statistics of the Gunning (1952) Fog index of "readability", which is based on the sentence-level frequency of complex and polysyllabic words.[12] The Fog index heuristically measures the number of years of formal education required to understand a document at first reading. By this metric, one needs at least a college degree to follow the median privacy policy in our sample, highlighting the (lack of) readability in most privacy contracts.

Figure 3.2 shows the histograms of privacy policy length and readability, which describe their full distribution. The distribution of length, which the figure shows on a

where N is the total number of documents. See Rajaraman and Ullman (2011), and Gentzkow, Kelly, and Taddy (2018) for more detailed treatments of TF-IDF.

[11] Roughly 92% of the 4,078 policies acquired are found using google searches, while the remaining 8% were recovered manually.

[12] These figures are similar to recent findings in the Computer Science literature (e.g., Fabian, Ermakova, and Lentz, 2017).The Gunning algorithm determines the average sentence length in the policies, and counts all words comprising three or more syllables. To get the final Gunning measure, the algorithm just adds the average sentence length and percentage of three-or-greater syllable words, and scales the result.

log scale, is skewed to the right, and a large number of policies feature between 1,000 and 10,000 words. The distribution of the Fog index suggests that hardly any policies are (heuristically) comprehensible to a reader with only a high school education (12 years), and that a few firms wrote policies with Fog indices above 22, meaning that they are estimated to require a PhD degree to read.

Given the apparent complexity of privacy policies, we sent 407 randomly selected policies (10% of unique scraped policies) to a legal expert for evaluation, to rate their degree of legal sophistication and quality. Of these policies, 54 did not contain meaningful legal text related to privacy or did not pertain to firms in our sample. For the remaining 353 policies, he assigned high (1), neutral (0) or low (-1) overall scores to each policy's degree of legal sophistication. The overall score is an amalgam of five dimensions of legal quality that we asked the expert to evaluate; we summarize the expert's detailed evaluation criteria in Appendix Section C.2.[13] To summarize, high scores were assigned to policies that clearly document the company's intended data collection and sharing practices, as well as the user's rights and ability to give consent. Figure 3.3a shows the distribution of overall scores. 42% of policies are classified as "Neutral", with a larger fraction (35%) classified as "Low," than "High" (23%) as regards their overall sophistication.

Figures 3.3b and 3.3c visualize important bigrams associated with policies that are ranked as high or low sophistication overall. To highlight bigrams that are specific to high and low rated policies, the importance of bigrams in this figure is normalized by the average importance in the population (i.e., those depicted in Figure 3.1). Based on these word clouds, we construct a simple measure of overall policy sophistication that can be extrapolated to our entire sample, i.e., to help classify those policies not

[13]These categories are: Data collection, user consent, responsible use, third-party sharing, and user rights. Table C.1 in the appendix shows that a regression of the overall score on the individual category scores yields an R^2 statistic of 81%.

reviewed by the expert. In particular, for each privacy policy, we compute the total frequency of the most important 100 bigrams in policies rated "High" by our expert (i.e., the unweighted sum across all the bigrams in the word cloud in Figure 3.3b) and subtract the equivalent metric for "Low" policies (i.e., the sum across Figure 3.3c). We term the result the "Legal Sophistication" index with different scores assigned to each of the 4,078 policies in our dataset using this simple scoring technique.[14],[15]

3.2.3 Data Extraction Behavior

We obtain data on each firm's data extraction behavior using the "openWPM" crawler developed in Englehardt and Narayanan (2016).[16] We emphasize that this measure is collected separately from, and has no direct relationship with the privacy policy data. There are null returns for which the crawler fails, and these reduce the sample size to 5,216 of 5,377 total firms.

Descriptive statistics of these tracking data are in Table 3.3. The first row shows the number of "first-party" cookies that each firm places on its own website. The second row shows the number of unique third parties (i.e., agents other than the firm itself) who place cookies on the firm's website that are classified as "tracking cookies." For an alternative measure of third-party activity, the third row shows the *total* number of third-party requests encountered on the firm's website, including, but not limited to cookies. The correlation between the two measures of third-party activity is 75%.[17]

[14]Let G and B be the set of bigrams appearing in Figure 3.3b and 3.3c respectively, and let $\hat{P}_{ij}$ be the TF-IDF frequency of bigram j in policy i. Then, our measure of quality is $Q_i = \sum_{j \in G} \hat{P}_{ij} - \sum_{j \in B} \hat{P}_{ij}$. We normalize this measure to have mean zero and variance one.

[15]This can be thought of as a very simple supervised learning model using the classifications in the training data. In Appendix Section C.9, we test how different the results are when we use more sophisticated approaches common in the NLP/machine learning literature.

[16]Their open-source privacy measurement software is available at . We obtain our data by scraping the results for each firm on , which uses OpenWPM.

[17]Using third-party tracking cookies instead of requests is more conservative, because tracking is

For simplicity, we use total cookies, defined as the sum of first-party and third-party tracking cookies, as our overall measure of data extraction in what follows. In Appendix Section C.7, we show that our main results also apply when we focus on either first- or third-party data extraction.

Figure 3.4 shows histograms of our data extraction measures. For all measures, the distribution of data extraction activity is skewed to the left, with many firms having zero or a small number of cookies / requests, and has a long right tail, with up to 200 third-party requests and up to around 60 in total.

3.3 Variation in Privacy Policy Text

This section presents our first set of stylized facts, which is based on a decomposition of the variation in firms' privacy policy text. We use the common *cosine similarity* measure, which quantifies the distance between the content of different text documents, to compare text documents across and within industries.

Concretely, the text of each firm's privacy policy can be described as a vector $P_i = (P_{i1}, ..., P_{iM})$ of word frequencies, where P_{im} is the frequency of word m in policy i, and M is the total number of words used in our sample of privacy policies. The cosine similarity C_{ij} between two policies is the cosine of the angle between their vector representations P_i and P_j:

$$C(P_i, P_j) = \frac{P_i \cdot P_j}{\|P_i\| \, \|P_j\|}$$

For an intuitive interpretation, suppose the only two possible terms are "apple" and "orange". If policy i only mentions apples, and policy j mentions only oranges, then

not always done via cookies. This approach is also quite accurate, because not all cookies are tracking cookies.

the angle between the two vectors is 90 degrees (i.e., they are orthogonal), and $C = 0$. If both policies mention only apples, then the angle is zero, and $C = 1$.[18] To measure aggregate variation, we compute the cosine similarity between each policy vector P_i and the centroid vector of all privacy policies in the sample, i.e., the "average" policy $\bar{P} = \left(\sum_j P_j\right)/N$. To isolate variation within industries, we compute the similarity between each policy and the associated industry-level centroid $\bar{P}_I = \left(\sum_{j \in I} P_j\right)/(\#I)$, where I is the set of firms in an industry.

Figure 3.5a shows the cumulative distributions of cosine similarities with the sample centroid, the centroid associated with each firm's SIC sector, as well as with the centroid associated with each firm's 2- and 3-digit SIC code bucket. The median cosine similarity between individual policies and the sample centroid is 0.57, which translates to a 55-degree median angle between policy vectors and the grand average policy. This measure rises to a median cosine similarity of about 0.62, corresponding to a 51-degree median angle, when we compare firms' policies and the industry-average vector at the 3-digit SIC level. Figure 3.5b shows the associated mean cosine similarities, with 95-percent (bootstrapped) confidence intervals.

The figures show that *within-industry* variation is only marginally smaller than *total* variation in privacy policies. Indeed, only about 4 degrees of the median angle between policies and the average policy can be "explained" by accounting for the industry-specific frequency of words. The upshot of this analysis is that privacy policies are not "boilerplates" that respond to industry-level regulation (under a pure "industry boilerplate" hypothesis, the cosine similarity between each firm's policy and the industry-level centroid would equal one). By contrast, we find that most of the

[18] Note, however, that this measure is nonlinear due to the cosine transformation: If a third policy k mentions apples and pears with equal frequency, then the angle is 45 degrees and $C_{ik} = \cos(45\text{deg}) = 0.71$. Since the cosine wave becomes steeper when moving from zero towards 90 degrees, the similarity measure is therefore more forgiving of small discrepancies between policies.

variation in firms' policies occurs within rather than between industries.

It is important to verify that these conclusions do not arise purely from idiosyncratic difference in wording. To check the robustness of our conclusions, we repeat the decomposition exercise above with a more specific measure of the semantic content of each policy in Figure C.1 of the appendix. We use Latent Semantic Analysis (LSA), which amounts to a singular value decomposition on the term-document matrix, to reduce the dimension of the textual data from about 18,000 words to 250 latent "topics". We expect this transformation to eliminate low-level differences in wording, and to focus on semantic content. After the transformation, we find that the level of all cosine similarities naturally rises,[19] but it remains the case that only a small fraction of the existing variation can be explained by industry effects.

While we have established that firms' privacy policies differ considerably, and that there is seemingly no convergence to a "boilerplate" policy, it is entirely possible that the documented variation across policies is simply idiosyncratic noise in the quality and quantity of verbal expression across firms. We now investigate whether data extraction practices and firm characteristics are systematically associated with privacy policy content.

3.4 Data Extraction Practices and Privacy Policies

In this section, we present our second set of stylized facts, which focuses on the association between firms' data extraction practices, as measured by total cookies in our data, and the content of their privacy policy text. Recall that our measure of cookies is obtained separately from privacy policy text, using the openWPM crawler of Engle-

[19]The median cosine similarity relative to the sample centroid is about 0.63 in latent topic space, while the median similarity to the SIC-3 industry centroid is about 0.66.

hardt and Narayanan (2016).

Figure 3.6 presents a simple, non-parametric visualization of the relationship between data extraction and privacy policies. In each panel, we sort firms into quintiles of data extraction intensity along the horizontal axis, and plot the associated means of privacy policy attributes on the vertical axis. The error bars denote (bootstrapped) 95-percent confidence intervals.

Figure 3.6a and 3.6b show that firms who use more cookies to track consumers are more likely to have a privacy policy on their website, and also more likely to display this policy visibly using a link on their homepage. Figure 3.6c and 3.6d show that, conditional on having a privacy policy, firms in the top quintiles of data extraction use significantly longer and more legally sophisticated text.[20] These facts, along with the substantial cognitive costs of reading privacy policies (e.g., McDonald and Cranor, 2008), suggest that there is a larger cognitive cost to consumers of reading the policies of firms in the top quintiles of data extraction. Figure C.2 in the appendix shows that the same patterns continue to arise if we sort firms by data extraction intensity after removing industry fixed effects, which is important, given the possibility of cross-industry variation in the importance of a web presence for sales and marketing purposes.

We make three further observations, which suggest that our results are not driven by the mechanical effect of firms whose business models are orthogonal to data collection (e.g., soup manufacturers). First, Figure 3.6 shows an association between data extraction practices and privacy policies even if we condition on firms that are relatively active data extractors. For example, the privacy policy attributes of firms in the third

[20] Figure 3.6e shows that there is no significant association between data extraction and the Gunning Fog index of readability. The contrast between this result and Figure 3.6d suggests that the Gunning method, which focuses on the number of polysyllabic words, cannot capture subtle differences in legal sophistication.

and fifth quintiles of total cookies are significantly different. Second, Figure C.3 in the appendix shows that similar patterns arise when we focus on firms in the services industry, who are more likely to have an inherent interest in consumer data. Finally, Table C.9 in the appendix shows that even after controlling for a broad set of attributes, more data extraction goes hand in hand with more sophisticated privacy policies.

Our preferred interpretation of these patterns is based on the recent evidence of consumer behavior, which suggests that some consumers are easily reassured by visible privacy policies, even if those policies do not contain information that genuinely protects them (e.g., Adjerid, Acquisti, Brandimarte, and Loewenstein, 2013; Hoofnagle and Urban, 2014; Athey, Catalini, and Tucker, 2017). Our empirical results on firm behavior are consistent with this "reassurance" effect, and more importantly, with a scenario in which firms are actively aware of it. In Section 3.6 below, we develop a simple signaling model inspired by the Gabaix and Laibson (2006) model of shrouded attributes, which rationalizes our empirical findings. In the model, firms are aware that some consumers have rational expectations, while other consumers are unsophisticated and "myopic," in the sense that they can be easily reassured by the mere presence of a visible privacy policy. In this environment, firms that choose low privacy policy visibility credibly signal their low data extraction capacity to sophisticated consumers, as doing so drives away myopic consumers, which is differentially more costly for the types of firms that extract consumer data. Data extracting firms make their privacy policies visible to facilitate data extraction from myopic consumers, and make these policies legally sophisticated to cover the firm against possible legal liability.[21] The patterns predicted by this simple theory of reassurance effects appear consistent with our empirical results.

[21]Notice that this signaling effect would be absent in a model where all consumers are rational. Indeed, with rational expectations, firms who do not extract data should have a strong incentive to display a simple privacy policy that states "we do not collect any data".

3.5 Firm Characteristics and Privacy

This section contains our third and final set of stylized facts. We investigate whether firms' economic characteristics can be used systematically to predict data extraction behavior and the nature of their privacy policies. We start by analyzing the relationship between the characteristics of privacy policies and two firm characteristics that seem intuitively important as a simple first step, namely, firm size and a measure of firms' technical sophistication. Our measure of technical sophistication is Peters and Taylor (2017)'s measure of "knowledge capital," which is essentially past accumulated R&D expenditures for firms assuming an industry-specific depreciation rate. To make this measure comparable across firms, we simply scale it by firms' total (i.e., tangible plus intangible) capital as in Table 3.1, and term the result the "knowledge share."

We first sort firms into quintiles of size, as measured by market value, and plot the associated characteristics of privacy policies and data extraction behavior. Figure 3.7a shows that larger firms are significantly more likely to extract consumer data. Moreover, Figures 3.7b-e show patterns that are consistent with our results in the previous section: Larger firms, who extract more data, are also more likely to have a privacy policy and to display it visibly. Conditionally on having a privacy policy, large firms write longer and more sophisticated policies.

Figure 3.8 sorts firms into firms with zero knowledge capital (about 55% of our sample), and then into four quartiles conditional on having positive knowledge capital. Figure 3.8a shows an interesting pattern. The average number of cookies rises with the knowledge share, but only for the first three buckets, i.e., for firms with a zero knowledge share or those in the first three quartiles of knowledge share. The third quartile of firms with positive knowledge share has the highest average number of

cookies in our sample. The firms in the top quartile of knowledge share, by contrast, engage in significantly less data extraction. Once again, the attributes of privacy policies follow the same, non-monotonic pattern, which is consistent with our results in Section 4. Indeed, firms in the third quartile of positive knowledge shares are most likely to have a privacy policy and to make it visible, and they also write the longest, most legally sophisticated privacy policy text.

Table 3.4 confirms these findings in multivariate regressions of data extraction and privacy policy attributes on firm characteristics. The top panel (a) of the Table contains regressions without industry fixed effects, while the bottom panel (b) of the table contains sector fixed effects at the level of SIC divisions.[22] The columns of the table correspond to the various attributes of the privacy policies that are on the left-hand-side of the regressions. The rows show the variables that are on the right-hand-side. The first right-hand-side variable is the log Market Value (i.e., size) of each firm. We further include the knowledge share, i.e., the share of the firm's knowledge capital as a fraction of its total capital, and its square, to capture the nonlinearity we detected earlier in Figure 3.8. The fourth right-hand side variable is the log Market Share measured as the firm's sales divided by industry sales at the 2-digit SIC code level; we include this because the classical theory (e.g., Spence, 1975) suggests that market power is a possible determinant of non-price attributes of firm behavior, such as privacy.

The table shows that for virtually all of the privacy policy attributes, there is a positive and statistically significant relationship with firm size, and that this relationship appears to hold within industries as well as in the specification without fixed effects. There is attenuation in the statistical significance of the coefficients in some cases with the introduction of the fixed effects, but no attenuation in the economic magnitude of

[22] See Figure 3.5 for the definition of SIC divisions. Table C.6 in the appendix shows that the estimated coefficients are of similar magnitude, albeit less precisely estimated, with SIC2 level fixed effects.

the coefficients, suggesting that this is primarily a power issue rather than an issue of between-industry variation being the proximate source of variation. Table C.3 in the appendix shows the robustness of our results when we control for a wider set of firm characteristics, including firms' marketing expenditures as a fraction of total assets, and firms' market-to-book ratios.

The knowledge share also continues to have a positive and statistically significant relationship with the policy attributes both with and without the inclusion of industry fixed effects, and the nonlinearity also shows up clearly in this case—the coefficient on the squared knowledge share is always negative and almost always statistically significant in the attributes for all regressions. Table C.4 in the appendix confirms that these patterns continue to arise, and indeed become stronger, when we focus on non-manufacturing firms. This once again suggests that our results are not driven by the mechanical impact of firms that have no interest in consumer data.

Overall, we find robust size effects, which are consistent with large firms engaging in more data extraction. This result can perhaps be explained easily in a model where there are economies of scale, or fixed costs, associated with the collection and processing of consumer data. We also find a persistent non-linear effect of technical sophistication, as measured by firm capital generated through research and development. The most sophisticated firms behave differently, engage in less data extraction than, and write more concise privacy policies from those with intermediate sophistication. This finding suggests that there may be multiple dimensions of technical sophistication. For instance, it is conceivable that the firms with the most research-related capital have alternative ways of acquiring consumer data (e.g., because they are themselves data intermediaries). This means that they can credibly signal to myopic consumers that they do not extract data, as they acquire these data without any need to interact directly with consumers. Another possibility is that such firms may simply have the

ability to process consumer data "in house" rather than sharing it with third-party intermediaries, obviating the need for an extensive privacy policy. Such features of firm technology might also explain why we find that the most sophisticated firms engage in both more first-party and more third-party data extraction.[23]

3.6 A Simple Model of Privacy Policies, Visibility and Signaling

In this section, we develop a simple, illustrative model of firms' privacy policy choices that rationalizes our empirical findings. From a modelling perspective, two of our empirical results are perhaps hardest to rationalize: First, as we emphasized in Section 3.4, we find a robust *positive* correlation between the visibility of firm's privacy policies and the intensity of their data extraction practices. The economic reason for this correlation is unclear: In a standard rational model, firms who have no inherent incentive to extract consumer data should have an incentive to post a highly visible policy with the legally binding statement that they do not collect any consumer data. Second, as we showed in Section 3.5, there are interesting nonlinear relationships between firms' technical sophistication and their privacy/data extraction policies. The model we present is geared towards understanding these features of the data alongside our other empirical findings. A key assumption in our analysis, guided by the existing survey and experimental evidence (Hoofnagle and Urban, 2014; Athey, Catalini, and Tucker, 2017), is that consumers are boundedly rational, and may be reassured by the mere presence of a privacy policy without perusing its full contents.

[23]See also Tables C.7 and C.8 in the appendix, where we confirm the same patterns using separate measures of first-party and third-party cookies.

3.6.1 Environment

We study a static model with a single firm and a unit mass of consumers. The firm offers a product to consumers at a fixed price $p > 0$. The firm can earn additional revenue from each transaction with consumers by extracting consumer data. The firm has a binary type $\theta \in \{0, 1\}$, which determines its ability to extract and use consumer data. The type θ is the firm's private information.

In particular, we assume that the firm with $\theta = 1$ can increase its revenue from each transaction with a consumer by $m > 0$ dollars by monetizing the consumer's data, while the firm with $\theta = 0$ cannot increase its revenue by monetizing data. We let π_0 denote the prior probability of facing a firm with type $\theta = 1$. Notice that the difference between the two types of firm in this environment is the marginal revenue that the firm can achieve by monetizing data. In this sense, the type $\theta = 0$ with low data-related marginal revenue can capture two different scenarios: The firm with $\theta = 0$ may not have the technical capacity to extract and process data, or the firm may already have access to abundant consumer data without extracting it by traditional methods. Both scenarios yield the same objective function for the firm, which effectively does not care about acquiring consumer data when $\theta = 0$. Mapping this idea to the data, one should interpret type $\theta = 0$ as representing either unsophisticated firms or very sophisticated firms that already possess a wealth of data. The latter scenario can arise, for example, if the firm is itself a data intermediary.

The firm in our model decides how visible to make its privacy policy. We denote the visibility of its policy by $v \geq 0$. For now, we focus on the firm's choice of visibility, and discuss the implications of the model for other observable attributes of privacy policies (e.g., content and length) after presenting the main results.

Consumers fall into two groups. A proportion $1-\alpha$ of consumers is *sophisticated*. Sophisticated consumers have rational expectations about the value of θ based on the firm's behavior in equilibrium, which we define in detail below, and act to maximize their utility given these expectations. To model the preferences of sophisticated consumers, we directly specify their behavior as a function of their beliefs about the firm's data extraction type. In particular, we assume that a fraction $\sigma(\pi)$ of sophisticated consumers are willing to buy the firm's product if they assign probability π to facing a data-extracting firm (i.e., to the event that $\theta = 1$). We assume that $\sigma'(\pi) < 0$, which reflects consumers' preference, ceteris paribus, for dealing with a firm that does not extract data. For example, Rainie, Kiesler, Kang, Madden, Duggan, Brown, and Dabbish (2013) provide survey evidence suggesting that some consumers stop using firms' services if they are concerned about data extraction. Furthermore, we assume that a fraction $\eta > 0$ of sophisticated consumers who buy from the firm are subsequently able to use technology such as "private browsing" to hide their data from the firm. See, for example, Kang et al. (2013; 2015) for recent empirical evidence that consumers take steps to avoid data collection.

The second group of consumers, which makes up proportion α of the population, is *myopic*. Myopic consumers' privacy decisions are guided only by the visibility of a privacy policy. We assume that a share $\mu(v)$ of myopic consumers behaviorally purchases the firm's goods are a function only of visibility. We assume that $\mu'(v) > 0$, which reflects the evidence that some consumers tend to be reassured by the mere presence of a privacy policy (e.g. Hoofnagle and Urban, 2014). We assume that the following boundary condition holds:

$$\lim_{v \to \infty} \alpha \left(\mu(v) - \mu(0) \right) < (1-\alpha) \left(\sigma(0) - \sigma(1) \right) \tag{3.1}$$

This condition places an upper bound on the number of myopic consumers that the firm can gain by increasing visibility v, which are on the left-hand side of the inequality. The upper bound, on the right-hand side of the inequality, measures the number of sophisticated consumers that the firm can gain by signaling that it does not extract data. We discuss the role of this condition in more detail after presenting our results below.

The timing of the game is as follows: First, each firm privately observes its type θ and determines the visibility v of its privacy policy. Consumers observe these choices. Finally, consumers form beliefs about θ, and decide whether they are willing to buy the firm's product based on their beliefs about θ (if sophisticated) or on the visibility v of the policy (if myopic). A perfect Bayesian equilibrium is defined by:

1. A strategy for the firm that specifies the visibility v as a function of the firm's type θ;

2. A belief $\pi(v) = Pr[\theta = 1|v]$ that sophisticated consumers use to evaluate their utility after observing v, and which satisfies Bayes' rule whenever possible;

3. The shares $\sigma(\pi)$ and $\mu(v)$ of sophisticated and myopic consumers, respectively, that choose to buy the firm's product after observing v.

In equilibrium, the firm maximizes expected profits given consumers' behavior.

3.6.2 Separating Equilibria

We demonstrate that, under mild conditions, there exists a separating equilibrium in which the visibility v of the firm's privacy policy is increasing in its data extraction type θ. Intuitively, we will argue that low visibility v is a credible signal of low data

extraction capacity θ, because doing so drives away myopic consumers, which is differentially more costly for data extracting types. This effect ensures that the standard "single crossing" for separation is satisfied.

We conjecture and verify the following equilibrium: Firms that do not extract data (type $\theta = 0$) choose visibility $v = 0$, and in which firms that extract data (type $\theta = 1$) choose visibility $v = v_1 > 0$. To support this equilibrium, we specify that consumers have beliefs $\pi = Pr[\theta = 1|v] = 0$ when they observe $v < v_1$ and $\pi = 1$ when they observe $v \geq v_1$. These beliefs clearly satisfy Bayes' rule on the equilibrium path. To check that firms are maximizing expected profits, we must verify that two incentive compatibility constraints are satisfied.

On one hand, a firm of type $\theta = 1$ must not want to deviate to an invisible policy with $v = 0$. Formally, this constraint is

$$(1-\alpha)\,\sigma(1)\,(p+(1-\eta)\,m)+\alpha\mu(v_1)\,(p+m) \geq (1-\alpha)\,\sigma(0)\,(p+(1-\eta)\,m)+\alpha\mu(0)\,(p+m) \tag{3.2}$$

The left-hand side of this inequality determines the firm's profit in equilibrium when $\theta = 1$. This is comprised of three terms. Sophisticated consumers learn its type based on the visibility of its policy, and therefore a share $\sigma(1)$ of them buy the firm's product, earning revenue $(1-\alpha)\,\sigma(1)\,p$, since the total mass of sophisticated consumers is $1-\alpha$. In addition, a share $(1-\eta)$ of these consumers do not hide their data, generating additional revenue $(1-\alpha)\,\sigma(1)\,(1-\eta)\,m$ from consumer data for the firm. Finally, a share $\mu(v_1)$ of myopic consumers buy the firm's product based on the equilibrium visibility v_1 of its privacy policy, generating revenue $\alpha\mu(v_1)\,(p+m)$. The right-hand side of the inequality evaluates the same revenues when the firm deviates to $v = 0$: Given this deviation, sophisticated consumers believe that $v = 0$, so that the share of $\sigma(0)$ of them choose to buy, while the share of myopic consumers that buy drops to

$\mu(0)$.

On the other hand, a firm of type $\theta = 0$ must not want to deviate to a visible policy with $v = v_1$. This constraint translates to

$$(1-\alpha)\,\sigma(0)\,p + \alpha\mu(0)\,p \geq (1-\alpha)\,\sigma(1)\,p + \alpha\mu(v_1)\,p. \tag{3.3}$$

The key difference to constraint (3.2) is that the firm with type $\theta = 0$ does not earn revenue from monetizing data. Hence, the revenue from each consumer who buys the firm's product is simply the product price p.

The following proposition establishes that it is possible to find a value v_1 of visibility for data-extracting firms such that both constraints are satisfied and, therefore, that a separating equilibrium exists.

Proposition 1. *There exists a range of values $v_1 \in [v^\star, v^{\star\star}]$ such that there exists a separating equilibrium in which firms that do not extract data (type $\theta = 0$) choose visibility $v = 0$, and in which firms that extract data (type $\theta = 1$) choose visibility v_1. Moreover, $v^\star$ is defined as the smallest value of v_1 satisfying (3.3).*

The proof makes use of the standard "single crossing" property in signaling games. Indeed, the key property of our game is that the marginal rate of substitution, in terms of the firm's revenue, between sophisticated and myopic consumers is smaller for firms who extract data. The MRS is equal to one for non-extractors with $\theta = 0$, since both types of consumer earn the firm revenues equal to the product price p. However, the MRS for data-extracting firms with $\theta = 1$ is equal to $\frac{p+(1-\eta)m}{p+m} < 1$. Intuitively, firms who extract data are more willing to give up sophisticated consumers in order to gain myopic ones, because myopic ones never hide their data. This feature means that non-extracting firms are willing to give up some myopic consumers by reducing the visibility of their privacy policy, because doing so credibly signals no data extraction

to sophisticated consumers. The assumption that sophisticated consumers hide their data with probability $\eta > 0$ is key in this context.

Moreover, the proposition makes use of the boundary condition (3.1). Intuitively, the boundary condition ensures that signaling (in the sense of low visibility v) is sufficiently costly to support separation: For large values of v, the gains in myopic consumers that the firm experiences by choosing visibility v are never enough to offset the potential gains in sophisticated ones that it can gain by signaling that $\theta = 0$.

3.6.3 Empirical Predictions

We close this section by discussing the relationship between the model's predictions and our empirical findings along three dimensions. First, the separating equilibria characterized above clearly predict that firms' data extraction activities are positively correlated with the visibility of their privacy policies. This is consistent with our observations in the data. In this section, we discuss the implications of the model for two further sets of our empirical results.

Second, we show empirically in the body of this paper that there is a positive correlation between data extraction and length and complexity of privacy policies. The framework above does not explicitly feature policy content as a choice variable for the firm. One can rationalize these patterns by means of the following simple extension: Assume, in addition that firms' legal teams ensure that, given a certain amount of data extraction, the content of the privacy policy is just detailed enough to cover it against the risk of future litigation. This assumption generates a mechanical, positive correlation between the firm's data extraction type θ and the length and complexity of its policy, which is consistent in the data. One possible complication arises in this context: If there is a mechanical relationship between data extraction and policy con-

tent, then sophisticated consumers may be able to infer the firm's data extraction type θ directly from the content of its policy, thus obviating the need for the firm to signal its type to sophisticated consumers via its choice of visibility. On balance, in order to generate a positive correlation between data extraction and *both* length and visibility of privacy policies, one needs to assume that at least some consumers, even among the sophisticated ones, do not bother reading the content of privacy policies. This additional assumption is in line with the empirical evidence that very few consumers read privacy policies, and that the cost of doing so is often prohibitive (e.g., McDonald and Cranor, 2008; Acquisti, Brandimarte, and Loewenstein, 2015).

Third, we show that there is a nonlinear relationship between firms' data extraction and policy visibility, on one hand, and their research-generated capital as measured by Peters and Taylor (2017), on the other hand, which we use as a proxy for firms' technical sophistication. This finding is consistent with the model if we allow for a more flexible interpretation of the firm's data extraction type θ. In particular, the "low extraction type" in the model ($\theta = 0$) describes a firm that cannot earn any marginal revenue by extracting consumer data. As we argued above, this type can capture either unsophisticated, who have a technologically low marginal value of using consumer data, or very sophisticated firms, who have a low marginal value because they already have access to abundant data regardless of consumers' actions. Both scenarios yield the same objective function for the firm, which effectively does not care about acquiring consumer data when $\theta = 0$. Thus, one can re-interpret the low extraction type $\theta = 0$ as a pool of low- and high-sophistication firms. By contrast, the high extraction type $\theta = 1$ captures firms with intermediate technical sophistication, who are able to extract consumer data and do not already have it. With this re-interpretation in hand, the model is consistent with the nonlinearity we find in the data. The interesting insight from this discussion is that, in order to rationalize the patterns in the data,

one needs to assume that there are multiple dimensions of technical sophistication, for example, the firms' ability to extract a flow of consumer data, as well as its existing stock of data.

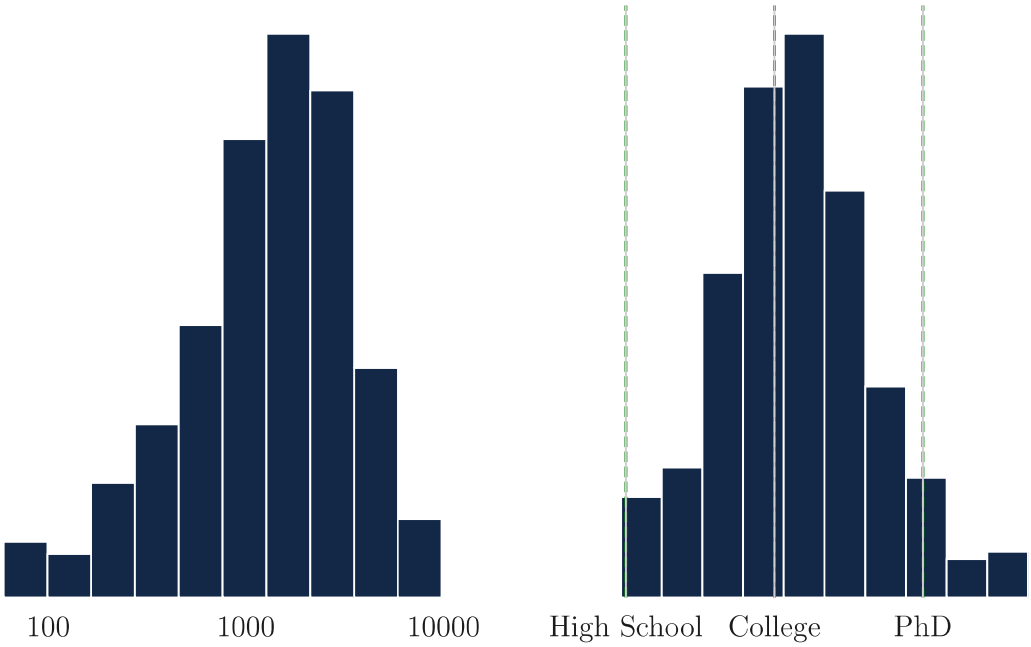
3.7 Conclusion

By analyzing a large set of US firms' privacy policies, we bring new facts and analysis to the study of the market for data privacy. There is significant variation in the ease of acquiring and finding firms' privacy policies, and when found, these policies do not follow a standard boilerplate. Instead, their text varies substantially both within and across industries. Perhaps surprisingly, the visibility, length, and legal clarity of policies all increase in a correlated manner with the extent to which firms extract data from consumers browsing their websites. And finally, we find that the variation in policy text is systematic, with large firms and those with high levels of knowledge capital exhibiting longer and more complex policies with ostensibly more clearly specified and legally sound protections outlined in their text. Our preferred interpretation of these facts, which we represent with a simple model, is that firms' privacy policies are a credible signal of their consumer data extraction intentions. In a simple model with "myopic" consumers that are reassured by the very presence of privacy policy text alongside more sophisticated consumers, high extraction firms find it differentially more costly to omit displaying their privacy policies, with the consequent risk of myopic consumers eschewing their websites. We view our findings in this paper as a first step towards a broader and deeper empirical and theoretical analysis of data privacy policies, and firms' data extraction behaviour, and we hope to continue to refine our insights about this important area going forward.

Figure 3.1. Word Cloud of the Privacy Policy Sample



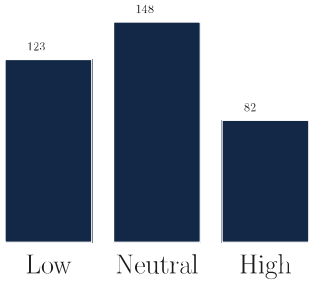
Note: Bigrams are scaled by their average TF.IDF score across all privacy policies in our sample ($N = 4078$).



(a) Word Count **(b)** Gunning Fox Index

Figure 3.2. Distribution of Privacy Policy Attributes

Note: See Table 3.2 for variable definitions. The labels on the horizontal axes in panel (b) mark the heuristic amounts of formal education needed to read documents, where we have coded the corresponding values of the Fog index as High School = 12, College = 17 and PhD = 22.



(a) Distribution of Overall Score



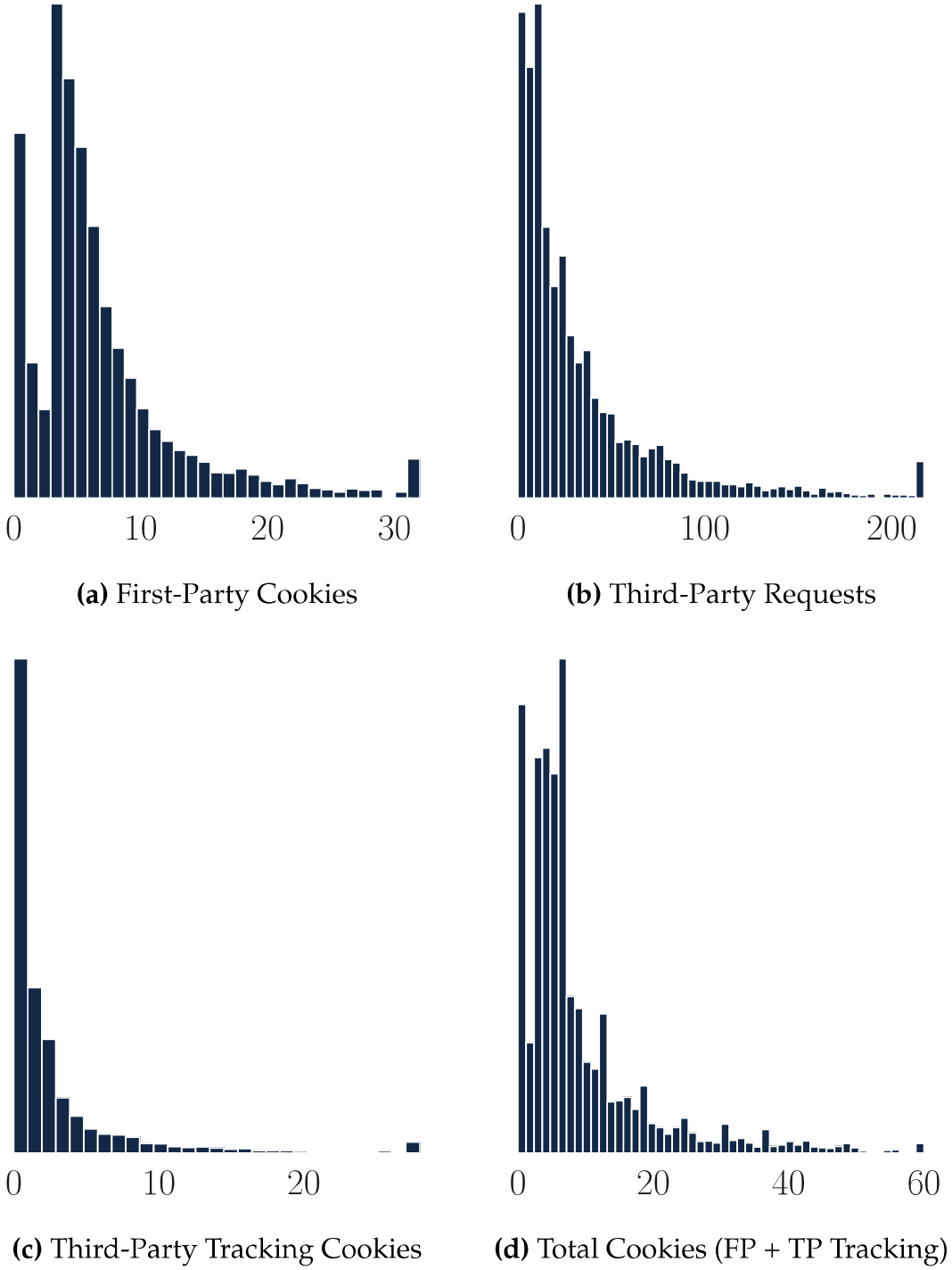
(b) High overall score



(c) Low overall score

Figure 3.3. Expert Evaluation of Privacy Policies

Panel (a) shows the distribution of overall ratings of legal sophistication among policies in the legal expert's sample ($N = 353$). Panels (b) and (c) show the word clouds of policies with high and low Overall ratings respectively. Bigrams are scaled by the difference between the average TF.IDF score among policies evaluated as high (low) by the legal expert, and the grand average TF.IDF score in the sample of policies evaluated by the legal expert.



(a) First-Party Cookies

(b) Third-Party Requests

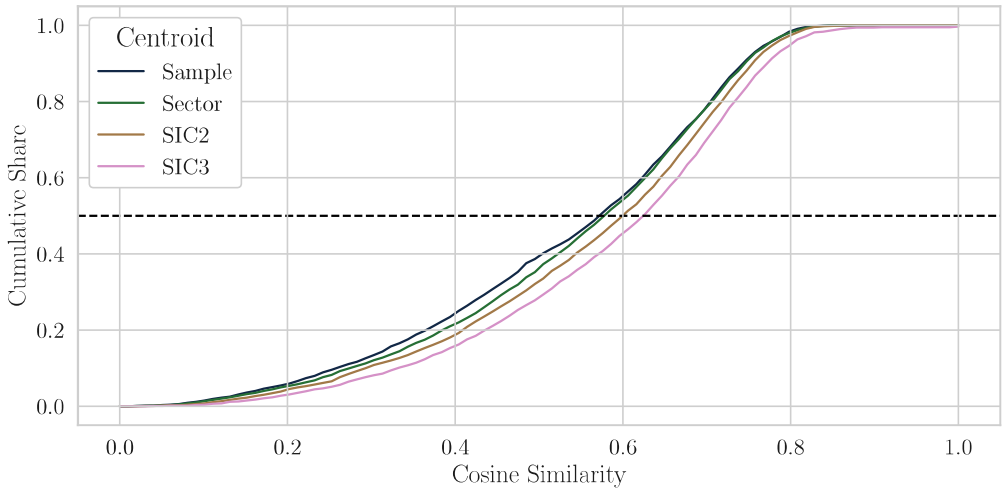
(c) Third-Party Tracking Cookies

(d) Total Cookies (FP + TP Tracking)

Figure 3.4. Distribution of Data Extraction Measures

Data extraction measures are obtained from privacyscore.org, which uses the openWPM web measurement method. For details on openWPM, see Englehardt and Narayanan (2016) and https://webtap.princeton.edu/software/. We winsorize the extraction measures at their first and 99th percentile.

(a) Cumulative Distributions of Cosine Similarities



(b) Mean Cosine Similarities

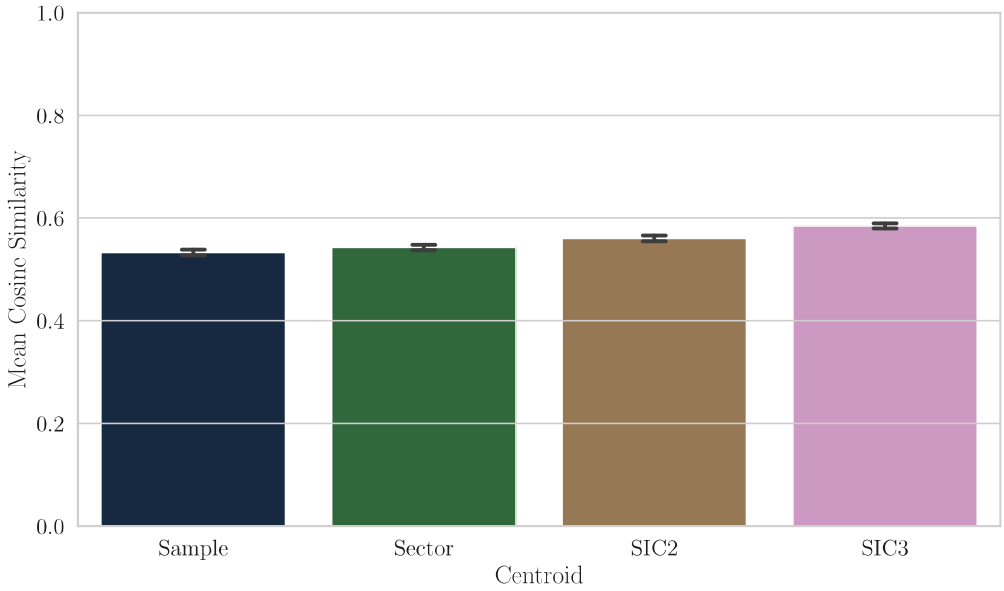
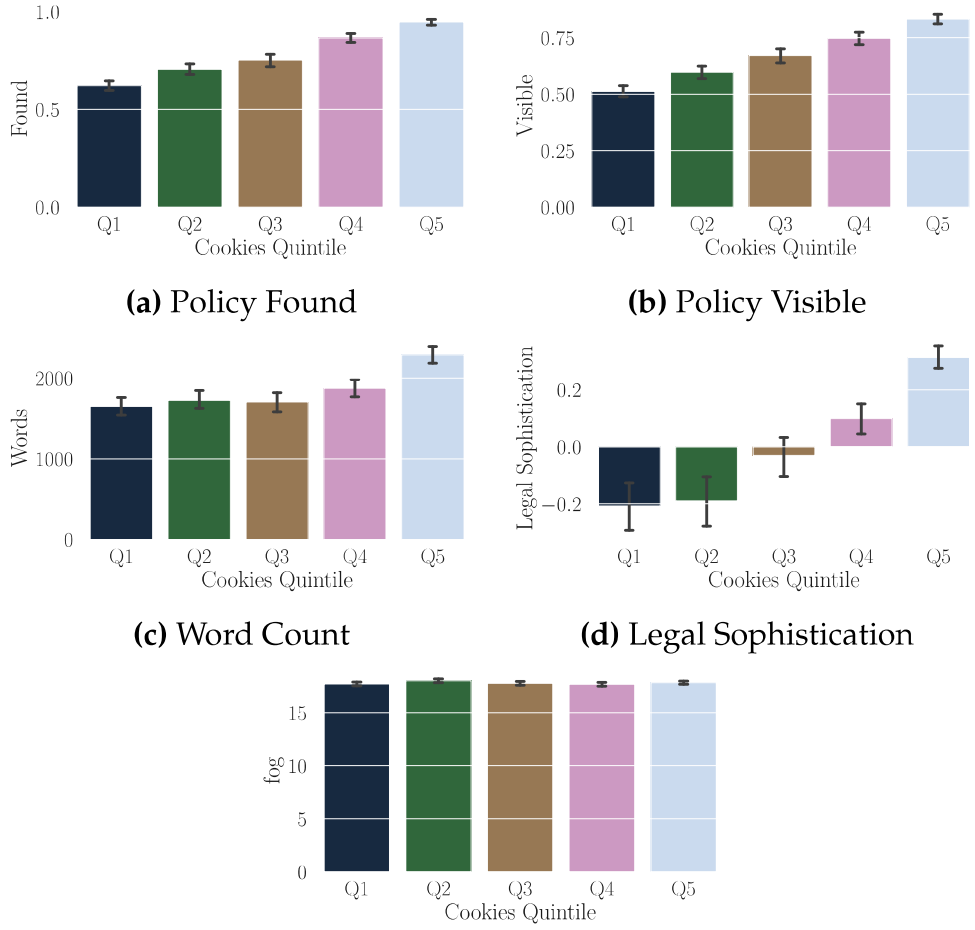


Figure 3.5. Variation in Privacy Policy Text

The Sample centroid is the mean TF-IDF frequency vector across all 4078 privacy policies. Sector centroids are the mean TF-IDF frequency vectors in the 12 SIC divisions, which are Agriculture, Forestry and Fishing (SIC 0100-0999), Mining (SIC 1000-1499), Construction (SIC 1500-1799), Manufacturing (SIC 2000-3999), Transport, Communications, and Utilities (SIC 4000-4999), Wholesale Trade (SIC 5000-5199), Retail Trade (SIC 5200-5999), Finance, Insurance, and Real Estate (SIC 6000-6799), Services (7000-8999), Public Admin (SIC 9100-9729), and Nonclassificable (9900,9999). SIC2 and SIC3 centroids are mean frequencies at the 2-digit and 3-digit SIC code level, respectively.



(a) Policy Found

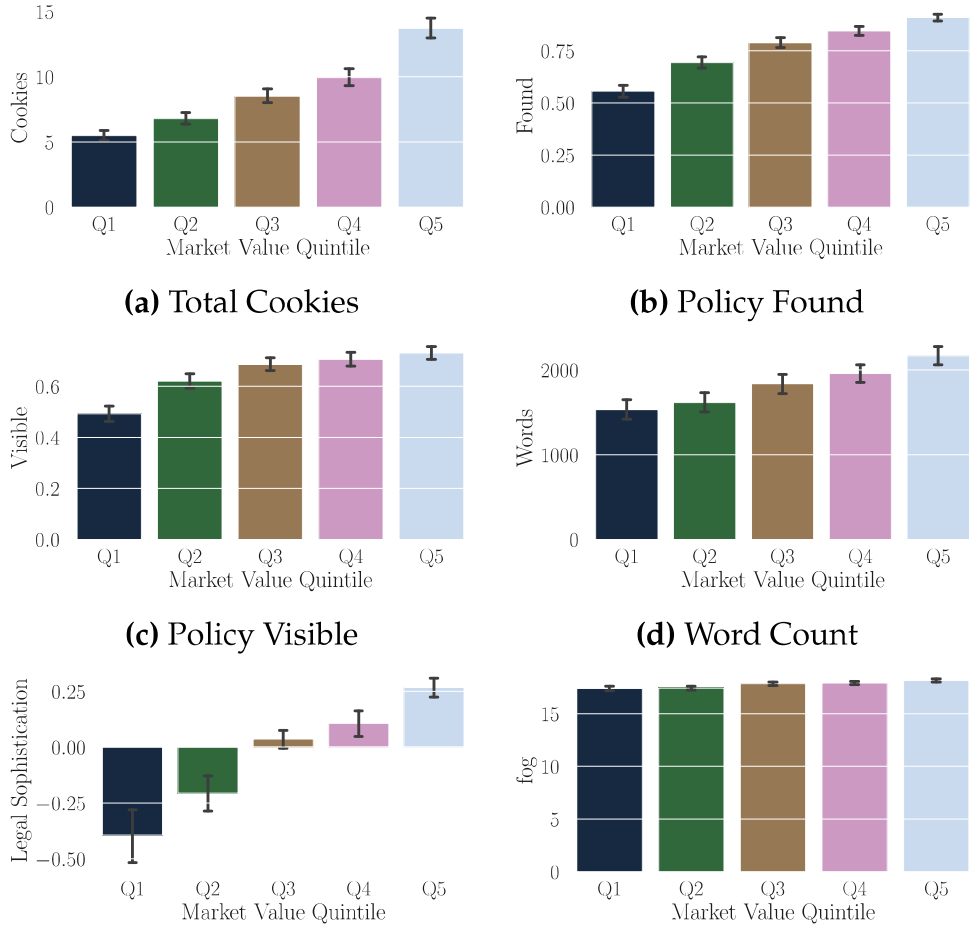
(b) Policy Visible

(c) Word Count

(d) Legal Sophistication

(e) Gunning Fog index

Figure 3.6. Average Attributes of Privacy Policies by Quintile of Total Cookies



(a) Total Cookies

(b) Policy Found

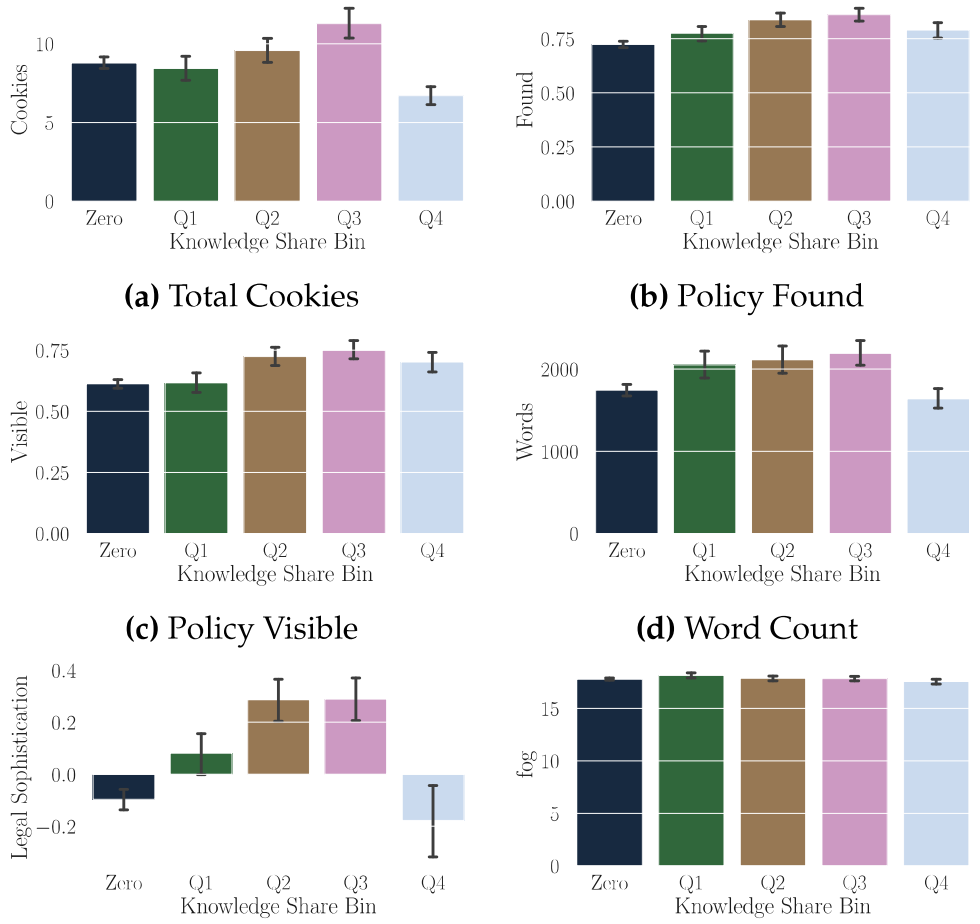
(c) Policy Visible

(d) Word Count

(e) Legal Sophistication

(f) Gunning Fog index

Figure 3.7. Average Attributes of Privacy Policies by Firm Size



(a) Total Cookies

(b) Policy Found

(c) Policy Visible

(d) Word Count

(e) Legal Sophistication

(f) Gunning Fog index

Figure 3.8. Average Attributes of Privacy Policies by Knowledge Share

Table 3.1. Descriptive Statistics of Firm Characteristics

	count	mean	median	std
Market Value	5,377	5.693	0.665	15.848
Market to book	5,345	2.207	1.408	2.216
Market Share	5,377	0.011	0.001	0.032
Intangible Share	5,306	0.159	0.043	0.216
Knowledge Share	5,140	0.079	0.000	0.160
R&D to Assets	2,621	0.029	0.010	0.051
Marketing to Assets	2,029	0.022	0.005	0.043

Note: Market value is measured in millions of USD. Market Share is the firm's sales divided by industry sales at the 2-digit SIC code level. Intangible Share is intangible assets divided by total assets, and Knowledge Share is the replacement value of knowledge capital divided by the sum of intangible assets and the replacement values of knowledge and organizational capital (see Peters and Taylor (2017)). We winsorize all variables at their first and 99th percentile.

Table 3.2. Descriptive Statistics of Privacy Policy Attributes

	count	mean	median	std
Policy Found	5,377	0.758	1.000	0.428
"Privacy" Visible on Homepage	5,377	0.647	1.000	0.478
Number of Paragraphs	4,078	30.972	23.000	27.134
Number of Words	4,078	1,858.741	1,433.000	1,645.912
Gunning Fog Index	4,078	17.792	17.695	2.579

Policy Found is equal to one if we found a privacy policy by automated scraping or manual collection. "Privacy" Visible on Homepage is equal to one if the firm's root web domain, as recorded in Compustat, contains a link with the word "privacy". For definitions of the Fog index, see Gunning (1952). We compute the Fog index using the "readcalc" package in Python (https://pypi.org/project/ReadabilityCalculator/). We winsorize the length variables and the Fog index at their first and 99th percentile.

Table 3.3. Descriptive Statistics of Data Extraction Measures

	count	mean	median	std
First Party Cookies	5,216	6.446	5.000	6.142
Third Party Requests	5,216	33.758	19.000	40.366
Third Party Tracking Cookies	5,216	2.446	1.000	4.802
Total Cookies	5,216	8.892	6.000	9.793

Data extraction measures are obtained from privacyscore.org, which uses the openWPM web measurement method. For details on openWPM, see Englehardt and Narayanan (2016) and https://webtap.princeton.edu/software/. We winsorize the extraction measures at their first and 99th percentile.

Table 3.4. Policy Attributes: Regressions

(a) Without Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.042***	0.048***	-0.006	0.043***	0.030	0.857***
	(12.22)	(12.13)	(-0.61)	(4.44)	(1.14)	(10.86)
Knowledge Share	0.847***	0.695***	2.405***	2.605***	0.501	14.562***
	(8.33)	(5.89)	(8.80)	(9.78)	(0.69)	(6.25)
Knowledge Share^2	-0.813***	-0.793***	-2.821***	-3.811***	-0.264	-22.301***
	(-4.90)	(-4.12)	(-6.30)	(-8.74)	(-0.22)	(-5.87)
Log Market Share	0.016***	-0.010***	0.087***	0.061***	0.100***	0.316***
	(5.41)	(-3.11)	(10.49)	(7.57)	(4.54)	(4.75)
Observations	5140	5140	3918	3918	3918	5140

(b) With Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.043***	0.042***	0.016	0.042*	0.054	0.880**
	(8.31)	(5.53)	(0.75)	(1.88)	(0.76)	(2.53)
Knowledge Share	0.659***	0.463**	1.502**	2.260***	0.759	14.321**
	(12.66)	(3.08)	(2.60)	(6.34)	(0.42)	(2.58)
Knowledge Share^2	-0.580***	-0.482***	-1.795**	-3.283***	-0.299	-20.522**
	(-12.91)	(-5.31)	(-2.75)	(-6.20)	(-0.18)	(-2.43)
Log Market Share	0.014**	-0.005	0.062*	0.062***	0.088	0.261
	(2.48)	(-0.41)	(2.14)	(5.73)	(0.78)	(1.13)
Observations	5140	5140	3918	3918	3918	5140

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Appendix A

Retail Investor Heterogeneity: Evidence from WallStreetBets

A.1 Retail Investors in Academic Research

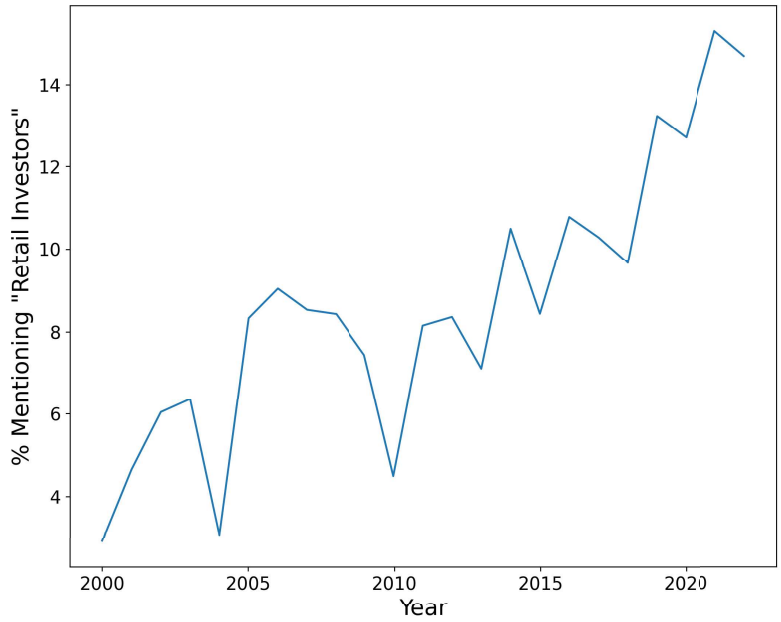


Figure A.1. Retail Investors in Academic Research

This figure plots the percentage of academic papers published in the Journal of Finance, Journal of Financial Economics, Quarterly Journal of Economics, and Review of Financial Studies which mention the bigram "Retail Investors" at least once.

A.2 Tickers removed

No.	Ticker	Latest Company/Fund	In CRSP?	Reason (if not obvious)
1	AI	Arlington Asset Investment	Yes	-
2	ATH	Athabasca Oil Corp	Yes	Accronym for "All Time High"
3	API	Agora Inc	Yes	Accronym for "Application Programming Interface"
4	BEAT	Heartbeam Inc	Yes	-
5	BTC	ETF Series Solutions	Yes	Also the ticker of Bitcoin
6	CEO	CNOOC	Yes	-
7	CDC	Victory Portfolios	Yes	-
8	COLD	Americold Realty Trust	Yes	-
9	DATA	Tableau Sotfware	Yes	-
10	DCF	BNY Mellon Fund	Yes	-
11	DD	DuPont	Yes	Accronym for "Due Diligence"
12	DM	Desktop Metal	Yes	Accronym for "Direct Message"
13	GDP	Goodrich Petroleum	Yes	-
14	GF	New Germany Fund	Yes	-
15	GOOD	Glastone Commercial	Yes	-
16	GOV	Government Properties Income	Yes	-
17	EOD	Allsrping Global	Yes	Accronym for "End Of Day"
18	ETH	Ethan Allen Interiors	Yes	Also the ticker of Ethereum
19	EV	Eaton Vance	Yes	Accronym for "Electric Vehicles"
20	HF	HFF	Yes	Accronym for "Hedge Funds"
21	IP	International Paper	Yes	Accronym for "Intellectual Property"
22	IPO	Renaissance Capital	Yes	-
23	IQ	IQIYI	Yes	-
24	IRS	IRSA Inversiones	Yes	-
25	IV	Mark IV Industries	Yes	
26	JAN	Jantzen	Yes	Ususually refers to January
27	LIFE	Atyr Pharma	Yes	-
28	LTD	L Brands	Yes	-
29	PSA	Public Storage	Yes	Accronym for "Public Service Announcement"
30	PE	Parsley Energy	Yes	
31	NYC	NYC Reit	Yes	-
32	MRNA	Moderna	Yes	Usually refers to the vaccine type
33	PS	PluralSight	Yes	-
34	RH	RH	Yes	Usually refers to Robinhood
35	ROI	WisdomTree Trust	Yes	Accronym for "Return on Investment"
36	SF	Stifel Financial Corp	Yes	-
37	TA	TravelCenters America	Yes	Accronym for "Technical Analysis"
38	TV	Grupo Televisa	Yes	-
39	UI	Ubiquiti	Yes	Accronym for "User Interface"
40	UK	Ucommune International	Yes	-
41	USA	Liberty All Star Equity	Yes	-
42	WIN	Windstream Holdings	Yes	-
43	WOW	WideOpenWest	Yes	-
44	WSB	W S B Holdings	Yes	-
45	YOLO	Advisors Shares (Cannabis ETF)	Yes	-

A.3 Sentiment

OpenAI offers four different GPT-3 models with increasing complexity and cost. In order of complexity, their names are Ada, Babbage, Curie and Davinci. Figure A.2 plots the out of sample accuracy of the three least expensive models (omitting Davinci) against the training size given to each of them. In order to create the training and validation set, I randomly selected a number of posts, settling on 500 of them, being careful to undersample positive posts, as they make the bulk of the sample. Finally, I coded them as either positive, neutral or negative. Giving to OpenAI a training set with x observations would therefore leave me with $500 - x$ observations for the validation set, thus reducing the validation set as I increase the training set. All three models perform reasonably well at all training size, even when giving only 10 training observations (4 positive, 3 negative, and 3 neutral). It is also noteworthy that the models hardly improve in accuracy once they are given more than 100 training observations, therefore providing further evidence that those models are indeed few short learners (Brown, Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, Herbert-Voss, Krueger, Henighan, Child, Ramesh, Ziegler, Wu, Winter, Hesse, Chen, Sigler, Litwin, Gray, Chess, Clark, Berner, McCandlish, Radford, Sutskever, and Amodei (2020)). Additionally, the models' accuracy increase with their complexity, with Curie performing the best, and Ada the worst.

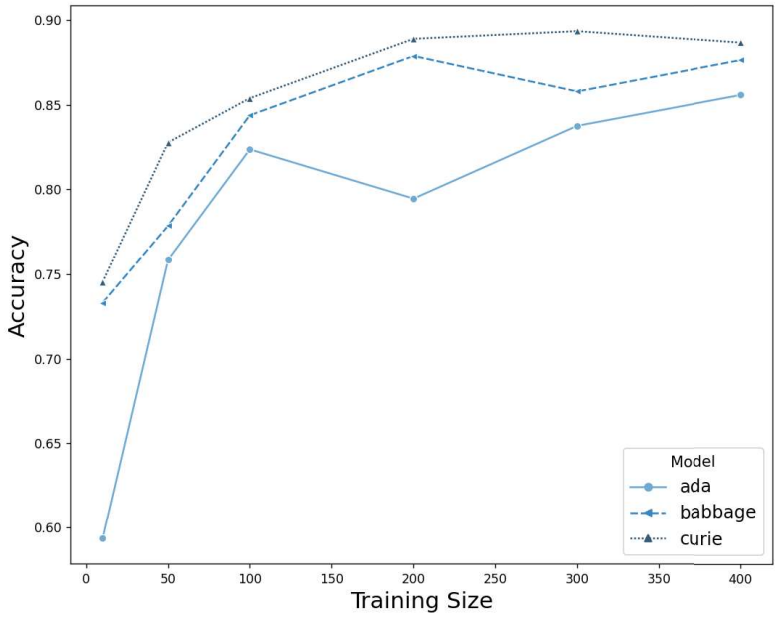


Figure A.2. Accuracy of OpenAI GPT-3 Models

This figure plots the accuracy of three GPT-3 models (in order of complexity: ada, babbage, and curie) for different training sizes x, ranging from 10 to 400 examples. The test sample consists in 500-x examples.

Given the good performance of Babbage[1], this is the model I select. Figure A.3 plots the term frequency-inverse document frequency (TF-IDF)[2] word clouds of posts with positive and negative labels with the words' color and size increasing with their TF-IDF. Words appearing more often in positive posts include some seen during the GME episode ("Hold", "[to the]Moon", "YOLO", ...) as well as more standard ones ("Calls", "Buy", "Long", ...). On the other hand, words associated with negative posts seem more standard, with "puts" and "shorts" being the most important. It is noteworthy however that many words - such as "Buy", or "Calls" for example - do appear signif-

[1] And its reasonable cost

[2] Once the stop words are removed, I divide the number of occurrences of any given word by the total number of words in the post, hence getting the term frequency (TF). I then multiply this TF by the inverse document frequency (IDF) which penalizes words which appear in many posts. The IDF is simply calculated as log of the inverse fraction of posts containing the word. Let P_{ij} be the number of times that the word j appears in post i and N the number of posts, then: $TFIDF_{i,j} = \frac{P_{i,j}}{\sum_k P_{i,k}} \cdot log\left[\frac{N}{\sum_n 1_{P_{n,j}>0}}\right]$

icantly in both positive and negative posts, further casting doubts as to the efficiency of a dictionary approach.



Figure A.3. TF-IDF of Posts' text with Positive and Negative Sentiment

Those world clouds represent the TF-IDF of words in posts either defined as positive or negative by the GPT-3 model.

A.4 Factor Model

A.4.1 Traditional Factor Model

In the case of a traditional factor model, one could, for each factor (k), regress:

$$Q_{i,n,t} = \beta_i^k C_{n,t}^k + \epsilon_{i,n,t}$$

Where:

- $Q_{i,n,t}$ is the overall talking share for stock n for individual i at time t. For example, if individual i talks about Apple three times on 01/02/2019 and AMC once on 03/17/2020 then:
  - $Q_{i,AAPL,01/02/2019} = 0.75$

  - $Q_{i,AMC,03/17/2020} = 0.25$
  - $\forall t, \forall n \notin [AAPL, AMC], Q_{i,n,t} = 0$

- $C^k_{n,t}$ is the characteristic k for stock stock n at time t, with:

  - $E(C^k_{n,t}) = 0$ because for each time t, the average characteristic is zero (as they are ranked and rescaled over $[-0.5, 0.5]$).
  - $Var(C^k_{n,t}) \approx 1/12$ because of the cross-ranking (uniform distribution over $[-0.5, 0.5]$)

Additionally note that $\forall i, \sum_{n=0}^{N} \sum_{t=0}^{T} Q_{i,n,t} = 1$ (the talking shares for any individual sum to 1).

Then:

$$
\begin{aligned}
\beta^k_i &= \frac{E[(Q_{i,n,t} - E[Q_{i,n,t}])\,(C^k_{n,t} - E[C^k_{n,t}])]}{Var(C^k_{n,t})} \\
&= 12 \cdot E[(Q_{i,n,t} - E[Q_{i,n,t}])C^k_{n,t}] \\
\hat{\beta}^k_i &= 12 \cdot \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N}\sum_{t=0}^{T}[(Q_{i,n,t} - \frac{1}{N}\frac{1}{T}\sum_{n=0}^{N}\sum_{t=0}^{T} Q_{i,n,t})C^k_{n,t}] \\
&= 12 \cdot \frac{1}{N \cdot T}\sum_{n=0}^{N}\sum_{t=0}^{T}\left[\left(Q_{i,n,t} - \frac{1}{N \cdot T}\right)C^k_{n,t}\right] \\
&\propto \sum_{n=0}^{N}\sum_{t=0}^{T}\left[\left(Q_{i,n,t} - \frac{1}{N \cdot T}\right)C^k_{n,t}\right]
\end{aligned}
\tag{A.1}
$$

A.4.2 This Model

In this model, there is a slight twist: the betas are calculated in the following manner:

$$\begin{aligned}\hat{\beta}_i^k &= \frac{1}{N \cdot T} \sum_{n=0}^{N} \sum_{t=0}^{T} \left[Q_{i,n,t} \cdot C_{n,t}^k - \frac{1}{I} \sum_{i=0}^{I} Q_{i,n,t} \cdot C_{n,t}^k \right] \\ &= \frac{1}{N \cdot T} \sum_{n=0}^{N} \sum_{t=0}^{T} \left[\left(Q_{i,n,t} - \frac{1}{I} \sum_{i=0}^{I} Q_{i,n,t} \right) \cdot C_{n,t}^K \right] \\ &\propto \sum_{n=0}^{N} \sum_{t=0}^{T} \left[\left(Q_{i,n,t} - \frac{1}{I} \sum_{i=0}^{I} Q_{i,n,t} \right) \cdot C_{n,t}^K \right] \end{aligned} \tag{A.2}$$

This is very similar to the previous equation, with the only difference being how the average $Q_{i,n,t}$ is calculated:

- Here, the average is taken over all individuals I on the day in question, t.
- Previously, the average was taken with respect to individual i over all stocks N and days T (and therefore equal to $1/(NT)$).

A.4.3 Why it matters

The expectation is no longer that individuals should talk about all stocks equally, it is that they should talk about each stock as much as their peers. This is a better expectation for - at least - the following two reasons. First, expecting Redditors to talk about stocks equally is an unreasonable assumption, most of the literature on retail investors points towards biases, and more particularly attention biases. We should therefore expect few stocks to gather most of the attention in any given day, and this is very much the case as shown in Figure 1.2. Second, the overall question this paper asks is whether some retail investors are different from the rest of their peers and equation A.2 gets to

the very heart of that.

In connection with this later point, working with a traditional factor model would cause additional issues; Redditors on WallStreetBets are for example prevented from talking about stocks with a market capitalization under $1B for most of the sample. We should therefore find them to have a bias towards large cap stocks. What is of interest, however, is whether some have an even greater bias towards large cap stocks than the rest of their peers.

A.5 Robustness Checks Correlations & PCA

A.5.1 Stock Level Correlation Matrix

As explained in Section 1.3.4, the correlations showed in Figure 1.8 could be due to correlations at the stock level. Figure A.4 alleviates those concerns, showing a similar correlation matrix, but done at the stock level. As explained in Section 1.3.2, the stock attributes are rescaled to have mean zero and support $[-0.5, 0.5]$. This figure shows that none of the correlations appear, or at the very least appear as strongly, as they do in Figure 1.8.

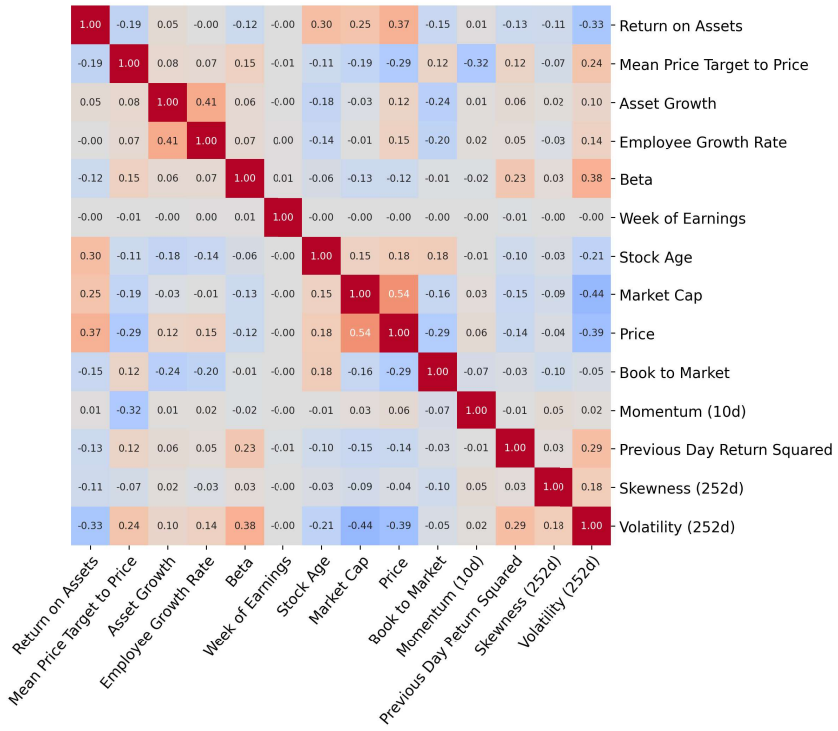


Figure A.4. Stock Level Correlation Matrix

This figure plots correlations between ranked and rescaled (mean zero and support $[-0.5, 0.5]$) selected daily stock attributes. The heat map reflects the intensity of the correlation, with red (blue) signifying positive (negative) correlation.

A.6 First Principal Component

As explained in Section 1.4, one possible concern could be the number of volatility related metrics. Having a large number of correlated metrics could artificially increase the power of explanation of a principal component, and here, in particular the first. In order to confirm that this is not the case here, the principal component analysis was run once again, removing all momentum variables but one, all skewness variables but one, and finally all volatility variables but one. Figure A.5 plots the t-statistics coming from running univariate regressions of all characteristic tilts and account level metrics

on the first principal component. The conclusion remains unchanged.



Figure A.5. First Principal Component Loadings

This figure plots world clouds of variable names based on their t-statistics from a univariate regression on the first principal components. The left panels show variables with a positive loading on the PC of interest while the right panels are those with negative loadings. Both the size and colour intensity of the fonts reflect the relative importance of each variable with the most important ones appearing larger and darker.

A.7 Talking Share Regressions for all 2,926 stocks

In section 1.5, I show the results of the talking share regressions for the top 1,000 stocks (ranked by the number of time they are posted about). Here, I show the same table, but for all 2,926 regressions.

Based on the average absolute values for the t-statistics only three account level variables appear significant at 95% level: the signal stock dummy, the number of stocks individuals talk about and the number of comments. On the characteristic tilt front, all but three of them are. Note, however, that given the number of right hand side variables in the model - twenty five - a 5% significance level may require a t-statistic higher than the traditional 1.96 (at worse, 3.09[3] if all variables are uncorrelated). Furthermore, few of those variables matter for more than a third of the stocks.

However, as explained previously in Section 1.5, the lacklustre significance is not entirely surprising, as the vast majority of stocks are only mentioned a few times.

[3] $CDF^{-1}\left[\left(1-(5\%)/2\right)\right)^{1/25}] = 3.09$

Table A.1. Stock Level Estimates

(a) Panel A: Account Characteristics

	Avg. $\|t-stat\|$	10%	50%	90%	% Significant (5% level)
Single Stock Dummy	5.52	-12.24	-2.78	0.34	55.88
No. Stocks Talked	4.52	-0.96	0.20	10.50	29.53
Reddit Account Age	1.79	-2.07	-0.08	2.59	22.52
No. Posts	1.41	-2.02	0.09	1.47	14.83
No. Comments	3.00	-6.80	-0.25	0.42	22.56
Pct. on Weekends	1.57	-2.00	-0.42	1.89	17.77
Pct. During Trading Hours	1.23	-1.66	-0.08	1.81	13.74
Pct. Entertainment	1.88	-2.60	-0.54	1.89	22.52
Pct. News	1.37	-2.18	-0.48	1.15	16.17
Gunning Fog Index	1.70	-2.13	-0.12	2.54	22.49
Spelling Errors	0.78	-0.53	-0.06	1.38	6.84

(b) Panel B: Characteristic Tilts

	Avg. $\|t-stat\|$	10%	50%	90%	% Significant (5% level)
Stock Age	3.79	-3.59	0.05	4.67	32.54
Return on Assets	3.59	-3.86	-0.01	4.73	33.83
Mean Price Target to Price	3.19	-2.07	0.52	5.63	33.19
Asset Growth (quarterly)	3.33	-3.56	-0.12	3.67	30.72
Book-to-Market	3.78	-3.61	0.18	5.50	34.14
Beta	2.26	-2.43	0.12	3.16	25.73
Week following Earnings	1.94	-1.54	0.01	3.76	22.04
Log Market Cap	4.61	-6.88	-1.04	2.67	43.78
Log Share Price	3.87	-3.40	0.48	5.45	36.16
Momentum (30d)	1.95	-2.97	-0.11	2.28	24.16
Skewness (252d)	3.71	-5.53	-0.27	3.42	35.71
Returns Volatility (252d)	3.79	-5.25	-0.23	3.78	35.44
Previous Day Return Squared	1.76	-3.08	-0.35	1.22	21.33
Earnings Volatility	4.50	-5.69	0.15	6.57	39.51

This table shows the average absolute value, as well as the 10%, 50% and 90% percentiles of the t-statistic coming from 2,926 stock level regressions in which the dependant variable is the talking share Q_i regressing how much an individual i talks about a given stock n, in percentage terms. The regression in question, for each stock n is $Q_{i,n} = \alpha_n + \sum_{k=1}^{K} \beta_{n,k} F_{i,k} + \epsilon_{i,n}$. The last columns shows for which percentage of the regression the variable in question is significant at 5%.

A.8 Returns with quadratics

An alternative specification to the regressions in Table 1.4 would be one in which the number of posts are not subsequently taken the logarithm of, but instead, one could introduce the square so as to capture non linear effects, as the effect is indeed unlikely to be linear. This is done in Table A.2 which, again, shows differential effects for the

different terciles. This is further proof of retail heterogeneity, as the metric by which one should compute aggregates matters.

Table A.2. Return Predictions

	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$
Positive Posts	0.000***	0.000*		
	(0.000)	(0.000)		
Negative Posts	-0.001***	-0.001*		
	(0.000)	(0.000)		
Log Comments	-0.000*	-0.000***		
	(0.000)	(0.000)		
Positive Posts$_{1^{st} \text{ Tercile of } PC_1}$			0.001***	0.001***
			(0.000)	(0.000)
Negative Posts$_{1^{st} \text{ Tercile of } PC_1}$			-0.006***	-0.006*
			(0.001)	(0.002)
Positive Posts$_{2^{nd} \text{ Tercile of } PC_1}$			0.000*	0.000
			(0.000)	(0.000)
Negative Posts$_{2^{nd} \text{ Tercile of } PC_1}$			-0.003**	-0.003*
			(0.001)	(0.001)
Positive Posts$_{3^{rd} \text{ Tercile of } PC_1}$			0.000*	0.000
			(0.000)	(0.000)
Negative Posts$_{3^{rd} \text{ Tercile of } PC_1}$			-0.001*	-0.001
			(0.001)	(0.001)
$r_{i,w}$	-0.041***	-0.041***	-0.040***	-0.040***
	(0.001)	(0.010)	(0.001)	(0.010)
Retail Order Imbalance	0.001*	0.001*	0.001*	0.001*
	(0.000)	(0.001)	(0.000)	(0.001)
Quadratic Controls	Yes	Yes	Yes	Yes
Number of Comments Controls	-	-	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Clustering	-	Firm & Month	-	Firm & Month
Obs	776552	776552	776552	776552
R-squared	0.002	0.002	0.002	0.002

Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$

The dependant variable is the return in the following week, $r_{i,w+1}$. The independent variables are the number of positive, and negative posts, as well as their squares, made in the current week, either fully aggregated or aggregated based on terciles of the individual's principal components. Controls include the number of comments (per type of individual), the returns in the current week, as well as retail order imbalance (Boehmer, Jones, Zhang, and Zhang (2021)). All regressions include firm and time fixed effects. The firm GameStop was removed from those regression.

A.9 Clustering

An alternative to separating individuals based on their first principal component would be to cluster them. For ease of interpretation, I focus on the first two principal components which on their own capture over 35% of the variance, as shown in Figure 1.9a. In spite of this arbitrary limitation, some conclusions can be drawn from this.

The first plot of Figure A.6 shows the density of those first two principal components, with at least three clusters clearly visible: one with positive loading on the first principal component (which captures interest in fundamentals), and two other with negative loadings on the first component but which differentiate themselves with positive (meaning high cash flow volatility) and negative (meaning older, more established stocks) loadings on the second principal component respectively. Since the main objective is to find clusters of individuals, rather than cluster all individuals, I mainly focus on areas of high density, and sub-sample to keep only points within higher density area. This is shown in the second graph of Figure A.6, where I only keep individuals in areas within the top 40% of densities.

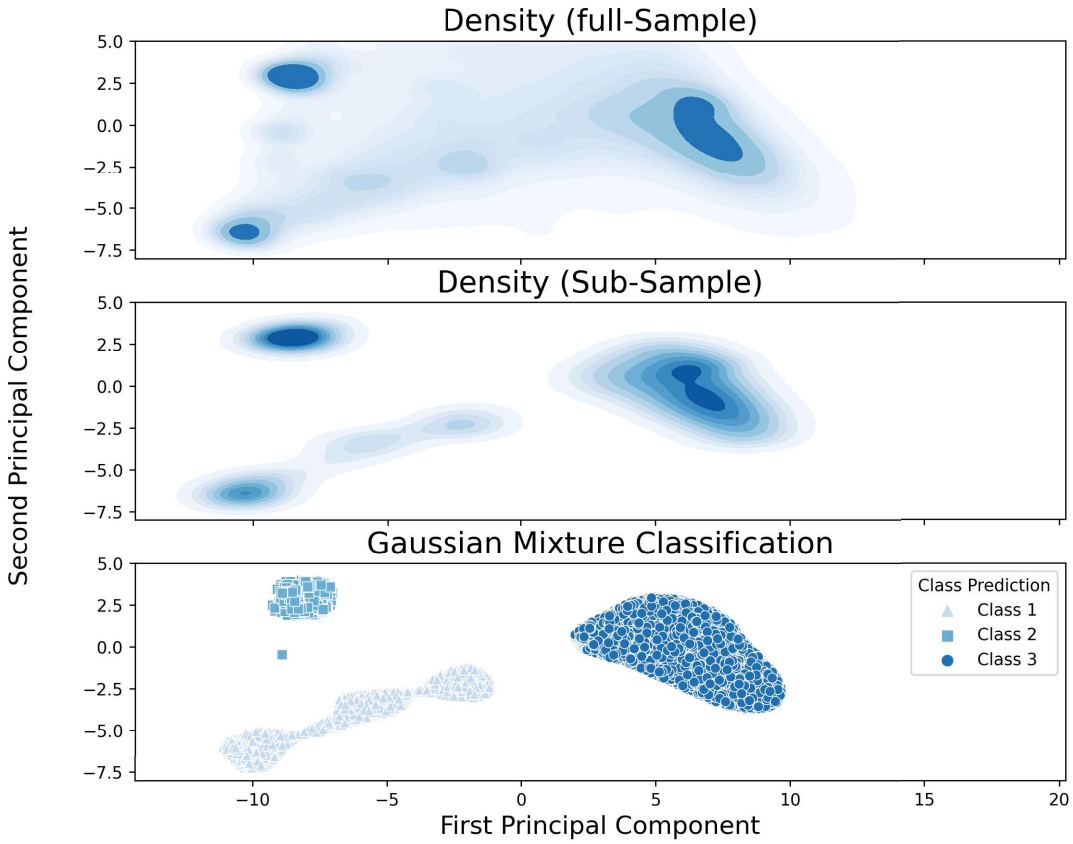


Figure A.6. Classification Result

The first figure show the density of the first principal component with respect to the second for all authors on WallStreetBets, with at the very least three clusters clearly visible. The second shows the same density when keeping only high density areas (top 40%). Finally, the third plot shows the result from applying a Gaussian Mixture Model classification to the observations in the previous figure.

This leaves three separate clusters which are easily detected by a Gaussian mixture model (GMM) which assumes that the data points are generated from a finite number of Gaussian distributions. To confirm that indeed those are separate clusters, I run nine different GMM on a training set, each with a different number of cluster - from one to nine - and calculate the likelihood on a separate test set. The results are showed in Figure A.7 with a clear kink at three clusters which are drown on the third plot of Figure A.6.

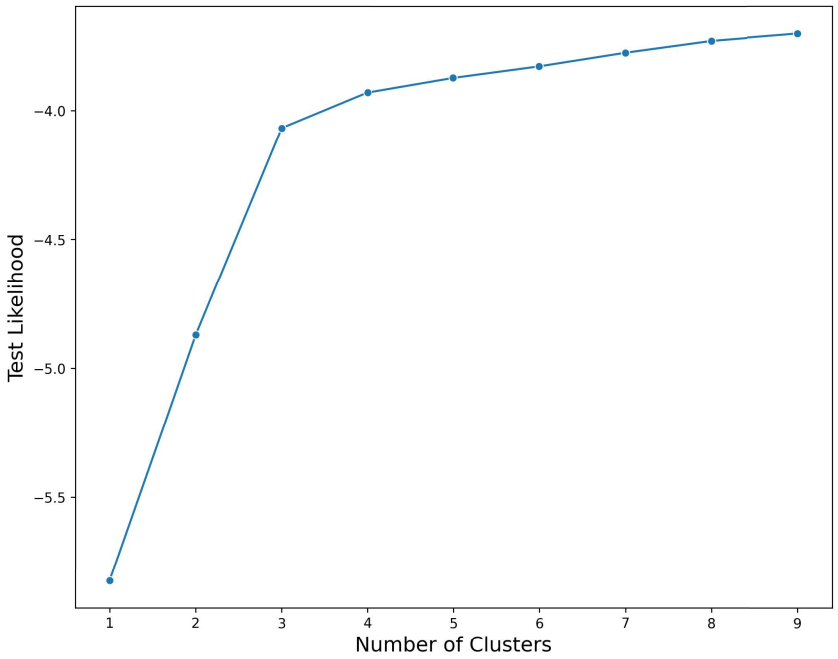


Figure A.7. Gaussian Mixture Test Likelihood

This figure plots the test likelihood for observations in an out of sample test set against the number of clusters selected for classification. The results show a significant decrease in the reduction of the test error after three clusters, thus suggesting this number as optimal.

Table A.3 shows the returns of the different clusters, with once again differences in performance across clusters. The third columns shows that first cluster is particularly useful in a prediction exercise when talking positively about a stock, three-times more than individuals in either the second or third cluster. Additionally, individuals in the second cluster appear particularly good at talking about stocks of which the returns will be negative in the following week: almost twice as much as individuals in the third cluster, with the effects remaining strongly significant when clustering the standard errors at the firm and months levels.

Table A.3. Clusters Return Predictions

	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$	$r_{i,w+1}$
Log Positive Posts	0.003***	0.003		
	(0.001)	(0.002)		
Log Negative Posts	-0.003*	-0.003		
	(0.001)	(0.002)		
Log Comments	-0.001***	-0.001		
	(0.000)	(0.001)		
Cluster$_1$ *Log Positive Posts*			0.013***	0.013
			(0.003)	(0.012)
Cluster$_1$*Log Negative Posts*			0.015*	0.015
			(0.006)	(0.015)
Cluster$_2$ *Log Positive Posts*			0.003	0.003
			(0.008)	(0.008)
Cluster$_2$ *Log Negative Posts*			-0.073***	-0.073**
			(0.013)	(0.025)
Cluster$_3$ *Log Positive Posts*			0.004***	0.004
			(0.001)	(0.002)
Cluster$_3$ *Log Negative Posts*			-0.004***	-0.004**
			(0.001)	(0.002)
$r_{i,w}$	-0.041***	-0.041***	-0.040***	-0.040***
	(0.001)	(0.010)	(0.001)	(0.010)
Retail Order Imbalance	0.001*	0.001*	0.001*	0.001*
	(0.000)	(0.001)	(0.000)	(0.001)
Number of Comments Controls	-	-	Yes	Yes
Time Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	Yes	Yes	Yes	Yes
Clustering	-	Firm & Month	-	Firm & Month
Obs	776552	776552	776552	776552
R-squared	0.002	0.002	0.002	0.002

Standard errors in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$

The dependant variable is the return in the following week, $r_{i,w+1}$. The independent variables are the log number of positive and negative posts (to which a 1 was added) made in the current week, either fully aggregated or aggregated based on the cluster to which the individual belongs. Controls include the number of comments (per type of individual), the returns in the current week, as well as retail order imbalance (Boehmer, Jones, Zhang, and Zhang (2021)). All regressions include firm and time fixed effects. The firm GameStop was removed from those regression.

A.10 Reversal Strategy Returns

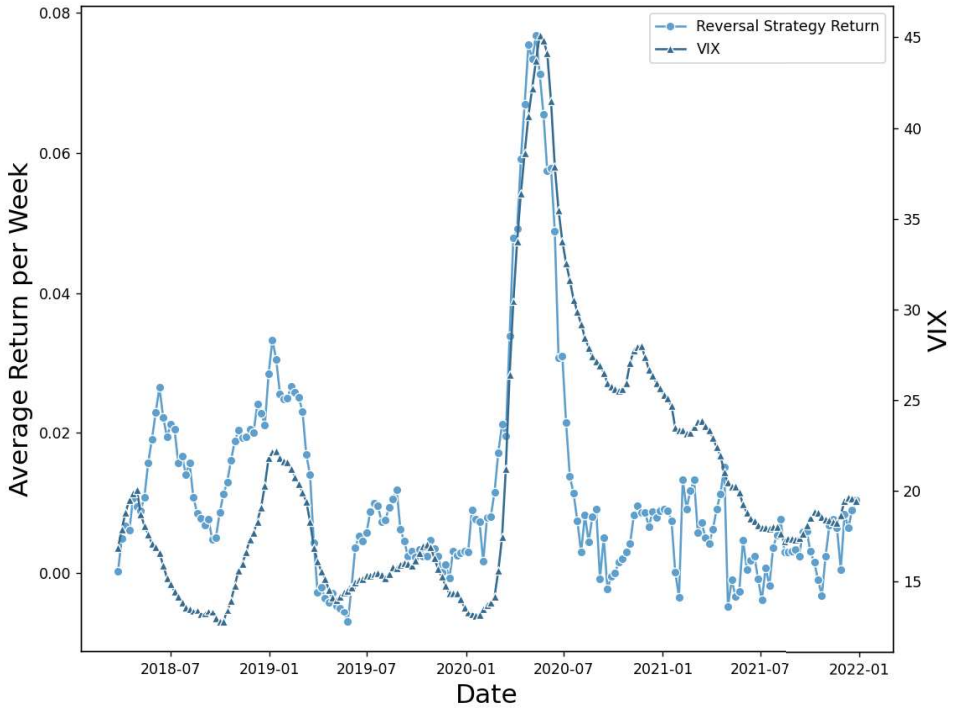


Figure A.8. Weekly return reversal strategy and VIX

This figure shows the 3-month moving average of the returns of a reversal strategy similar to Nagel (2012), against the 3-month moving average of the VIX. Each week, the reversal strategy returns are calculated as the returns of a strategy that weights stocks proportionally to the negative of market-adjusted returns on days week-1, with weights scaled to add up to $1 short and $1 long. Returns are hedged against conditional market factor exposure.

A.11 Details of the variables

A.11.1 Firm Level Variables

Table A.4. Firm Level Variables

No.	Acronym	Firm Characteristic	Data Source	Frequency
1	absacc	Absolute accruals	Compustat	Annual
2	acc	Working capital accruals	Compustat	Annual
3	age	Stock Age (days)	Compustat	Daily
4	agr	Asset growth	Compustat	Annual
5	agrq	Asset growth quarterly	Compustat	Quarterly
6	at	Assets	Compustat	Annual
7	atq	Assets quarterly	Compustat	Quarterly
8	baspread	Bid-ask spread	CRSP	Daily
9	beta	Beta	CRSP	Daily
10	betasq	Beta squared	CRSP	Daily
11	bm	Book-to-Market	Compustat + CRSP	Daily*
12	bm_ia	Industry adjusted Book-to-Market	Compustat + CRSP	Daily*
13	capx	CAPEX	Compustat	Annual
14	capx_at	CAPEX to assets	Compustat	Annual
15	cash	Cash holdings	Compustat	Annual
16	cashq	Cash holdings quarterly	Compustat	Quarterly
17	cashdebt	Cash flow to debt	Compustat	Annual
18	cashpr	Cash productivity	Compustat + CRSP	Daily
19	ceq	Common/Ordinary Equity	Compustat	Annual
20	cf	Cash flow	Compustat	Annual
21	cfp	Cash flow to price ratio	Compustat + CRSP	Daily
22	cfp_ia	Industry adjusted cash flow to price ratio	Compustat + CRSP	Daily**

Table A.4. Firm Level Variables (continued.)

No.	Acronym	Firm Characteristic	Data Source	Frequency
23	chato	Change in asset-Turnover	Compustat	Annual
24	chatoia	Industry adjusted change in asset-turnover	Compustat	Daily**
25	chcsho	Change in shares outstanding	Compustat	Annual
26	che	Cash	Compustat	Annual
27	chempia	Industry adjusted change in employees	Compustat	Daily**
28	cheq	Cash	Compustat	Quarterly
29	chinv	Change in inventory	Compustat	Annual
30	chpm	Change in Profit Margin	Compustat	Annual
31	chpmia	Industry Adjusted change in Profit Margin	Compustat	Daily**
32	chtx	Change in tax expense	Compustat	Annual
33	chtxq	Change in tax expense	Compustat	Quarterly
34	cinvest	Corporate investment	Compustat	Quarterly
35	convind	Convertible debt indicator	Compustat	Annual
36	currat	Current Ratio	Compustat	Annual
37	depr	Depreciation / PP&E	Compustat	Annual
38	divi	Dividend Initiation	Compustat	Annual
39	divo	Dividend Omission	Compustat	Annual
40	dltt	Long term debt	Compustat	Annual
41	dlttq	Long term debt	Compustat	Quarterly
42	dolvol	Dollar Trading Volume	CRSP	Daily
43	dv	Dividend paying	Compustat	Annual
44	dvt	Dividend	Compustat	Annual
45	dy	Dividend to Price	Compustat + CRSP	Daily
46	egr	Growth in Common Shareholder Equity	Compustat	Annual
47	ep	Earnings to price	Compustat + CRSP	Daily
48	gma	Gross profitability	Compustat	Annual
49	gmaq	Gross profitability	Compustat	Quarterly

Table A.4. Firm Level Variables (continued.)

No.	Acronym	Firm Characteristic	Data Source	Frequency
50	grCAPX	Growth in CAPEX	Compustat	Annual
51	grltnoa	Growth long term net operating assets	Compustat	Annual
52	grltnoaq	Growth long term net operating assets	Compustat	Quarterly
53	herf	Industry sales concentration	Compustat	Annual
54	hire	Employee growth rate	Compustat	Annual
55	ib	Income Before Extraordinary Items	Compustat	Quarterly
56	idiovol	Idiosyncratic Returns Volatility	CRSP	Daily
57	ill	Illiquidity	CRSP	Daily
58	indmom	Industry Momentum	CRSP	Daily
59	intan	Intangibles	Compustat	Annual
60	intanq	Intangibles	Compustat	Quarterly
61	intan_at	Intangibles to assets	Compustat	Annual
62	invest	Capital expenditures and inventory	Compustat	Annual
63	investq	Capital expenditures and inventory	Compustat	Quarterly
64	lev	Leverage	Compustat + CRSP	Daily*
65	lgr	Growth in long-term debt	Compustat	Annual
66	lt	Liabilities	Compustat	Annual
67	maxret21d	Maximum daily return (last 21 days)	CRSP	Daily
68	meanptg	Mean price target	IBES	Daily
69	meanptg_price	Mean price target to price	IBES + CRSP	Daily
70	medptg	Median price target	IBES	Daily
71	medptg_price	Median price target to price	IBES + CRSP	Daily
72	mktcap	Size	CRSP	Daily
73	mktcap_ia	Industry Adjusted Size	CRSP	Daily
74	mom5d	Momentum (5 days)	CRSP	Daily
75	mom10d	Momentum (10 days)	CRSP	Daily
76	mom30d	Momentum (30 days)	CRSP	Daily

Table A.4. Firm Level Variables (continued.)

No.	Acronym	Firm Characteristic	Data Source	Frequency
77	mom60d	Momentum (60 days)	CRSP	Daily
78	mom90d	Momentum (90 days)	CRSP	Daily
79	niadj	Adjusted Net Income	Compustat	Annual
80	niadj_at	Adjusted Net Income to Assets	Compustat	Annual
81	nincr	Number of earnings Increase	Compustat	Quarterly
82	numdown1m	Number of price targets Down (1 month)	IBES	Daily
83	numup1m	Number of price targets up (1 month)	IBES	Daily
84	numest	Number of price targets estimates	IBES	Daily
85	operprof	Operating Profitability	Compustat	Annual
86	pchcapx	Change in Capital Expenditures	Compustat	Annual
87	pchcapx_ia	Industry Adjusted Change in Capital Expenditures	Compustat	Daily**
88	pchcurrat	% change in current ratio	Compustat	Annual
89	pchdepr	% change in depreciation	Compustat	Annual
90	pchgm_pchsale	% change in gross margin - % change in sales	Compustat	Annual
91	pchquick	% quick ratio	Compustat	Annual
92	pchsale_pchinvt	% change in sales - % change in inventory	Compustat	Annual
93	pchsale_pchrect	% change in sales - % change in A/R	Compustat	Annual
94	pchsale_pchxsga	% change in sales - % change in SG&A	Compustat	Annual
95	pchsaleinv	% change sales-to-inventory	Compustat	Annual
96	pctacc	Percent Accruals	Compustat	Annual
97	pdret	Previous day return	CRSP	Daily
98	pdret_sq	Previous day return squared	CRSP	Daily
99	prc	Price	CRSP	Daily
100	ps	Financial statements score	Compustat	Annual
101	ptghigh	Max price target	IBES	Daily
102	ptghigh_price	Max price target to price	IBES + CRSP	Daily
103	ptglow	Min price target	IBES	Daily

Table A.4. Firm Level Variables (continued.)

No.	Acronym	Firm Characteristic	Data Source	Frequency
104	ptglow_price	Min price target to price	IBES + CRSP	Daily
105	quick	Quick ratio	Compustat	Annual
106	rd	R&D increase	Compustat	Annual
107	rdq	R&D increase	Compustat	Quarterly
108	rd_mve	R&D to market capitalization	Compustat	Daily*
109	rd_sale	R&D to sales	Compustat	Annual
110	rd_saleq	R&D to sales	Compustat	Quarterly
111	realestate	Real estate holdings	Compustat	Annual
112	retvol5d	Returns volatility (5 days)	CRSP	Daily
113	retvol10d	Returns volatility (10 days)	CRSP	Daily
114	retvol21d	Returns volatility (21 days)	CRSP	Daily
115	retvol60d	Returns volatility (60 days)	CRSP	Daily
116	retvol252d	Returns volatility (252 days)	CRSP	Daily
117	roa	Return on asset	Compustat	Annual
118	roaq	Return on asset quarterly	Compustat	Quarterly
119	roavol	Earnings volatility	Compustat	Quarterly
120	roe	Return on equity	Compustat	Annual
121	roeq	Return on equity quarterly	Compustat	Quarterly
122	roic	Return on invested capital	Compustat	Annual
123	sale	Sales/Turnover	Compustat	Annual
124	salecash	Sales to cash	Compustat	Annual
125	saleinv	Sales to inventory	Compustat	Annual
126	salerec	Sales to receivables	Compustat	Annual
127	secured	Secured debt	Compustat	Annual
128	securedind	Secured debt indicator	Compustat	Annual
129	sgr	Sales growth	Compustat	Annual
130	sgrq	Sales growth	Compustat	Quarterly

Table A.4. Firm Level Variables (continued.)

No.	Acronym	Firm Characteristic	Data Source	Frequency
131	shrout	Shares Outstanding	CRSP	Daily
132	sin	Sin Stocks	Compustat	Annual
133	skew5d	Skewness (5 days)	CRSP	Daily
134	skew10d	Skewness (10 days)	CRSP	Daily
135	skew21d	Skewness (21 days)	CRSP	Daily
136	skew60d	Skewness (60 days)	CRSP	Daily
137	skew252d	Skewness (252 days)	CRSP	Daily
138	sp	Sales to price	Compustat + CRSP	Daily
139	std_dolvol	Vollatility of liquidity ($ trading volume)	CRSP	Daily
140	std_turn	Vollatility of liquidity (share turnover)	CRSP	Daily
141	stdacc	Accrual Volatility	Compustat	Quarterly
142	stdcf	Cash flow volatility	Compustat	Quarterly
143	stdevptg	Price targets standard deviation	IBES	Daily
144	stdevptg_price	Coefficient of variation	IBES + CRSP	Daily
145	tang	Debt capacity/firm tangibility	Compustat	Annual
146	tb	Tax income to book income	Compustat	Annual
147	turn	Share turnover	CRSP	Daily
148	week_after_earnings	Week after earning's release	Compustat	Daily
149	week_before_earnings	Week before earning's release	Compustat	Daily
150	xrd	R&D	Compustat	Annual

*: Market Capitalization calculated at the daily level
**: Annual/Quarterly variable adjusted at the daily level
Note: The format of this table is taken from Gu, Kelly, and Xiu (2020), with many characteristics from Green, Hand, and Zhang (2017).

A.11.2 Reddit Variables

Table A.5. Reddit Variables

No.	Characteristic	Definition
1	Account Age	Age of the account as of December 31th 2021
2	Number of Posts	Total number of posts on WSB
3	Number of Comments	Total number of comments on WSB
4	Number of Posts and Comments	Sum of (2) and (3)
5	Number of tickers talked about	Total number of tickers mentioned on WSB
6	Weekend Poster	% of Posts/Comments made on weekends
7	Posts/Comments during trading hours	Number of posts/comments made between 9:30am and 4pm
8	Posts/Comments just before stock market opens	Number of posts/comments made between 9:00am and 9:30am
9	Posts/Comments just after stock market closes	Number of posts/comments made between 4:00pm and 5:00pm
10	Mainly interested in entertainment posts/comments	% posts and comments in entertainment flairs
11	Mainly interested in information posts/comments	% posts and comments in information flairs
12	Mainly interested in news posts/comments	% posts and comments in news flairs
13	Total number of words written	Sum of all words written on WSB
14	Number of spelling errors	-
15	Percentage of spelling errors	-
16	SMOG	SMOG grade of all posts and comments
17	FOG	Gunning Fog index of all posts and comments

Appendix B

A Bad Bunch: Asset Value Under-Reporting in the Mumbai Real Estate Market

B.1 Tables & Figures

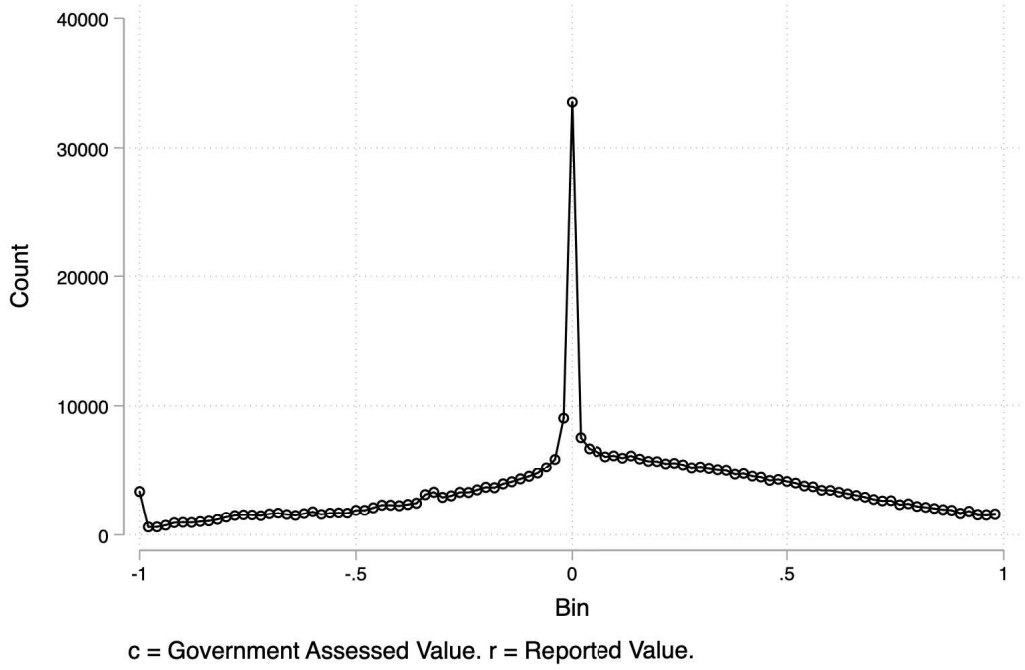
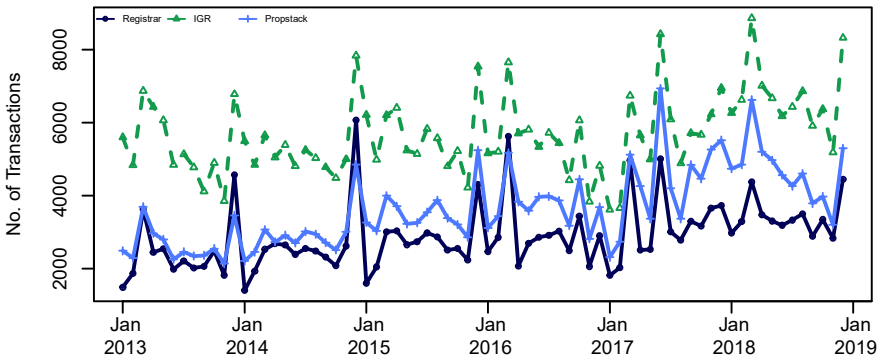
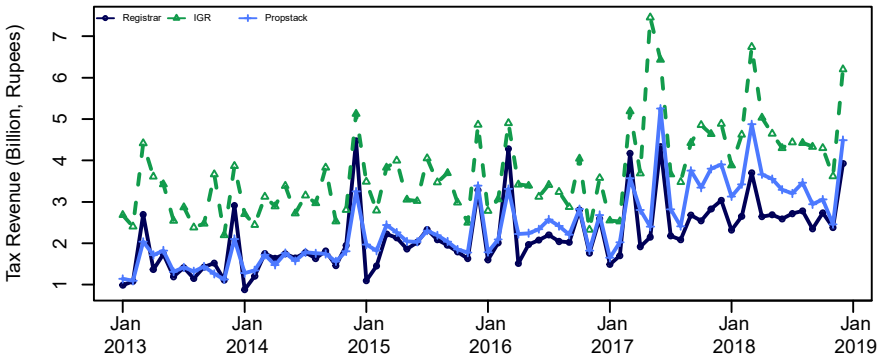


Figure B.1. Bunching Estimates for Sao Paulo, Brazil

This figure replicates Rocha, Scot, and Feinmann (2023) who show similar bunching patterns using the ITBI municipal transaction tax on properties in Sao Paulo, Brazil. The transaction tax rate is 3%, and it is charged on the higher of the buyer's reported value or a government assessed guidance value. The line shows the distribution of reported values across 2% reported value bins, where a reported value bin is measured as a deviation from the government assessed guidance value.



(a) Transactions



(b) Tax Revenue

Figure B.2. Sample Comparison to Aggregate Tax Revenues

This figure plots the monthly time series of the total number of transactions in panel (a) and the total tax revenue from these transactions in panel (b). The blue line with circles plot the numbers obtained from aggregating the extracted Registrar data, the green triangle is the sum reported by the Inspector General of Registrations for a region that Mumbai and Mumbai suburban areas belong to, that is larger than our sample, and the light blue line with "+" plots the aggregated information from Propstack analytics. The overlapping data sample period ends in January 2019, although our full sample is between 2013–2022.

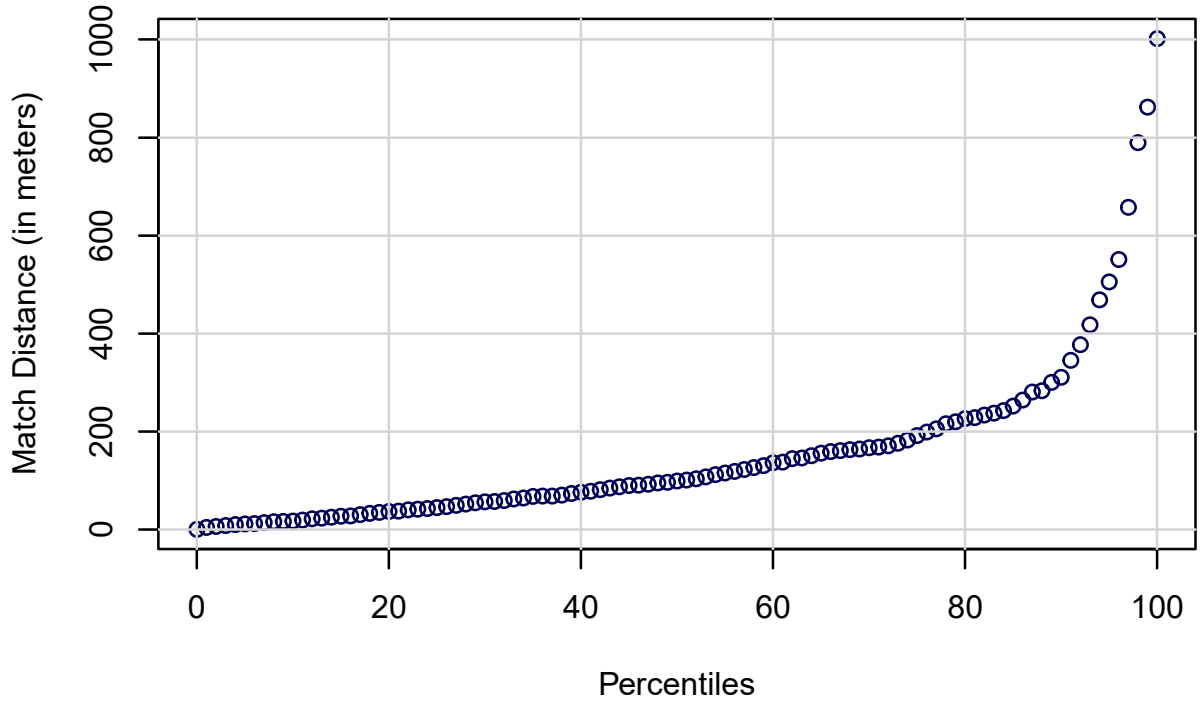


Figure B.3. Propstack-Propequity Match Quality

This figure plots the empirical distribution of the match distance (in meters) between a transaction in our Propstack data (source of administrative data on reported and guidance values) and the match from our Propequity data (source of estimated price data). 80% of the transactions matched to Propequity price data are within 200 meters, and 95% of the transactions are matched to Propequity transactions within 500 meters.

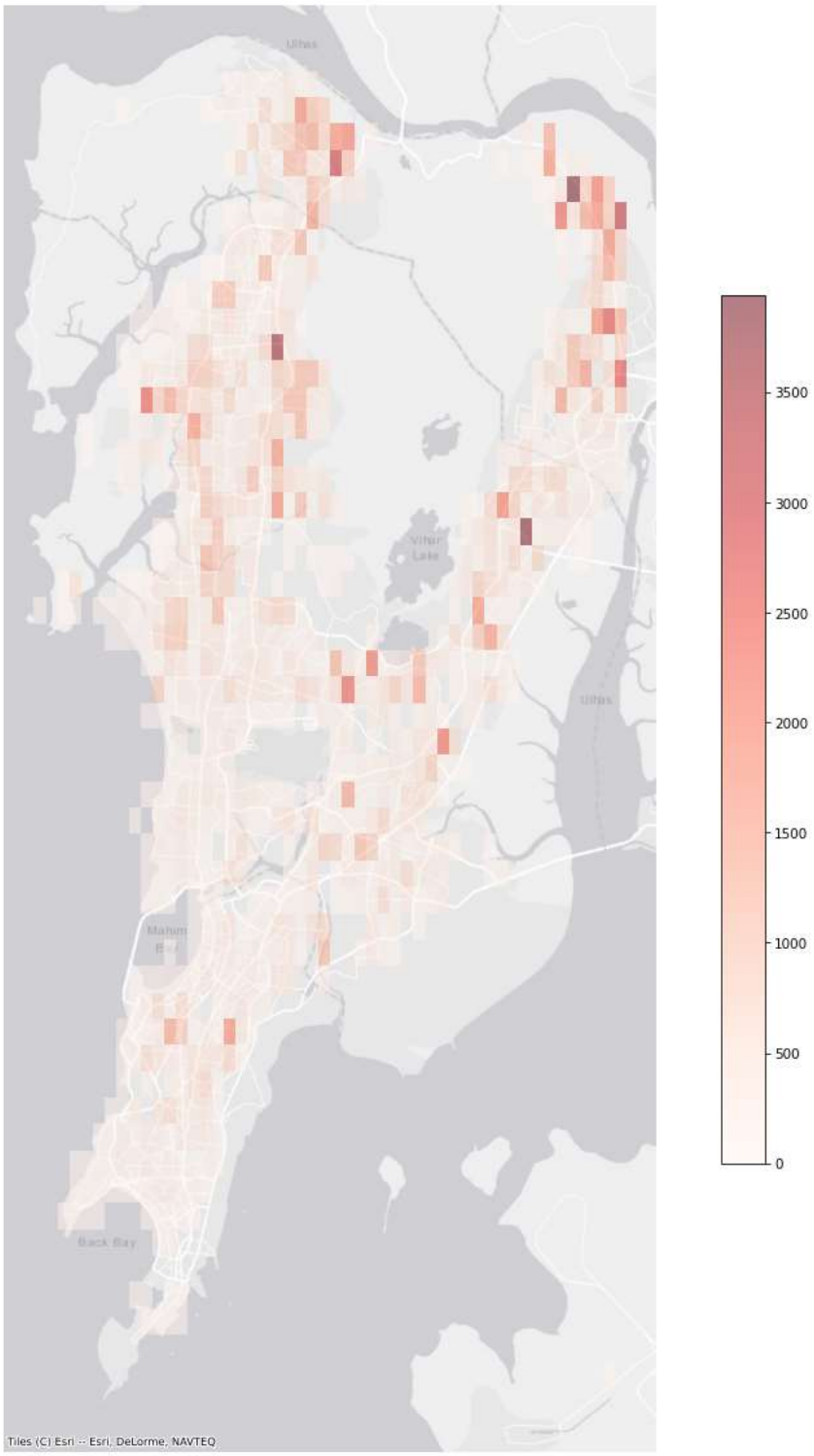


Figure B.4. Heatmap of Transactions in our Final Sample

This heatmap presents the spatial distribution of the final set of transactions in our sample in Mumbai and Mumbai Suburban regions between 2013–2022.

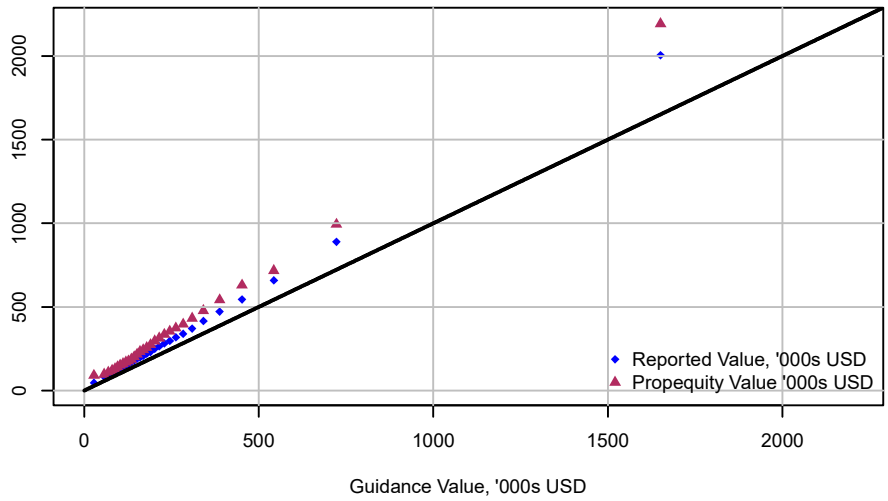


Figure B.5. Correlation between Reported Value, Propequity (Estimated Market) Values, and Guidance Values

This figure presents a binned scatterplot of the average reported values (blue diamonds) and estimated market values from the Propequity data (maroon triangles) within guidance value bins. The black line is the 45-degree line.

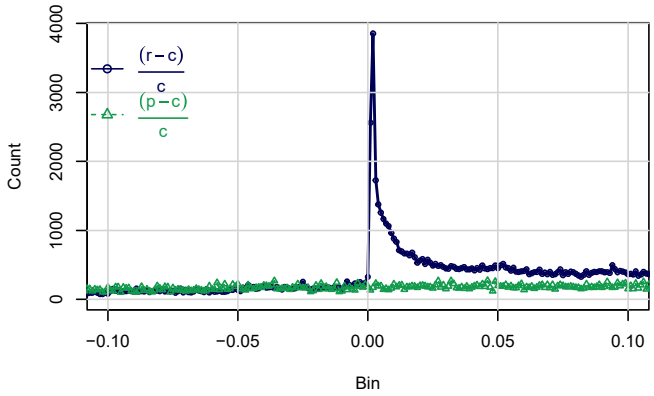
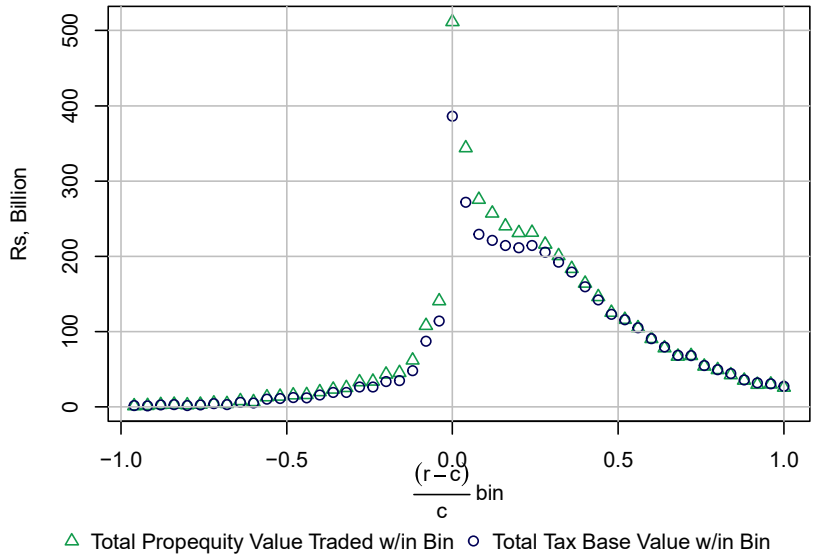
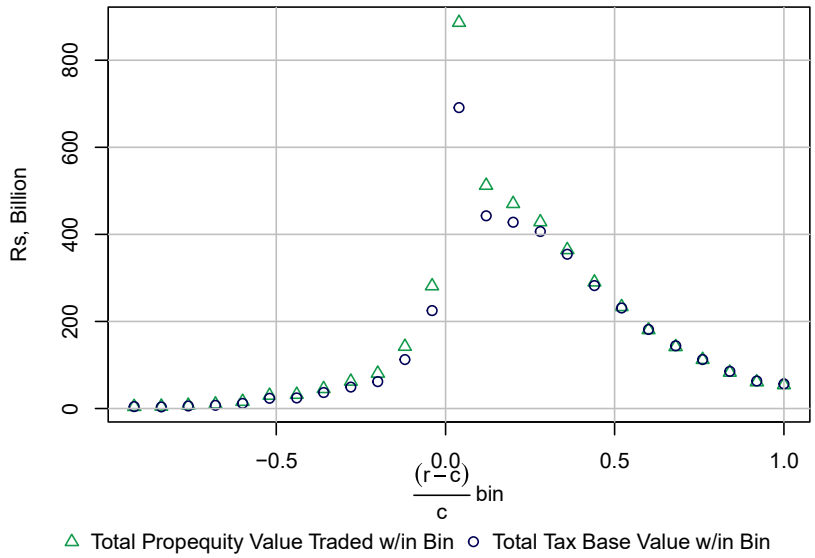


Figure B.6. Bunching of Reported and Propequity Values Around Circle Values in 0.1% Bins

The blue line/circle shows the distribution of reported values across 0.1% reported value bins, where a reported value bin is measured as a deviation from the government assessed value. The green line/triangle shows the distribution of our noisily measured estimate of the market price (the Propequity values) for the same underlying set of transactions reported in the blue line/circle.



(a) Bin width of 4%



(b) Bin width of 8%

Figure B.7. Robustness: Aggregate Tax Base and Propequity Values by Reporting Behavior Bins

Panel A reports aggregated values with a bin width of 4% and Panel B with a bin width of 8%. See Figure 2 for details.

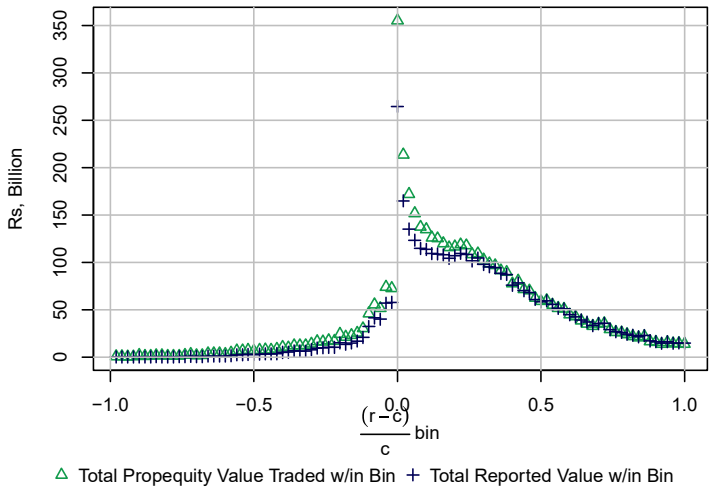


Figure B.8. Aggregate Reported and Propequity Values by Reporting Behavior Bins

This figure reports the aggregated reported and Propequity (estimated market) values by reporting behavior bins. It differs from Figure 2.2 in that this figure aggregates reported values even if they are lower than a transaction's guidance value.

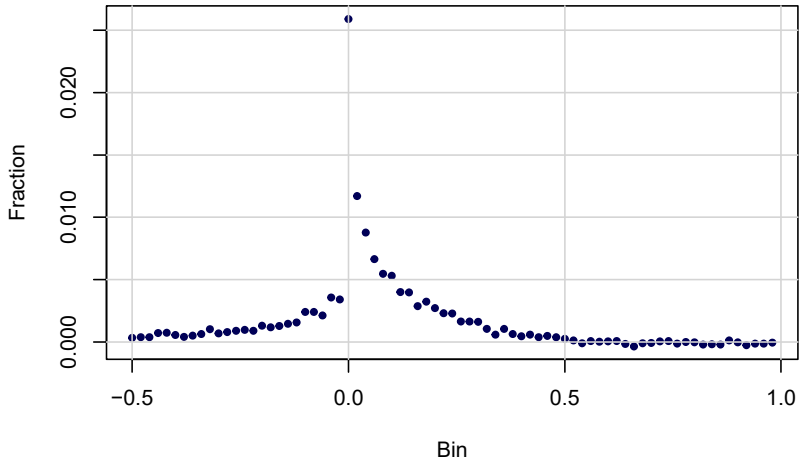


Figure B.9. Truthful Reporting Hypothesis:
Measurement Error Contribution by Reporting Behavior Bin

Each point is equal to aggregate difference between our estimated market values and the tax-base value within a $\frac{r-c}{c}$ bin divided by the total aggregated tax-base value across the whole sample. Assuming that the tax-base values are the true market values, the figure then shows the pattern of measurement error required to rationalize our data under the truthful reporting hypothesis. The points sum to the aggregate measurement error amount as discussed in the decomposition in the text.

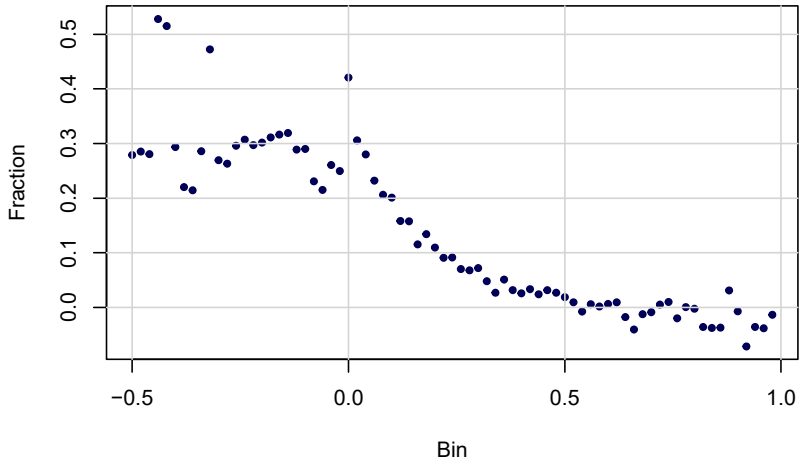


Figure B.10. Estimated Measurement Error Assuming Truthful Reporting: Full Sample

The blue circles show our estimate of measurement error in our estimated market price data under the assumption of completely truthful reporting (i.e. that the tax base value for each transaction is actually the true market value). The estimated measurement error is the sum of the differences between our estimated market values and the tax base value within each bin, divided by the total tax base value in each bin. This version of the figure does not trim large estimated market values.

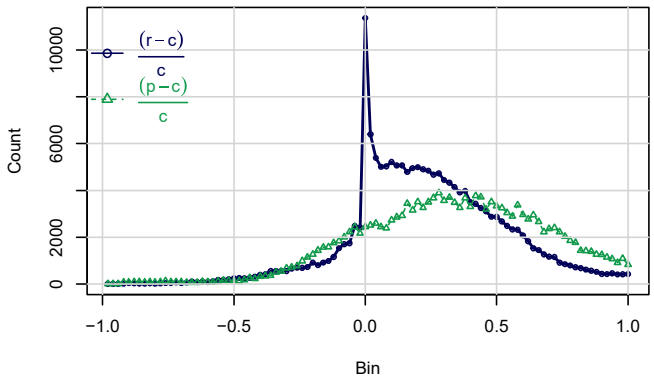


Figure B.11. Exact Match: Bunching of Reported and Propequity Values Around Circle Values

This figure replicates our main findings using 60% of the all transactions in our sample where we have an exact project match for our estimate of the market price (the Propequity values). The blue line shows the distribution of reported values across 2% reported value bins, where a reported value bin is measured as a deviation from the government assessed value. The green line shows the distribution our noisily measured estimate of the market price (the Propequity values) for the same underlying set of transactions reported in the blue line.

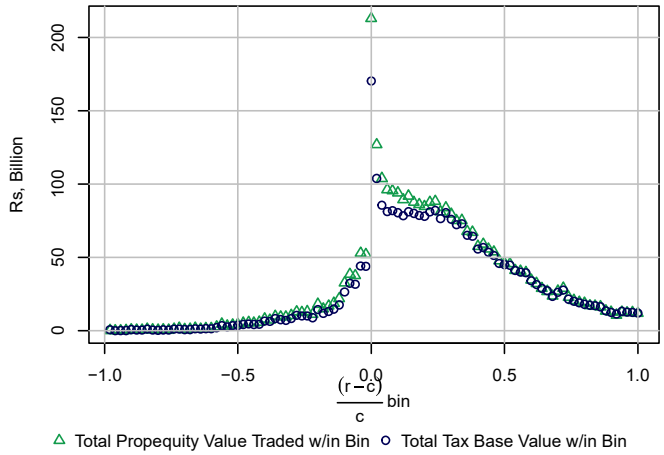


Figure B.12. Exact Match: Aggregate Reported and Propequity Values by Reporting Behavior Bins

This figure replicates our main findings using 60% of the all transactions in our sample where we have an exact project match for our estimate of the market price (the Propequity values). The green triangles show the aggregated reported value within 2% reported value bins, where a reported value bin is measured as a deviation from the government assessed value. The blue circles show the aggregate noisily measured estimate of the market value (the Propequity values) for the same underlying set of transactions reported in the green triangles.

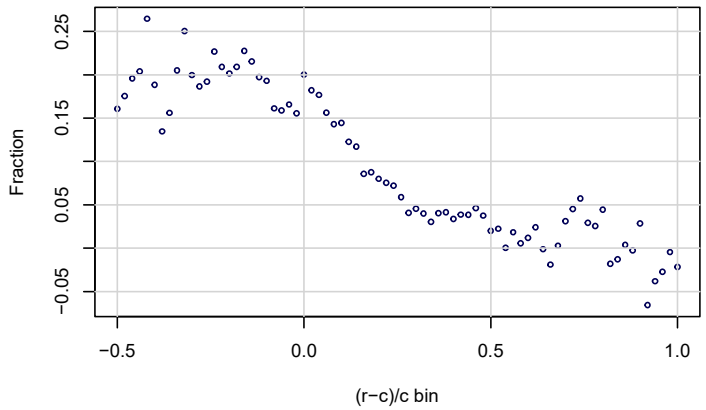


Figure B.13. Exact Match: Under-Reporting Rate by Reporting Behavior Bins

This figure replicates our main findings using 60% of the all transactions in our sample where we have an exact project match for our estimate of the market price (the Propequity values). The blue circles show the estimated under-reporting rate within 2% reported value bins, where a reported value bin is measured as a deviation from the government assessed value.

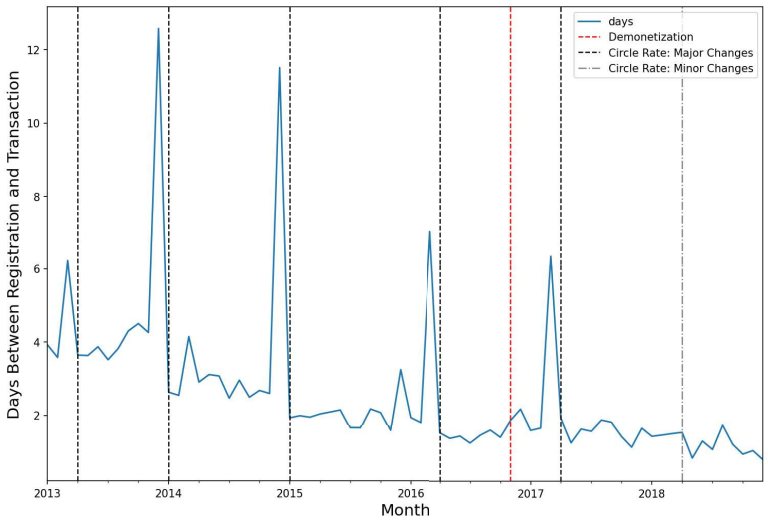


Figure B.14. Agreement Date Backdating

This figure plots the agreement month on the x-axis and the average number of days between the registration and transaction dates on the y-axis.

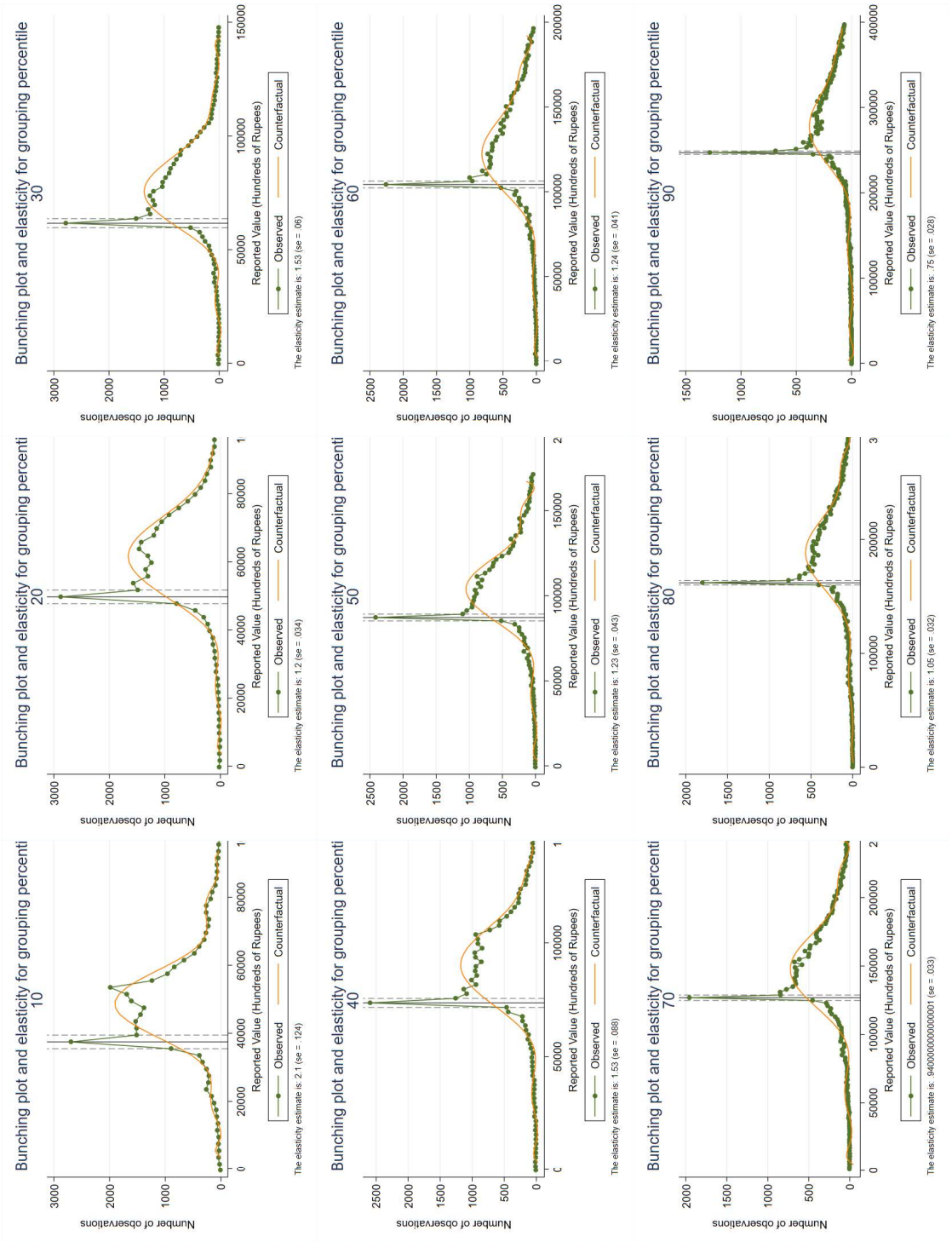


Figure B.15. Bunching Estimates for Guidance Value Bins

This figure presents the bunching elasticity estimates for various Guidance Value bins, which are reported in 2.2.

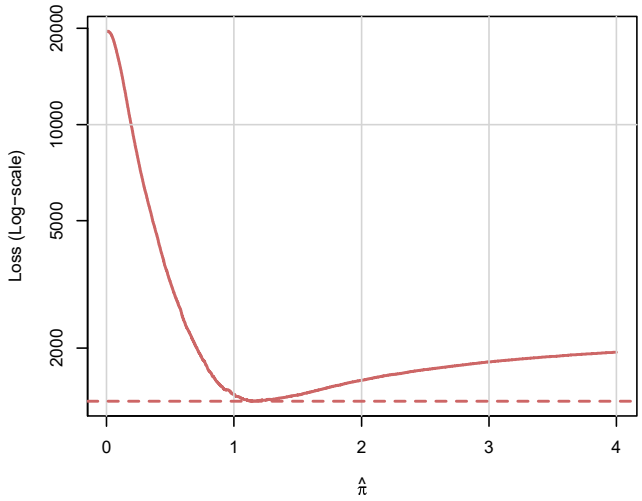


Figure B.16. Loss Function: $\hat{\pi}$

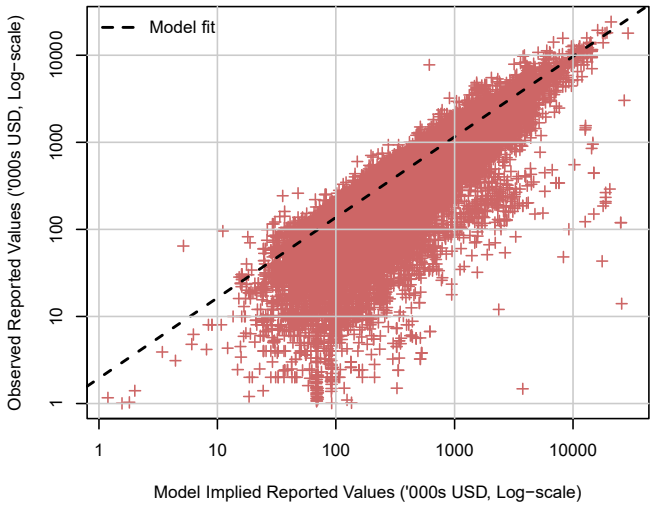
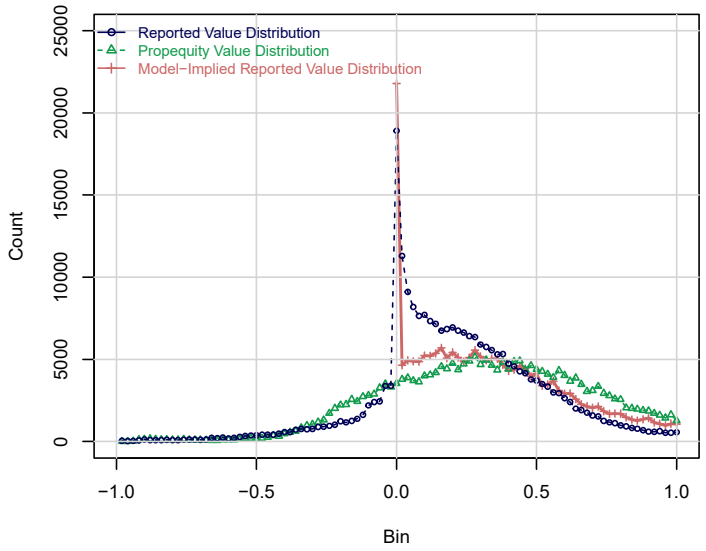
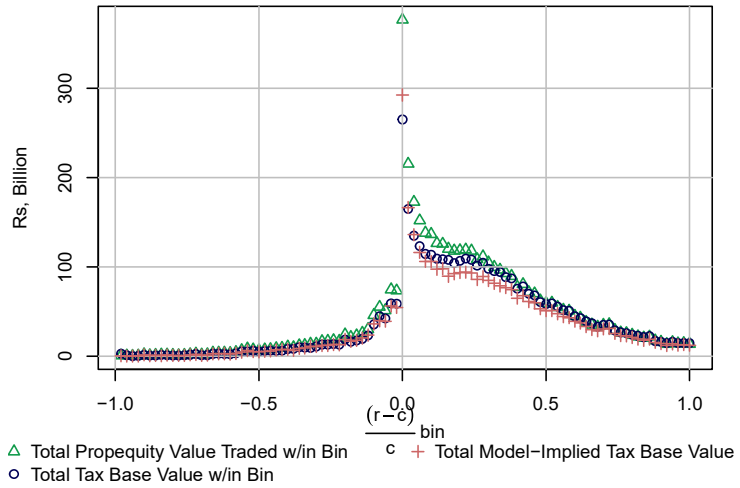


Figure B.17. Model fit



(a) Count Distribution



(b) Value Distribution

Figure B.18. Model-Implied Reported Value Distribution

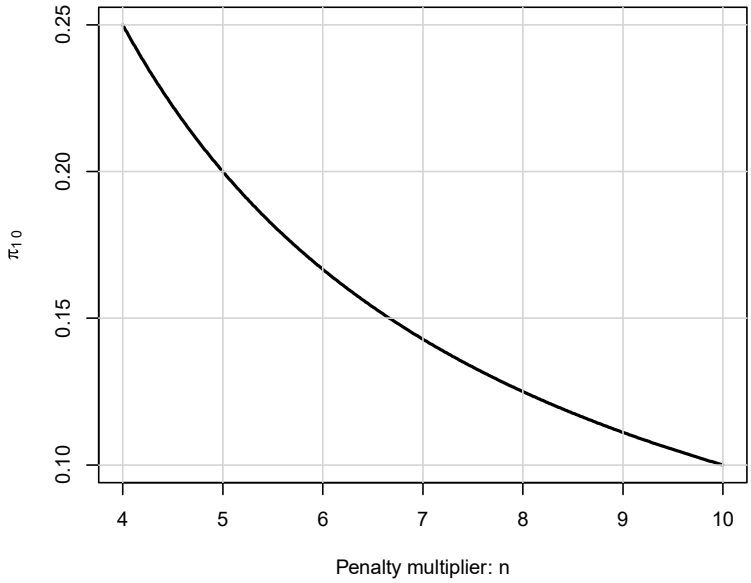


Figure B.19. Efficiency Frontier

The figure plots the combinations of $\pi_{1,0}$ and n where there is no under-reporting and maximum tax revenue generated by the authorities.

				2BHK COST SHEET-
Saleable Area		1237	1242	12
Basic Rate		6099	6099	
Basic Cost		7544463	7574958	
Car Parking Charges (1 car park slot)		400000	400000	
Preferential Location Charges (Rate per sqft)		50	50	
Preferential Location Charges Cost		61850	62100	
Basic Unit Value		**8006313**	**8037058**	**8061**
				Additi
Corpus Fund Rs 40/- per sq.ft		49480	49680	
Infrastructure charges Rs 75/- per sq.ft		92775	93150	
Maintenance Charges for 12 months @ Rs. 3.50/sq.ft		51954	52164	
CMWSSB & TNEB Charges		100000	100000	
Documentation Charges		25000	25000	
Club Amenities charges		50000	50000	
GST@18%		57551	57657	
Total Additional Charges Including GST@18%		**426760**	**427651**	**428**
				PAYME
Application Amount		300000	300000	
Allotment stage Payment (To be paid on or before15th day from Booking Date)	**10%**	500631	503706	
1st Quarterly Instalment - 11th Quarterly Instalment (To be paid on or before 45th day from Booking Date)	**85%**	6805366	6831499	
12 Quarterly Installment upon Intimation on Possession (Inclusive of Additional charges)	**5%**	827076	829503	
Total Payable		**8,433,073**	**8,464,709**	**8,**

Figure B.20. Example Price Sheet from Propequity

This extract presents the detailed breakdown of the costs covered by our data provider. Our estimate of the market value includes both the property purchase cost and other ancillary services.

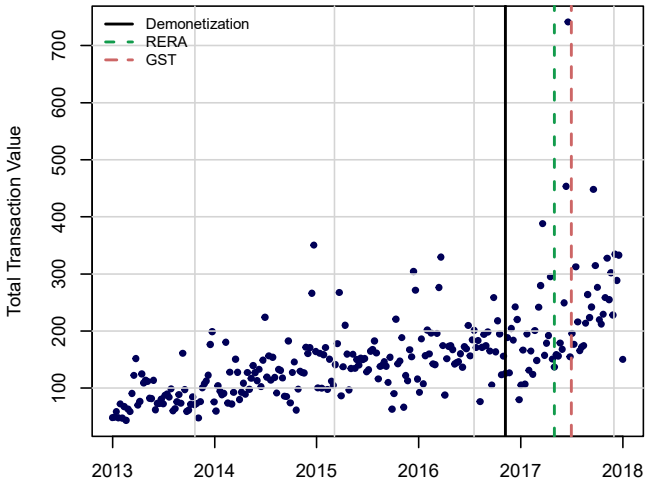


Figure B.21. Weekly Total Estimated Transaction Value

See text for description of demonetization, and Appendix B.5 for description of RERA and GST policies.

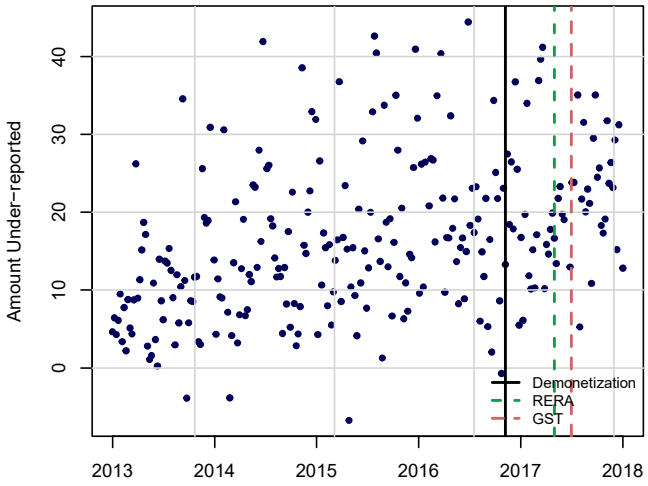


Figure B.22. Weekly Estimated Amount Under-Reported (US$ Millions)

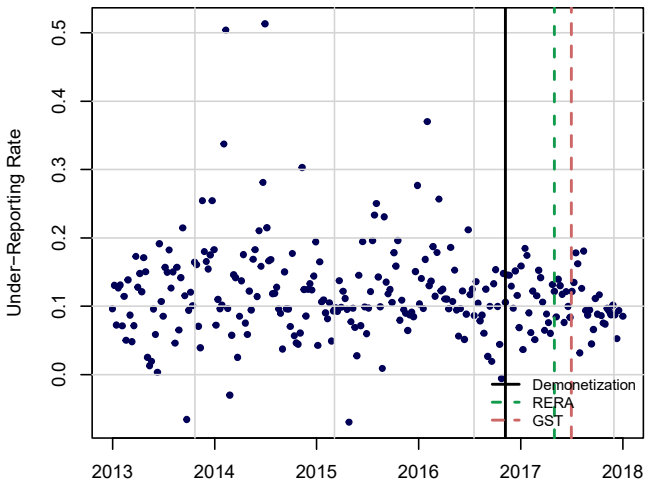


Figure B.23. Weekly Estimated Under-Reporting Rate

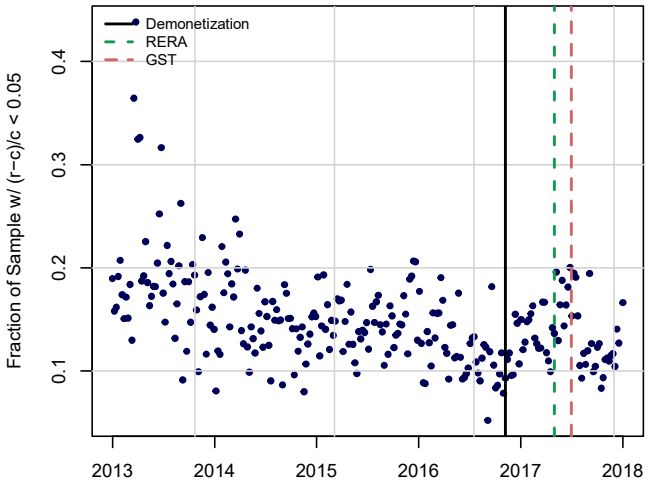


Figure B.24. Fraction of Sample with $(r-c)/c < -.05$

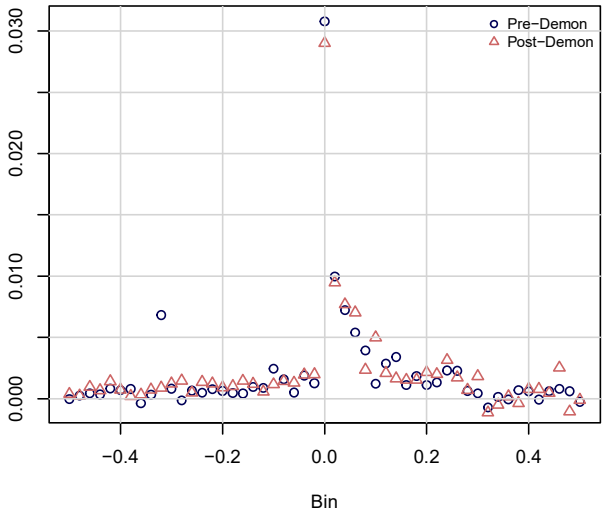
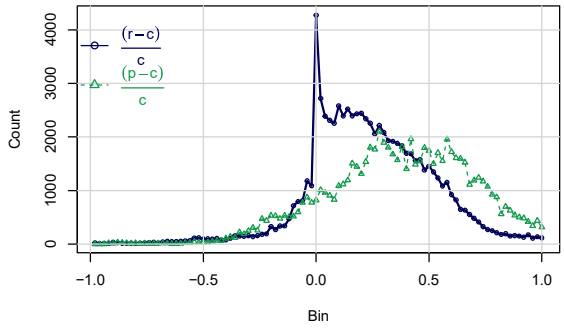
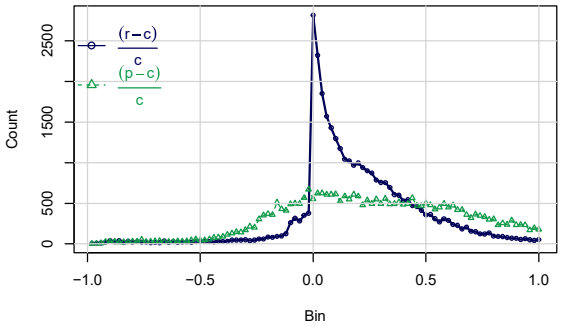


Figure B.25. Contribution of Bin to under-reporting in (r-c)/c Bins 180 Pre- and Post- Demonetization

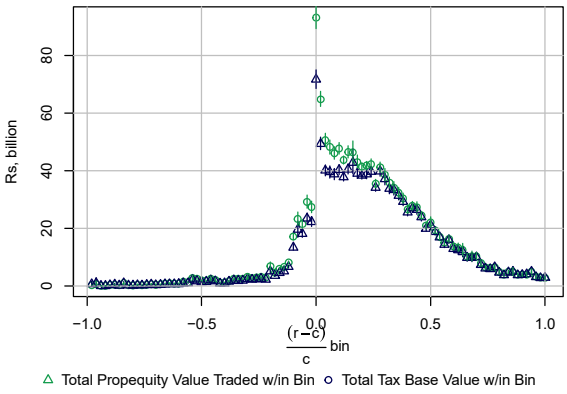
Each point is equal to the total amount of under-reporting within a $\frac{r-c}{c}$ bin divided by the aggregated estimated amount of transactions in the given time-period (pre-demon or post-demon). The sum of all points in the Pre-Demon (Post-Demon) series adds up to the aggregate under-reporting rate in the Pre-Demon (Post-Demon) Period.



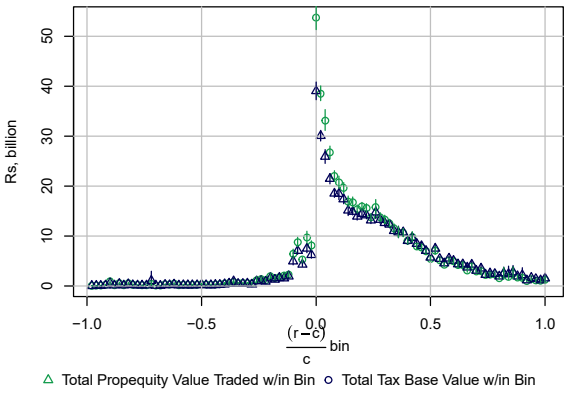
(a) Developer Sales



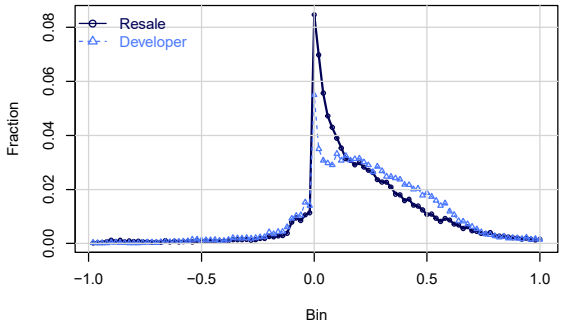
(b) Resale



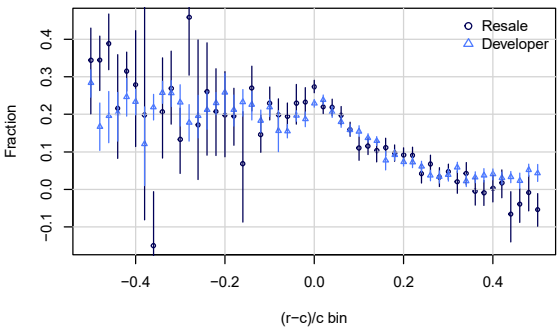
(c) Developer Sales



(d) Resale



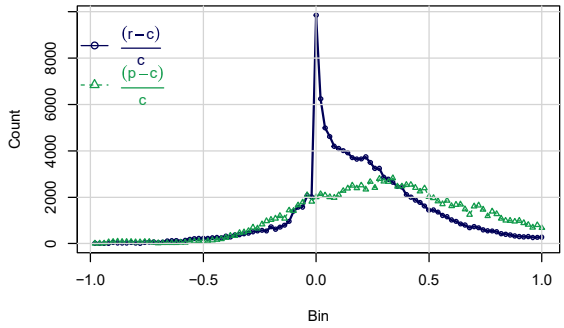
(e) Reported Counts (Density) Comparison



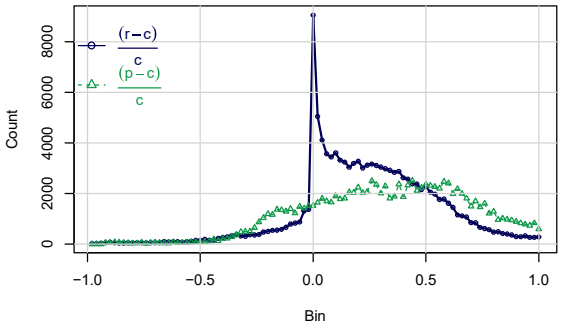
(f) Under-Reporting Rates

Figure B.26. Developer vs. Resale Heterogeneity (Since Jan 2019)

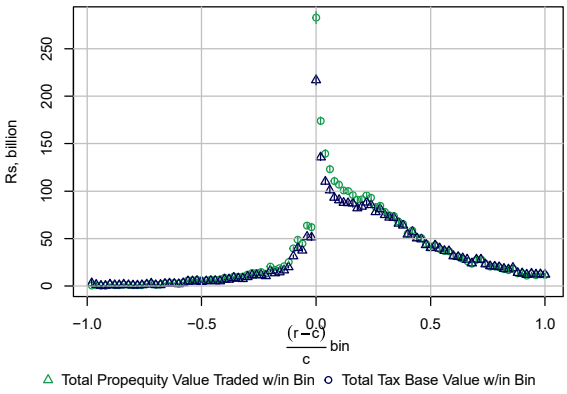
See Figures 2.2a, 2.2b and 2.2c for detailed descriptions. This restricts the sample to begin from January 2019 when all transactions are flagged as developer or resale transactions.



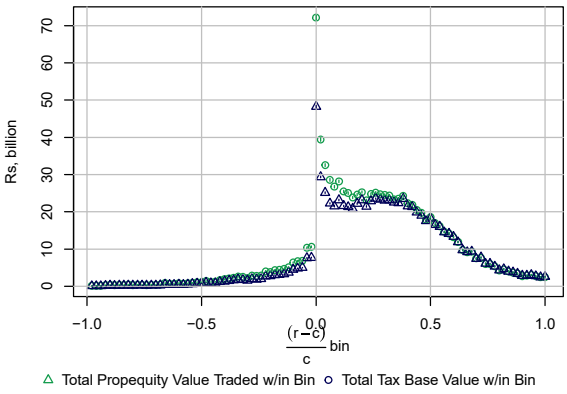
(a) Above Median Circle Value



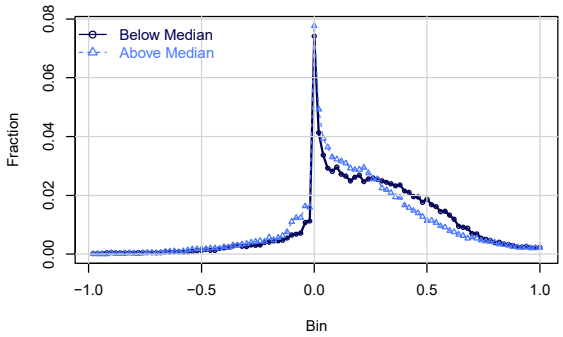
(b) Below Median Circle Value



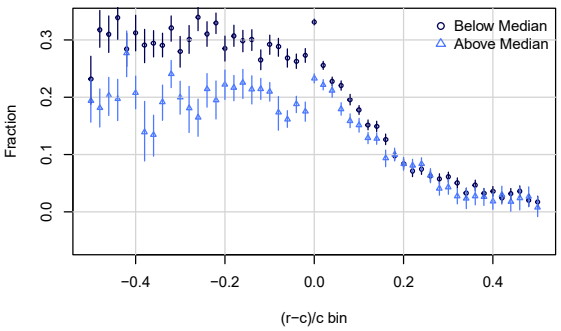
(c) Above Median Circle Value



(d) Below Median Circle Value



(e) Reported Counts (Density) Comparison



(f) Under-Reporting Rates

Figure B.27. Above vs. Below Median Circle Value Heterogeneity

See Figures 2.2a, 2.2b and 2.2c for detailed descriptions.

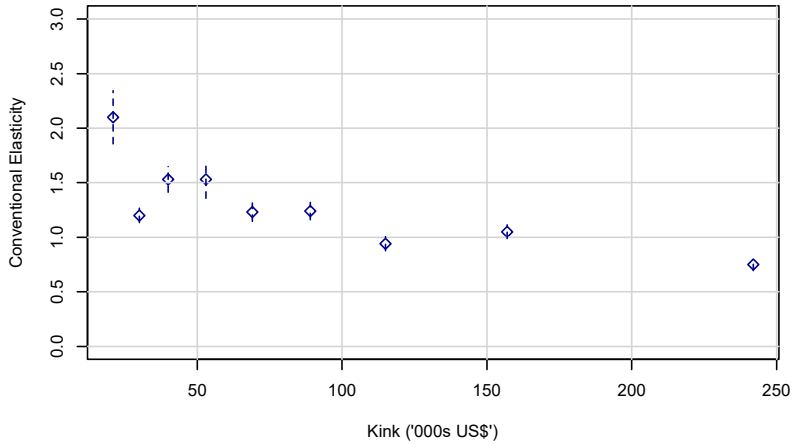


Figure B.28. Reported Value Elasticity to Transaction Tax Rate

This figure plots the reported value elasticity to transaction tax rate by deciles of the guidance value distribution. These estimates are presented in Table 2.2 of the paper.

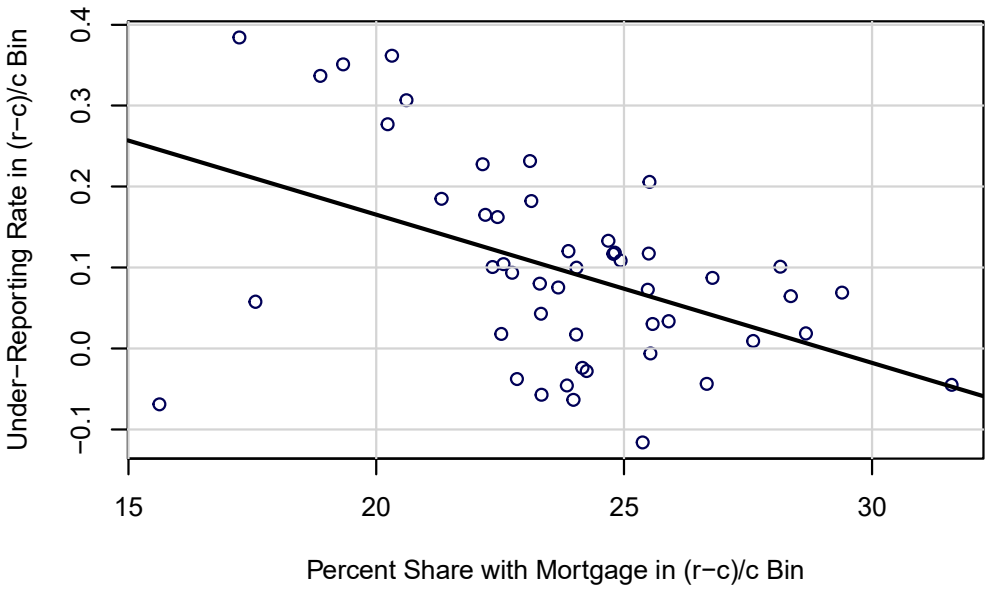


Figure B.29. Incidence of Mortgages and Under-reporting

This figure plots the percentage of mortgage-based transactions in each $(r-c)/c$ bin (x-axis) and the under-reporting rate for each $(r-c)/c$ bin (y-axis). The black line presents the fitted line from a linear regression.

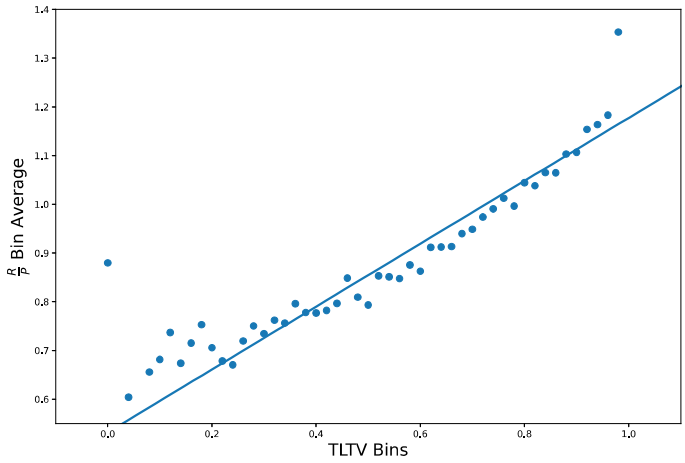


Figure B.30. Equilbrium Relationship between $\frac{R}{P}$ and $TLTV$

This figure plots the average fraction of reported value to the propequity value $\frac{R}{P}$ (y-axis) for bins of the true LTV ($TLTV$) as implied by the propequity value p.

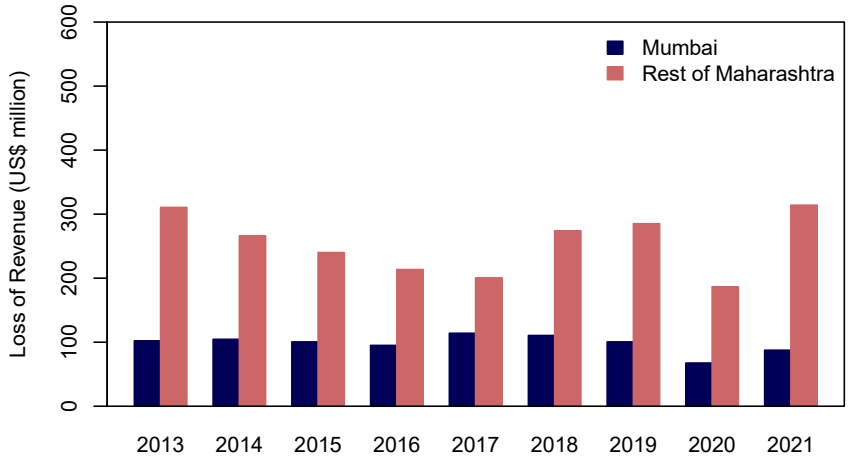


Figure B.31. Estimating Total Revenue Loss for the Government

This figure presents the annual total revenue loss for the government as a result of under-reporting, from the Mumbai Metropolitan Region and the rest of Maharashtra.

Table B.1. Guidance Value Systems in Cities around the World

City	Country	Transaction Tax Rate	Additional tax	Assessment Method	Registration Fee
Delhi	India	6% men, 4% women, 5% joint (stamp)		Centralized	1%
Sao Paulo	Brazil	3%		Centralized	0.75%
Mumbai	India	5% men, 4% women (stamp)		Centralized	1% or ₹30,000, whatever is lower.
Buenos Aires	Argentina	3.6%; if the property is for residential use, valued under ARS 975,000, and the client's first purchase, stamp duty is waived.	10.5% VAT for residential buildings, 21% for other buildings	Centralized	0.20%
Kolkata	India	4%-5% stamp		Centralized	1%
Lagos	Nigeria	2% stamp	8% consent fee; 5% VAT	Decentralized	3%
Rio de Janeiro	Brazil	3%		Centralized	0.75%
Moscow	Russia	0.3% land tax	20% VAT	Centralized	0.1%-1%
Paris	France	For properties more than 5 years old, stamp duty is 5.8%, or 5.09%. For properties less than 5 years old, stamp duty is 0.7%.	For properties more than 5 years old, additional 20% VAT	Decentralized	5.10%
Bogota	Colombia	1%	Transaction tax is called a registration tax (impuesto a registrar)	Centralized	0.50%
Jakarta	Indonesia	5%	1% deed tax	Centralized	0.20%
Chennai	India	7% stamp		Centralized	4%
Lima	Peru	3%		Centralized	0.81%

Table B.1. Guidance Value Systems in Cities around the World (continued.)

City	Country	Transaction Tax Rate	Additional tax	Assessment Method	Registration Fee
Hyderabad	India	4% stamp		Centralized	0.50%
London	United Kingdom	Progresssive stamp duty. 3% higher in each bracket if buyer owns another residence.		Centralized	Progressive fixed fee.
Tehran	Iran	10% for new buildings. 3%-5% otherwise	0.5% stamp	Decentralized	0.10%
Chicago	United States	3.75 dollars per $500		Decentralized	First registration 250-500 dollars for vacant buildings. Semiannual fee of similar ammount afterward
Ho Chi Minh City	Vietnam	No transfer tax	5% VAT	Decentralized	0.5% and VND20,000
Luanda	Angola	2% IPT. Does not apply if transacted property is first property owned by buyer and used for personal and long-term residential purposes	0.3% stamp (on true value)	Decentralized	AOA 105,600 (for certain commercial real estate; unclear for other types of properties)
Ahmedabad	India	4.9% stamp		Centralized	1% for properties exceeding ₹30 Lakh. Otherwise, women pay 0%
Kuala Lumpur	Malaysia	1-4% progressive stamp		Decentralized	MYR100

Table B.1. Guidance Value Systems in Cities around the World (continued.)

City	Country	Transaction Tax Rate	Additional tax	Assessment Method	Registration Fee
Hong Kong	China	15% BSD	5-20% SSD; 1.5-8.5% AVD	Decentralized	230-450 dollars
Riyadh	Saudi Arabia	5%		Decentralized	
Surat	India	4.90%		Centralized	1% (men only)
Madrid	Spain	6.00%	new property: 1.5% stamp duty + 10% VAT if the seller is a company)	Centralized	0.02%-0.175%
Pune	India	5-6% men; 4-5% women	1% metro + 1% Local Body Tax	Centralized	For properties below Rs 30 lakh - 1% of the property value. For properties above Rs 30 lakh - Rs. 30,000
Toronto	Canada	0.5-2.5%	Ontario land transfer tax	Decentralized	
Belo Horizonte	Brazil	3%		Centralized	
Singapore	Singapore	1-3%	5-15% additional buyer stamp duty dependent on citizenship	Decentralized	SGD70 (US$52)
Philadelphia	United States	3.47%	additional 1% tax for commonwealth	Decentralized	$25
Atlanta	United States	$1 per $1,000		Centralized	$100
Barcelona	Spain	7%-11%	10% VAT (new house)	Centralized	400 to 750 EUR per deed; stamp duty (registration) 1.5%

Table B.1. Guidance Value Systems in Cities around the World (continued.)

City	Country	Transaction Tax Rate	Additional tax	Assessment Method	Registration Fee
Saint Petersburg	Russia	0.3% land tax	20% VAT	Centralized	0.1%-1%
Washington Met.Area	United States	1.1 % of consideration or fair market value for residential property transfers less than $400,000 and 1.45% of consideration or fair market value on the entire amount, if transfer is greater than $400,000.		Centralized	

Table B.2. Under-reporting Before Circle Rate Changes

Dep Var: Under-reporting Rate	(1)	(2)	(3)	(4)
Month Before Policy Change	0.061***	0.054***	0.064***	0.068***
	(0.000)	(0.000)	(0.000)	(0.000)
Mean Dep Var.	0.06	0.06	0.06	0.06
Intercept	Yes	Yes	Yes	Yes
Time-trend	No	Yes	Yes	Yes
Month of year FE	No	No	Yes	Yes
Year FE	No	No	No	Yes
No. Obs.	260,614	260,614	260,614	260,614

The table reports the regression results estimating the average under-reporting rate the month before circle rate changes.

B.2 Data

B.2.1 Transactions Data

Our primary dataset on reported values and guidance values comes from Propstack Analytics and is described in the main text. For our analysis of transactions associated to mortgages, however, we required more detailed information on buyer characteristics to match between transactions data and mortgage. We obtained the underlying transaction documents for this sub-sample analysis for the purpose of matching transactions to mortgages.[1]

The main underlying data source on transactions is the publicly available individual property transaction reports released by the Office of the Inspector General of Registration and Controller of Stamps (IGR), Department of Revenue, Government of Mahrashtra, India. This state apparatus plays an important role in collecting state

[1]This section also describes the underlying documents where the Propstack Analytics data is ultimately sourced.

government revenues from across the state using various fiscal instruments available in the state government's toolkit. The state is split into 8 regional divisions and we obtain data for the Mumbai regional division which is comprised of Mumbai City and Mumbai Suburban districts. Our study area currently covers 437 square kilometers out of the 6,640 square kilometers Mumbai Metropolitan Region. We currently focus on this region because we can reliably obtain transaction data that can be mapped to geo-spatial information relevant for our study.

The *eSearch* facility set up by the IGR enables access to transaction-level data for all properties transacted in Greater Mumbai. Every transaction report is in Marathi, the most commonly spoken language in Maharashtra. Figure B.32 presents an example of the original document downloaded from the IGR *eSearch* facility. Figure B.33 presents the transaction report translated into English using Google's translation services. The details available in each transaction report provides a consistent information set for all real-estate transactions for Greater Mumbai. This information set also serves as the basis for the government to make policy decisions on real-estate transaction taxes.

Each transaction report obtained from the *eSearch* facility begins with a document number, and the name of the registrar office (the local IGR office for a region). The more substantive information is in the form of a table starting with the name of the local village where the property is located[2]. The first row of the table in Figure B.33 lists the type of transaction. All real-estate transactions in Maharashtra are classified as "Agreement", "Agreement to Sale", "Sale deed" and "Transfer Deed" types. We filter all downloaded transaction reports to these deed types to form our core data set.

The second row lists the reported price at which the transaction took place. In this

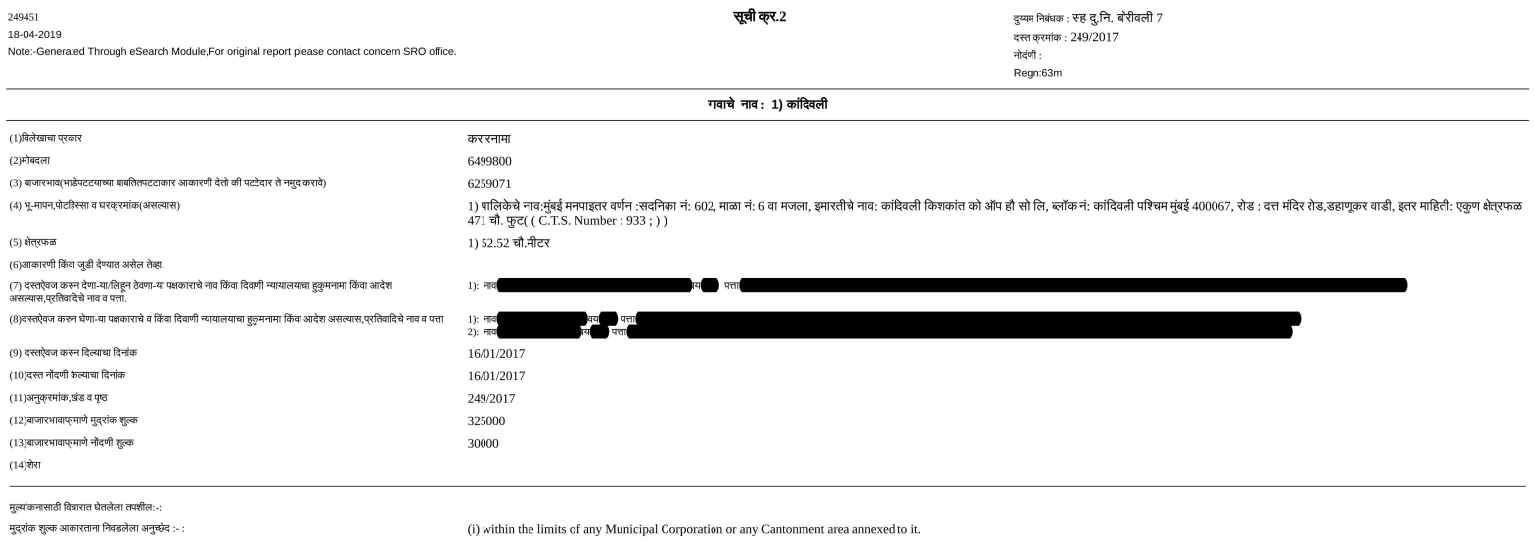
[2]Historically, the Mumbai region was formed of seven islands or fishing villages, which then expanded rapidly over time. The village tag to geographies is more of an artefact of historical documentation than a reference to the economic or social conditions of different regions in Mumbai.

case, the reported transaction price is ₹7,500,000. The third row lists the price as per the government issued guidance value, known as the *policy circle rate* that is determined annually by a legally predetermined process. The policy circle rate determines the floor price at which the government will deem this property to be sold for taxation purposes. The value of this property according to government determined circle rates is ₹4,434,062. Row 12 provides the computed stamp tax paid on this transaction of ₹375,000, determined as the prevailing stamp tax rate, in this case 5%, on the reported transaction value. The circle rate plays an important role in that it sets the lower bound for the stamp tax revenue generated for a property of this type. In the event this property's reported transaction price is below ₹4,434,062, the stamp duty payable will be 5% of this guidance value, after which the law facilitates a process by which the related parties can file for revision. The fourth row of this table provides the property address, and other measurement details in terms of the area of the house, the land registry survey number, and other information relevant for determining the circle rate. The fifth row of this table provides us with the property area, and the next few rows provide details of the two parties to the transaction. Row 9 reports the transaction date for the document, and Row 10 the actual date of formal registration for the sale. These two dates can be different as the law allows for a grace period of 3 months from the actual transaction date during which time they are legally bound to register the sale with the registrar. The last row provides data on the registration fee paid which is capped at ₹30,000 or 1% of the reported value, whichever is lower.

We validate the coverage of our transaction reports data from the IGR *eSearch* facility by matching the total real-estate stamp duty number of documents filed and revenue generated in each year from our transactions data to the official aggregate numbers. Figure B.34 presents this comparison. The top panel of the figure shows the official number of documents filed with the IGR in each month in green against the num-

ber of documents in our transaction data (orange for the registrar data we manually sourced for matching to mortgages and blue for Propstack). The time-series are very highly correlated. The bottom panel of Figure B.34 shows the official aggregate tax revenue collected from stamp duty in each month against the aggregate stamp revenue from our transaction data. The total revenue figures from both sources include stamp revenues and the registration fees for all transactions in a given month. Once again, the time-series are very highly correlated, especially in the second half of our sample period

Although we capture a majority of the transactions in Greater Mumbai, the differences between our aggregate revenue numbers and the official figures arise primarily due to two reasons. The official figures for Mumbai also includes a suburban area of Navi Mumbai, which we do not include in our sample. Moreover, we count revenues in the month the transaction was registered, and this may not necessarily be the same as the official approach, especially for transactions that may be executed at the end of the month, but fulfilled in the early periods of the following month.



249451
18-04-2019
Note:-Generated Through eSearch Module,For original report pease contact concern SRO office.

सूची क्र.2

दुय्यम निबंधक : सह दु.नि. बोरीवली 7
दस्त क्रमांक : 249/2017
नोंदणी :
Regn:63m

गावाचे नाव : 1) कांदिवली

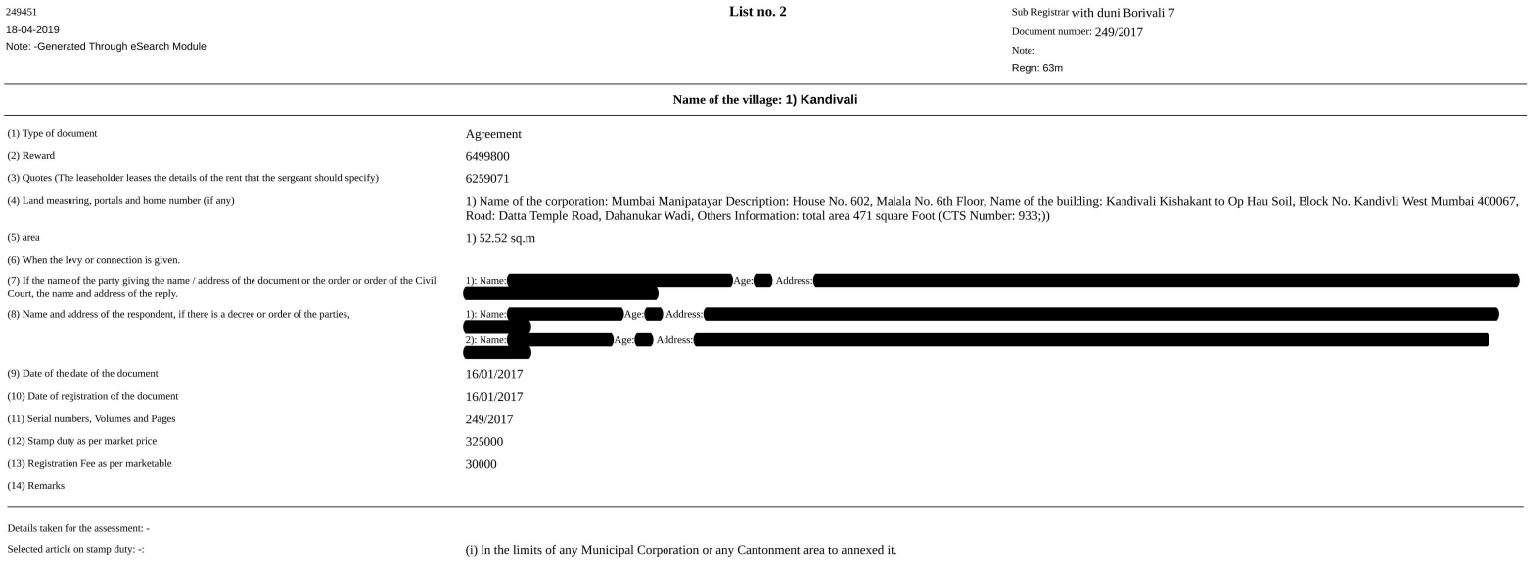
(1)विलेखाचा प्रकार	करारनामा
(2)मोबदला	6499800
(3) बाजारभाव(भाडेपटटयाच्या बाबतितपटटाकार आकारणी देतो की पटटेदार ते नमुद करावे)	6259071
(4) भू-मापन,पोटहिस्सा व घरक्रमांक(असल्यास)	1) पालिकेचे नाव:मुंबई मनपाइतर वर्णन :सदनिका नं: 602, माळा नं: 6 वा मजला, इमारतीचे नाव: कांदिवली किशकांत को ऑप हौ सो लि, ब्लॉक नं: कांदिवली पश्चिम मुंबई 400067, रोड : दत्त मंदिर रोड,डहाणूकर वाडी, इतर माहिती: एकुण क्षेत्रफळ 47ः चौ. फुट((C.T.S. Number : 933 ;))
(5) क्षेत्रफळ	1) 52.52 चौ.मीटर
(6)आकारणी किंवा जुडी देण्यात असेल तेव्हा.	
(7) दस्तऐवज करुन देणा-या/लिहून ठेवणा-या पक्षकाराचे नाव किंवा दिवाणी न्यायालयाचा हुकुमनामा किंवा आदेश असल्यास,प्रतिवादिचे नाव व पत्ता.	1): नाव[redacted] पत्ता[redacted]
(8)दस्तऐवज करुन घेणा-या पक्षकाराचे व किंवा दिवाणी न्यायालयाचा हुकुमनामा किंवा आदेश असल्यास,प्रतिवादिचे नाव व पत्ता	1): नाव[redacted] पत्ता[redacted] 2): नाव[redacted] पत्ता[redacted]
(9) दस्तऐवज करुन दिल्याचा दिनांक	16/01/2017
(10)दस्त नोंदणी केल्याचा दिनांक	16/01/2017
(11)अनुक्रमांक,खंड व पृष्ठ	249/2017
(12)बाजारभावाप्रमाणे मुद्रांक शुल्क	325000
(13)बाजारभावाप्रमाणे नोंदणी शुल्क	30000
(14)शेरा	

मुल्यांकनासाठी विचारात घेतलेला तपशील:-:

मुद्रांक शुल्क आकारताना निवडलेला अनुच्छेद :- : (i) within the limits of any Municipal Corporation or any Cantonment area annexed to it.

Figure B.32. Original Transaction Document: An Example

This figure presents the original transaction document obtained from the IGR *eSearch* facility https://freesearchigrservice.maharashtra.gov.in.



249451
18-04-2019
Note: -Generzted Through eSearch Module

List no. 2

Sub Registrar with duni Borivali 7
Document number: 249/2017
Note:
Regn: 63m

Name of the village: 1) Kandivali

(1) Type of document	Agreement
(2) Reward	6499800
(3) Quotes (The leaseholder leases the details of the rent that the sergeant should specify)	6259071
(4) Land measuring, portals and home number (if any)	1) Name of the corporation: Mumbai Manipatayar Description: House No. 602, Malala No. 6th Floor, Name of the building: Kandivali Kishakant to Op Hau Soil, Block No. Kandivli West Mumbai 400067, Road: Datta Temple Road, Dahanukar Wadi, Others Information: total area 471 square Foot (CTS Number: 933;))
(5) area	1) 52.52 sq.m
(6) When the levy or connection is given.	
(7) If the name of the party giving the name / address of the document or the order or order of the Civil Court, the name and address of the reply.	1): Name: [redacted] Age: [redacted] Address: [redacted]
(8) Name and address of the respondent, if there is a decree or order of the parties,	1): Name: [redacted] Age: [redacted] Address: [redacted] 2): Name: [redacted] Age: [redacted] Address: [redacted]
(9) Date of thedate of the document	16/01/2017
(10) Date of registration of the document	16/01/2017
(11) Serial numbers, Volumes and Pages	249/2017
(12) Stamp duty as per market price	325000
(13) Registration Fee as per marketable	30000
(14) Remarks	

Details taken for the assessment: -

Selected article on stamp duty: -: (i) In the limits of any Municipal Corporation or any Cantonment area to annexed it

Figure B.33. Translated Transaction Document: An Example

This figure presents the translated version of the original document in Figure B.32 using Google Translation services.

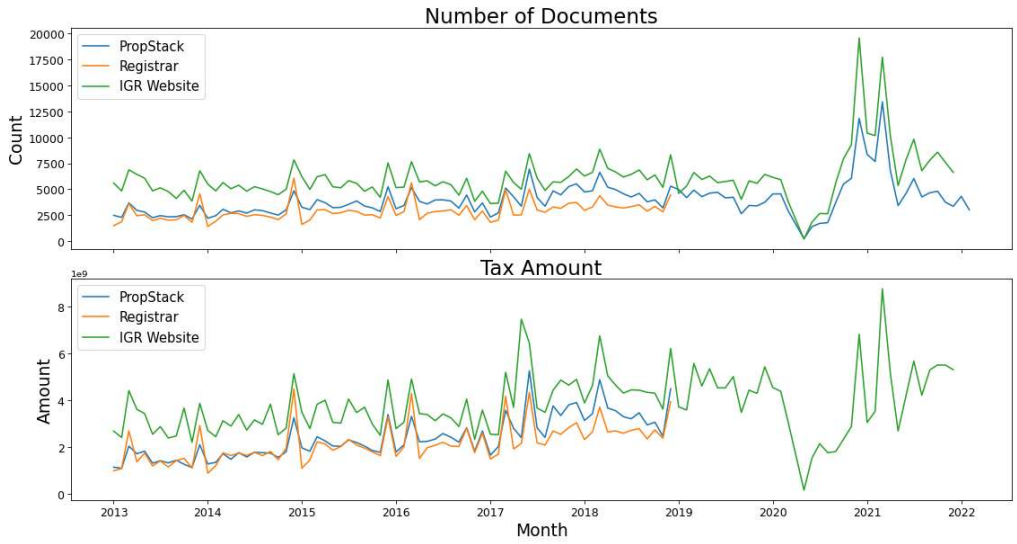


Figure B.34. Coverage Validation

This figure reports the total real-estate stamp duty revenue generated in each year from our transactions data to the official aggregate numbers.

B.2.2 Circle Rate Scheduled Changes

Figure 2.3 shows how reported values and under-reporting behavior evolved over time, particularly in reference to pre-determined dates when circle rates were changed. Circle rates are set at the sub-zone level, a geographic area of approximately .67 square kilometers on average. Table B.3 presents the summary statistics on the circle rate variation in Greater Mumbai for our sample period. At the start of our sample period we have 727 sub-zones, which increase to 747 sub-zones in 2015, and then stabilize at 734 for the remainder of our sample period.[3] In the early years of our sample, nearly all sub-zones underwent changes in circle rates. The average change in each year vary from 0% in 2018 to 14.4% from the previous year in 2015 (Column 3). The cross-

[3]New sub-zones are formed by either dividing existing sub-zones into multiple new ones, or by fusing different parts of multiple sub-zones to form new ones. We keep track of all of the changes in the geo-spatial files, thus identify which regions form to create the new sub-zones, and the old sub-zones they belonged to.

sectional distribution is also large. At the lowest end of the distribution are sub-zones with ₹7330 as the circle rate per square meter of property area (in 2014), to ₹653,240 per square meter of property area in 2018. Figure B.35 presents the geo-spatial variation at the sub-zone level in circle rates at the beginning of our sample (Panel A) and at the end of the IGR sample in 2018 (Panel B). The circle rates have been re-scaled to the mean sub-zone, with the darker red indicating sub-zones with high circle rates and sub-zones in lighter shades of yellow indicating those with the lowest circle rates in Greater Mumbai.

Table B.3. Circle Rate - Summary Statistics

	Sub-zones			Cross-sectional Distribution (×1000₹)					
Year	#	% with Change	% Change	Mean	1%	25%	50%	75%	99%
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
2014	727	-	-	139.59	7.33	81.35	109.00	161.85	580.89
2015	747	100.00	14.44	160.21	14.72	94.60	126.20	189.55	619.25
2016	734	97.49	10.54	172.75	30.04	103.92	134.45	201.05	652.03
2017	734	68.46	6.98	178.40	11.62	109.78	145.15	209.88	653.24
2018	734	0.00	0.00	178.40	11.62	109.78	145.15	209.88	653.24

This table reports the summary statistics on variation in circle rates across sub-zones in Greater Mumbai. Column 1 reports the total number of sub-zones in each year of our sample. Column 2 reports the percent of sub-zones that witnessed a change in circle rates compared to the previous year. Column 3 presents the average percentage change in circle rates relative to the previous year. Columns (4–9) present the cross-sectional distribution of circle rates in 1000s of rupees per square meter of property area.

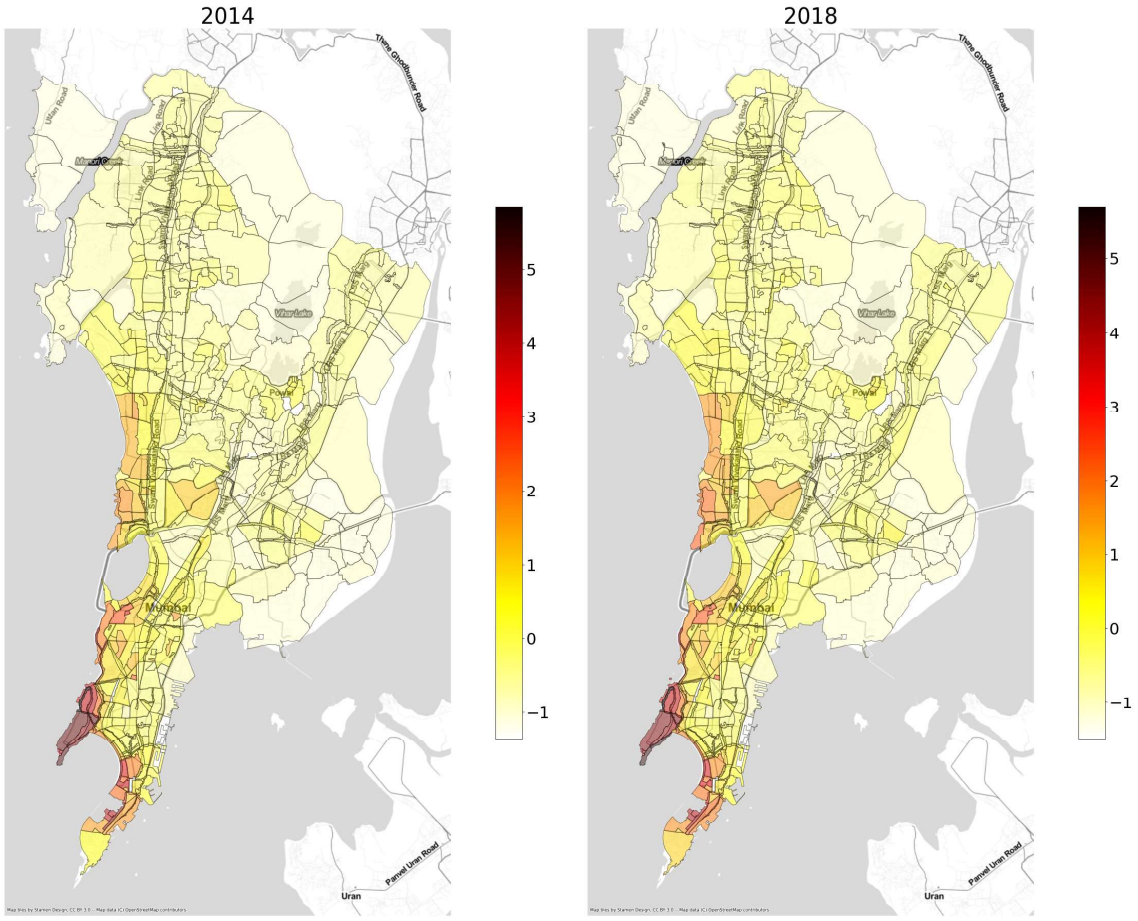


Figure B.35. Sub-Zones

Panel A presents a heatmap of the sub-zones at the start of our data, and Panel B at the end of our sample period. The circle rates are rescaled to the mean sub-zone with darker shades representing the sub-zones with the largest circle rates (in Southern Mumbai) and subzones in white the lowest (northern periphery of Greater Mumbai).

B.3 Simulation to Illustrate Effects of Measurement Error in p

Figure 2.1c shows the distribution of $\frac{r-c}{c}$ and $\frac{m-c}{c}$ for the no under-reporting with measurement error case. The $\frac{r-c}{c}$ blue distribution is exactly the same, by construction,

as that in the high under-reporting with measurement error case. However, the green $\frac{m-c}{c}$ is constructed to reflect bunching in the market value distribution around circle assessed values as opposed to under-reporting behavior. Such bunching could result if the tax authority sets circle assessed values to correctly match market values, or indeed, if buyers and sellers anchor on circle prices when negotiating sale prices. The red curve plots p, which is assumed to be a noisy measure of m.

In the simulation shown in Figure 2.1c we set $p = m + \epsilon$, where ϵ is distributed normally with mean 0 and standard deviation 5. The key insight is that if we do not observe the true distribution of market values $\frac{m-c}{c}$, and cannot correct for measurement error, then it is difficult to distinguish the high under-reporting, no measurement error case from the zero under-reporting, high measurement error case by inspecting bunching behavior alone. The way we have set up the simulation, bunching in reported values is the same magnitude in both cases, and the distribution of our measured market prices p is a smoothed version of the bunching in both m and r around c.

B.4 Bunching Elasticity Estimation

We estimate the elasticity of reporting property values with respect to the transaction tax rate. We employ the conventional method developed by Saez (2010). For our main estimates we assume the tax rate increases from zero below the kink (i.e the guidance value) to five percent above the kink. The formal setup underlying our elasticity estimation is as follows. A household maximizes the following utility function in choose how much to report r for a given house purchase:

$$\max_{r}(R - \tau r) + [r - \frac{r^{(1+\frac{1}{\epsilon})}}{(1+\frac{1}{\epsilon})m^{\frac{1}{\epsilon}}}]$$

where τ is the transaction tax rate (set to .05), m is set to the value the household would report in the absence of any transaction tax, and ϵ is the elasticity of reported value with respect to the transaction tax rate.

The first order condition yields:

$$r = m(1-\tau)^{\epsilon}$$

Substituting $\tau = 0$ we obtain the definition of m as the reported value when there is no transaction tax. This could be truthful reporting, or it could be lower than truthful reporting in the case where there are benefits of under-reporting beyond avoiding the transaction tax. As shown in Saez (2010), differentiating this gives the definition of the ϵ as the percent change in reporting due to a percent change in the tax rate.

Combining this first order condition with equality conditions from the marginal buncher and non-buncher (see Chetty, Friedman, Olsen, and Pistaferri, 2011), we have the following relationship between the underlying elasticity of reporting and the bunching mass B.

$$\hat{e} \approx \frac{\hat{B}}{z^* \cdot h_0(z^*) \cdot \log(\frac{1-t_0}{1-t_1})} \tag{B.1}$$

The estimation procedure involves two steps, first estimating a counterfactual income density based on the income density excluding data points near the kink, and then using the counterfactual density to estimate the excess mass from which the elasticity is recovered.[4] To estimate the counterfactual density, we fit a polynomial of a specified degree to the observed reporting density, excluding the data in a specified window

[4]This description closely follows the implementation of the conventional bunching estimator discussed in Anagol, Lockwood, Davids, and Ramadorai (2022).

around the kink, using the following specification:

$$C_j = \sum_{i=0}^{q} \beta_i^0 \cdot (Z_j)^i + \sum_{i=R_l}^{R_u} \gamma_i^0 \cdot \mathbf{1}[Z_j = i] + \epsilon_j^0. \tag{B.2}$$

Here, q denotes the order of the polynomial, and R_l and R_u denote the lower and upper bounds of "bunching window" near the kink, which is excluded from the polynomial estimation. The convention in Chetty, Friedman, Olsen, and Pistaferri (2011) is to set a symmetric bunching window, such that $R_l = -R_u$. Based on visual inspection of the plots we set the bunching window as one bin to the left of the kink, the kink bin, and one bin to the right of the kink. When estimating the polynomial regression, we follow Chetty, Friedman, Olsen, and Pistaferri (2011) and impose an "integration constraint" such that the total count of observations across the empirical distribution equals the integral of observations under the counterfactual density across the plotted region.[5]

The second step is to compute the excess mass of reported values around the kink relative to this counterfactual density. Using equation (B.2), we compute the counterfactual mass in each bin within the bunching window, $\hat{C}_j^0$. Subtracting this predicted mass from the observed density yields the estimated excess number of individuals who report values near the kink relative to this counterfactual distribution:

$$\hat{B} = \sum_{i=R_l}^{R_u} C_j - \hat{C}_j^0 = \sum_{i=R_l}^{R_u} \hat{\gamma}^0. \tag{B.3}$$

Standard errors for $\hat{e}$ are estimated using a bootstrap procedure. We resample with replacement from the underlying distribution of transactions 10 times, re-estimating

[5] Kleven (2016) notes that imposing an integration constraint may bias the elasticity estimate: "This approach may introduce bias, especially in relatively flat distributions in which interior responses do not affect bin counts (except at the very top of the distribution away from the threshold being analyzed). It would be feasible to implement a conceptually more satisfying approach that does not have this potential bias, but for the reasons stated above, it will matter very little in most applications."

the elasticity each time, and defining the standard error as the standard deviation of the distribution of $\hat{e}$ estimates.

B.5 Developer Regulation and Under-Reporting

There are two other major economic reforms during our sample period that this analysis allows us to investigate. The first is the passage of the Real Estate Regulation Act (RERA), which was a national law that was implemented ("notified") on April 18, 2017 (five months after the demonetization policy was announced). The broad goal of this act was to improve the functioning of the market for newly built apartment homes; the main provisions included requiring developers to set aside money in an escrow account to complete the building of real estate projects 2) requiring any project using 500 meters or more of land space or selling 8 or more units to register and provide updated data on completion times to MahaRERA (the newly established real estate regulator in Maharashtra), 3) procedures and time-lines for developers to respond to customer complaints.[6] Overall, the purpose was to curtail the ability of developers to sell properties before or during construction and then delay/abandon projects without reasonable compensation to buyers.

RERA did not include any specific provisions regarding the reported values of transactions within projects. We argue there are at least two plausible reasons RERA may affect under-reporting. First, if "fly-by-night" developers are also likely to be the ones who engage in collusive deals to under-report property values, it is plausible this regulation would reduce the amount of under-reporting through a change in the composition of developers selling properties. Second, RERA gave buyers stronger recourse

[6]For a full description of RERA provisions see: https://maharerait.mahaonline.gov.in/PDF/FAQMergedPDF.pdf.

in the case where a developer failed to deliver on time; to the extent that the financial compensation a developer pays is linked to the reported value, buyers have a stronger incentive to report the true values to obtain the full protective value of RERA. We expect RERA to primarily affect the market for new home sales; projects completed by May 1, 2017 were not affected by RERA.[7] Nonetheless, in Figure B.23 we do not see any major changes in estimated under-reporting rates in April/May of 2017 (demarcated by the rightmost, purple vertical line) for developer sales over and above the demonetization impacts.

A second potentially relevant policy change was implemented on July 1, 2017, namely, the introduction of the national Goods and Services Tax (GST), which centralized value-added-tax system administered at the central level.[8] It is possible that this policy created "input tax credits" for major construction supplies such as steel and cement relative to the previous fragmented VAT system, which means developers should report the cost of these inputs to the government. This paper trail of input costs may make it more difficult to under-report new sales of real estate. Again, this incentive primarily affects the market for new home sales. We find no major changes in under-reporting rates around the GST reform in Figure B.23.

B.6 Matching Transactions to Mortgages

As noted above, our mortgage analysis required us to match individual transactions data we obtained separately from the Propstack database.[9] For these transactions data that we independently source, we do not have access to cleanly labeled real estate

[7] Completion was determined by whether the project had received an "occupancy certification" from the local housing authority.

[8] See Panigrahi (2021) for a more detailed policy description.

[9] The Propstack database of transactions does not enough identifying information for matching to mortgages.

development names, so we must match them to the estimated market prices using location. Here we describe the matching of our independently sourced transaction documents to Propequity estimated market prices.

Our main approach is to match these independently sourced transaction documents to the "CTS" level. A CTS is the smallest administrative geo-spatial unit in Greater Mumbai for the IGR.[10] This CTS number is important for property registration, mortgages, and determination of the stamp duty to be payable when a property is being bought or sold. We primarily use the CTS of a sales transaction to geo-locate the project. We obtain GIS information for CTS in the Mumbai division from the Urban Development Research Institute (UDRI). We extract the shapefiles for each CTS using ArcGIS, and use these polygons to identify the CTS location for each of our IGR transactions.

Panel A of Appendix Table B.4 presents the details of sample attrition resulting from the process of geo-tagging each transaction report. We begin with the set of transaction reports that contain non-missing information for reported value, guidance values, and the property area. We also exclude reported values less than ₹1000 and area of under 10 square meters. After these initial filters, we have 215,121 transactions in our sample period. 6,864 reports have no property description in the transaction reports, making it impossible to identify its location. Of the remaining 208,257 transactions, we identify the location for these properties using three different approaches. First, if the property description in the transaction report contains the CTS number, we use it to match to the CTS geo-location using our spatial polygons obtained from UDRI. Second, some properties mention several CTS numbers in their property description. This happens when large apartment blocks straddle multiple CTS, and in these instances, we map the property to the first CTS number in the property description.

[10] CTS stands for the Chain and Triangulation Survey Number in the Mumbai suburban district, and the Cadastral Survey Number in Mumbai division. A set of CTS numbers form a sub-zone, and then aggregate upwards to the Mumbai division of the IGR.

Lastly, if there are no CTS numbers available, we use the property description to obtain the latitude-longitude information from Google Maps or Bing, and then match it to the geo-spatial data to identify the CTS number for these properties. Using the three approaches to locate the transactions in Mumbai, we successfully match the data for 187,999 transactions, or 87.3% of the full sample of transaction reports obtained from the IGR website.

To complete the match to Propequity buildings, we next match Propequity projects to the CTS level based on the projects latitude and longitude. Finally, we use the average price of Propequity projects in a transaction's CTS to estimate the transactions "market" price.

Table B.4. Data Validation: GIS Tagging

	Number of Observations
Registrar	215,121
Without Property Description	-6,864
With Property Description	208,257
Does Not Match to a CTS	-20,258
Final Sample	**187,999**
- Perfect CTS Match	48,351
- Match on the First Number of the CTS	86,139
- Google/Bing based Match	53,509

This table reports the sample attrition due to GIS tagging of all transaction data from the IGR, and validating the GIS tag by using circle rate information in the IGR transaction reports.

We now describe how we matched these independently source transactions documents to mortgages. We start with the 215,121 independently sourced transactions (see Table B.4) and 125,195 mortgage transaction documents in the Mumbai central and Mumbai suburban districts.[11] Table B.5 describes our matching procedure. We

[11]Note for this matching exercise we start with the full number of sales transactions we downloaded

match mortgages to transactions using the PAN (tax identification number) of the property buyer (from the transaction document) and the borrower (form the mortgage document), as well as the area (in square meters) of the property. 121,410 of the mortgage documents have a usable PAN number and area information.

Matching directly based on PAN and area we are able to match 42,996 mortgage transactions to a transaction document. Of these 42,996 matching mortgages, 10,385 are duplicate or extraneous documents for the same transaction. For this set of mortgage to transaction matches, we take the chronologically first matching mortgage and drop the others. This leaves us with 35,706 transactions associated with a mortgage (where we only keep the first mortgage). Of these 35,706 transactions associated with a mortgage, 32,611 are transactions that had a property description and a CTS number (i.e. these are transactions that are in our transaction analysis sample of 187,999 transactions). Finally, amongst those 32,611 transactions with mortgages, 31,119 have usable loan information and therefore can be used to calculate a loan to value ratio. To summarize, we currently have 187,999 sales transactions. Of these, 31,119 are matched to a mortgage and are assigned a loan to value ratio based on that mortgage. 1,419 are matched to a mortgage but we do not observe the loan to value ratio. This leaves 187,999 - (31,119 + 1,419) = 155,461 transactions that are currently not matched to a mortgage and therefore assigned a loan to value ratio of zero.

We note that this matching procedure leaves 78,414 mortgages completely unmatched to transactions. In order to asses whether we are systematically missing underlying transactions, we present a comparison of the distribution of successful and unsuccessfully matched mortgages in Figure B.36. We find that the coverage is not disproportionately missing any part of the property size and value distributions, but caveat these results with the fact that there may be measurement error in the loan-to-value

for this region before removing observations without a property description and without a CTS.

ratio of transactions that we were unable to find a mortgage match for.

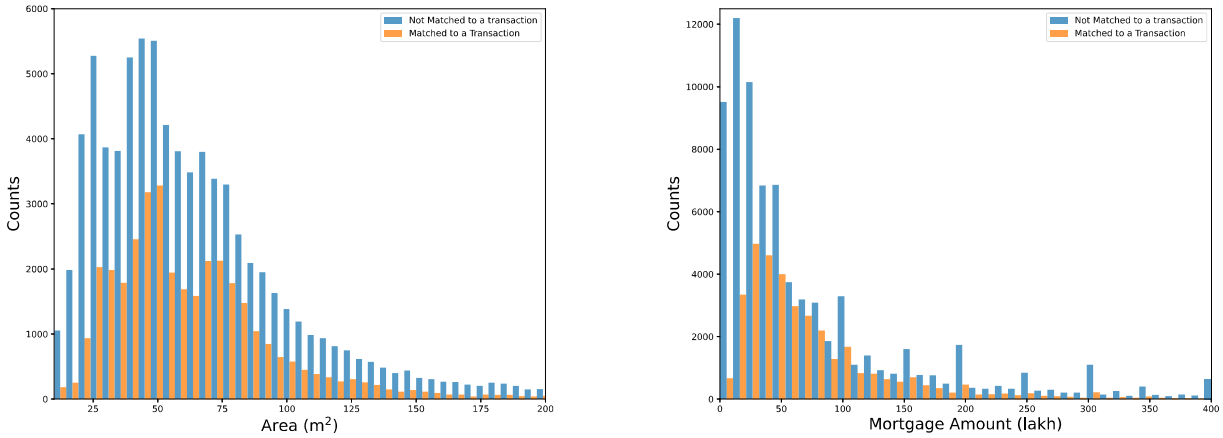


Figure B.36. Mortgage Balance Tests

Panel A (B) represents an histogram of property size (value) for mortgages that either match or do not match to a transaction in the sample.

Table B.5. Mortgage Matching Summary Statistics

1. Number of Mortgages in Mumbai Post 2012	127,195
1.a - With a PAN	124,773
1.b - With a PAN and SQM	121,410
2.a Number of Matching Mortgages	42,996
2.b Number of non Matching Mortgages	78,414
3. Number of Earliest Matching Mortgages	35,706
3.a - With Transaction in our Sample	32,611
3.b - With Transaction in our Sample and LTV ≤ 1	31,119

Appendix C

Privacy Policies and Consumer Data Extraction: Evidence From U.S. Firms

C.1 Proof of Proposition 1

Proof. Let $v^\star$ be the smallest value of visibility v_1 that deters type $\theta = 0$ from deviating. From (3.3), we have

$$(1-\alpha)\left(\sigma\left(0\right)-\sigma\left(1\right)\right)p = \alpha\left(\mu\left(v^\star\right)-\mu\left(0\right)\right)p, \tag{C.1}$$

and the boundary condition (3.1) ensures that there is a unique $v^\star > 0$ satisfying this equality. Clearly, any $v_1 \geq v^\star$ satisfies (3.3). Moreover, evaluating (3.2) at $v^\star$ and rearranging, we have

$$\alpha\left(\mu\left(v^\star\right)-\mu\left(0\right)\right) \geq (1-\alpha)\left(\sigma\left(0\right)-\sigma\left(1\right)\right)\frac{p+(1-\eta)\,m}{p+m},$$

and substituting (C.1), we get

$$1 \geq \frac{p+(1-\eta)\,m}{p+m}$$

which holds as a strict inequality for all $\eta > 0$. Hence, (3.2) holds strictly at $v_1 = v^\star$ and, by continuity, also in a range of values $v_1 \in [v^\star, v^{\star\star}]$, as required. □

C.2 Dimensions of Expert Evaluation

C.2.1 Dimension 1: Data Collection

'Data Collection' clauses describe data gathering techniques, specifying what type of data is collected, and when or how it is collected and stored. Scoring was based on 1) clarity of the clauses, and 2) comprehensiveness of the data collected.

A high score meant that the data collection clauses were clear, and/or that the data collected was purposefully minimal. Sometimes these might conflict wherein a policy would state they collect comprehensive data for 'the sake of clarity and transparency – when this was the case, a policy was given a high score when the types of data collected were reasonable and in-line with industry standards and necessity.

A low score meant that the data collection clauses were either unclear, collected data so comprehensively to the point that it seemed unreasonable or excessive, or did not specify what type of data would be collected in sufficient detail that the user would not understand what data they are providing. A neutral score meant that the data collection clauses were sufficiently though not especially clear, and resembled a standard policy.

Examples:

- High score policy: Trinity Biotech – limited data collection, specific reference to their website mechanisms, implied exclusion of other types of collection
- Low score policy: Intuit – extensive data collection including location, camera, and contact data

C.2.2 Dimension 2: Consent

'Consent' clauses specified where the policy was presuming the consent of the user, and was sometimes also used to identify where the organization expressly mentioned the legal basis they relied on for data processing.

Scoring was based on how onerous the presumed consent was on the user.

A high score meant that the consent clause stated it would ask specifically for consent

for different processes, and would proactively notify the user of any changes to the policy.

A low score meant that the consent clause presumed the user's consent from their continued use (sometimes aggressively disclaiming their liability), and/or required the user to frequently check and review the policy with each use.

A neutral score meant that the consent clauses were resembled a standard policy. This often meant that consent was presumed but was not aggressively framed.

Examples:

- High score policy: WWE – no presumption of consent, will proactively inform users of changes via email or clear notice prior to the change taking effect
- Low score policy: Zynerba – presumed consent and onerous on user by requiring them to check the policy with each use

C.2.3 Dimension 3: Responsible use

'Responsible Use' clauses describe how the organization will use or interact with the data, specifying any services, security measures, marketing, or other internal use that the data will be subject to. By nature, this dimension casts a wider net than the others.

Responsible Use also covers the use of particular tracking or monitoring techniques such as cookies or other third-party software. These exist on a boundary between Data Collection and Responsible Use, but was grouped with Responsible Use because the clauses are often found separate from other data collection clauses and will detail the function of those tools.

Scoring was based on 1) whether the use was either limited and favourable for the

user or extensive and favourable for the organization, and 2) the extent of the use of additional tracking and monitoring tools.

A high score meant that the responsible use clauses proactively offered the user clear benefits and robust security assurance, and/or limited and specific use of user data. Further, a high score would indicate reasonable or restricted additional tracking and monitoring tools.

A low score meant that the responsible use clauses specified extensive use of user data, and/or subjection to heavy advertising and additional services. Further, a low score would indicate extensive additional tracking and monitoring tools.

A neutral score meant that the responsible use clauses were reasonable and resembled a standard policy.

Examples:

- High score policy: Palo Alto Networks – clear and specific explanation of use with limited use of additional tracking and monitoring tools
- Low score policy: Insight – extensive uses of user data and extensive use of additional tracking and monitoring tools including third-party tools

C.2.4 Dimension 4: Third-parties

'Third-Parties' clauses describe how the organization will share user data with third-parties, and what liability they accept or reject for that sharing. There are some categories of third-party sharing that are unavoidable, for example for the purposes of law enforcement, and then other reasons such as contract fulfillment, marketing purposes, and business interests that were assessed.

Scoring was based on 1) how clearly the sharing protocols were explained, 2) whether the sharing was restricted or not, and 3) whether the organization retained any liability or responsibility over the shared data.

A high score meant that third-party sharing was clearly explained, minimal in practice, and purely out of necessity. Further, it might indicate that the organization retained liability and responsibility over the shared data. While it is standard to disclaim liability for the actions of third-parties, some policies that scored well made an attempt to exercise some responsibility over the shared data to protect the user.

A low score meant that third-party sharing was unclear or poorly explained, leaving the user to wonder about the safety and use of their data. Further, it might indicate that sharing was extensive and not necessarily for the interest or benefit of the user. While it is standard to disclaim liability for the actions of third-parties, some policies that scored poorly did not make any attempt to protect the user or their data.

A neutral score meant that the third-party clauses were reasonable and resembled a standard policy, often disclaiming liability in unassuming terms.

Examples:

- High score policy: Vuzix – third-party sharing was clearly explained, limited to legitimate reasons for sharing, and at least attempts to impose some standards on sharing partners to protect user data in good faith
- Low score policy: Aps – extensive sharing with third-parties for marketing purposes and no attempt to impose standards on their sharing partners

C.2.5 Dimension 5: User-rights

'User-Rights' clauses describe what protection and remedies users have in response to the organization's use of their personal data. This includes clauses about data retention and deletion, information redress, requests for access, and complaint procedure. Although the GDPR does not directly apply to these American websites, the GDPR still inspired this dimension insofar as it provides users with the tools and language to understand what rights they may exercise over their data. Clauses that explained opt-out clauses were also included.

Scoring was based on 1) whether or not users were granted any rights over their data, 2) how clearly these rights were explained, and 3) how simple it was for users to exercise these rights.

A high score meant that significant and comprehensive rights were conferred onto the user over their data. It might also indicate that the rights were clearly explained and that the organization was forthcoming in providing users with a straightforward avenue to address any issues.

A low score meant that no rights were conferred at all onto the user, or if they were, they were minimal, poorly explained, or difficult and inaccessible for users to actually put in effect.

A neutral score meant that the user-rights clauses offered some reasonable form of redress that is neither particularly forthcoming nor overly minimalistic.

Examples:

- High score policy: Huron Consulting Group – rights are clearly identified and conferred onto the user. They are laid out in order and all in one place at the end

of the policy.

- Low score policy: Marcus Millichap – some user rights loosely explained throughout policy but the totality of the user's exercisable rights is unclear and not found in one place as with the vast majority of other policies.

C.2.6 Dimension 6: Overall

This dimension is the culmination of the other dimensions, and introduces perhaps the most interesting part of the data but also the most subjective to human error. The entire policy was considered as a whole, more than simply the sum of the other dimensions, because it accounted for each policy's tone, clarity, and style, all assessed from the perspective of whether a lay-user would be able to understand the policy and be able to exercise their data rights from it.

Scoring was based on 1) the overarching tone and legibility of the policy, and 2) how many of the other dimensions were positively or negatively scored.

Most policies that had a high overall score were clear throughout and had minimal shortcomings, whereas most policies that poor overall were lacking throughout.

There was room for discrepancy in some policies that received mixed evaluations across the dimensions. For example, a policy with too many negatively scoring dimensions would have been difficult to read or use as a whole, but where only one aspect of an otherwise good policy fell-short, it could still be a relatively effective policy for users.

By contrast, a policy that was relatively clear or at least measured up to standard in most dimensions, but was very poor in a crucial dimension such as user-rights, the policy may be scored negatively overall because it would be difficult for a user to

apply towards protecting their data rights.

Examples:

- High score policy: Image Sensing – clearly laid out with different sections for each crucial aspect of the policy
- Low score policy: Tabularasa Healthcare – policy offers almost nothing meaningful for the user and is only there to vaguely disclaim any liability

C.3 Legal Quality: Overall Score and Individual Categories

Table C.1. Overall Score and Individual Categories

	(1) Overall
Data Collection	0.125***
	(3.89)
Consent	0.337***
	(8.17)
Responsible Use	0.241***
	(7.21)
Third-Parties	0.317***
	(9.38)
User-Rights	0.404***
	(15.99)
Constant	-0.020
	(-1.05)
Observations	356
R^2	0.806

Note: t-statistics in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$

C.4 Variation in Policy Text with LSA

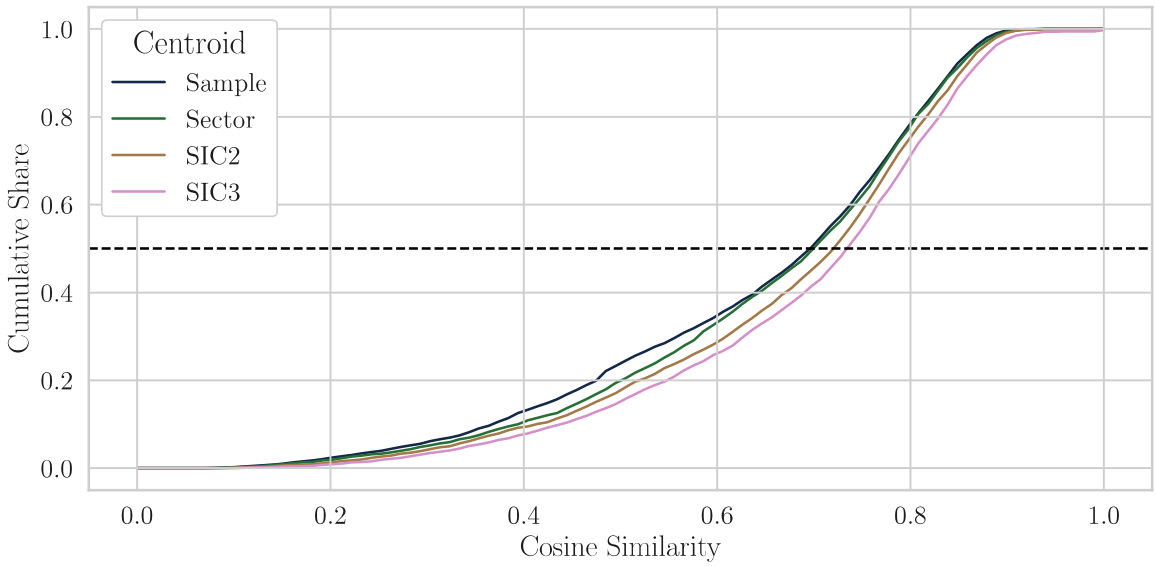
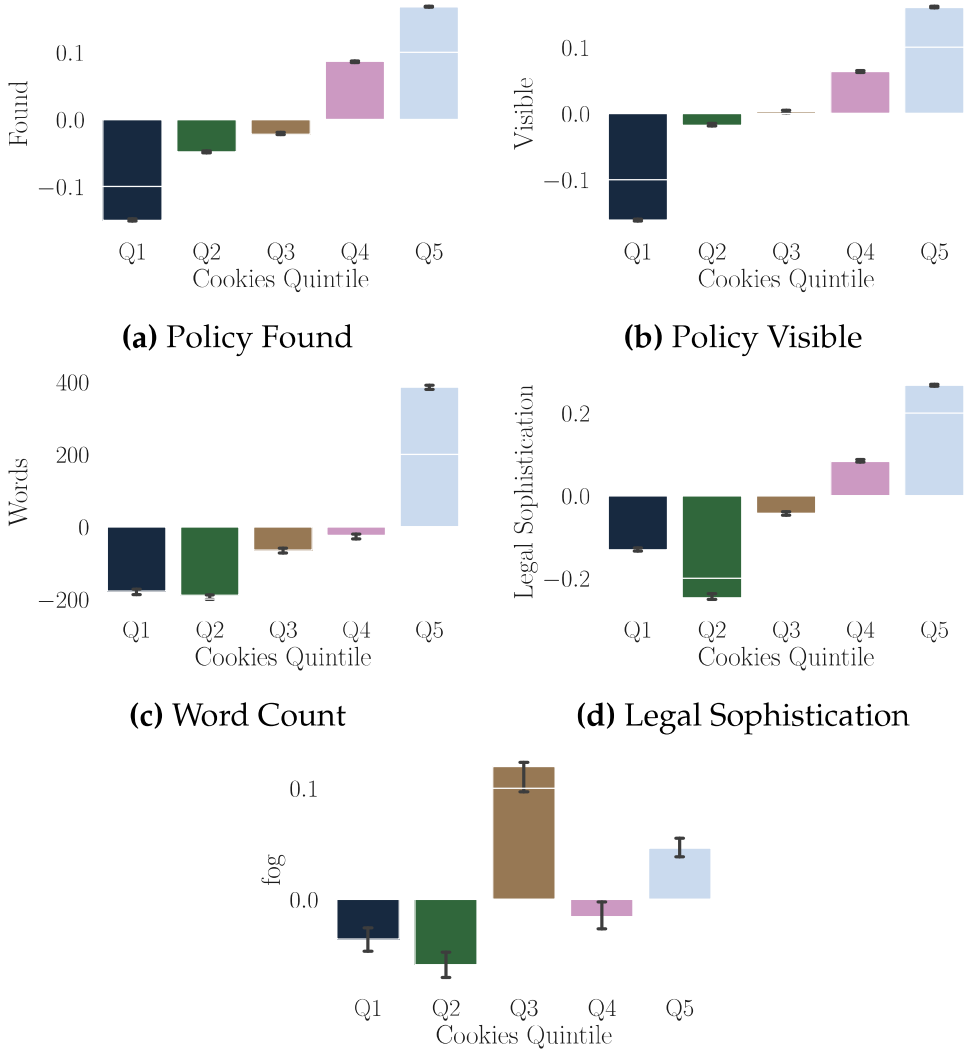


Figure C.1. Variation in Policy Text in 100 Latent Dimensions

C.5 Additional Results on Policy Attributes



(a) Policy Found

(b) Policy Visible

(c) Word Count

(d) Legal Sophistication

(e) Gunning Fog Index

Figure C.2. Privacy Policies and Cookies

Table C.2. Policy Attributes: Controlling for Policy Length

(a) Without Sector Fixed Effects

	(1) Policy Visible	(2) Legal Sophistication	(3) Fog Index	(4) Cookies
Log Market Value	0.018***	0.045***	0.035	0.792***
	(4.35)	(5.09)	(1.42)	(7.82)
Knowledge Share	0.474***	1.682***	-1.622**	8.454***
	(4.10)	(6.80)	(-2.36)	(2.98)
Knowledge Share^2	-0.655***	-2.728***	2.226**	-16.003***
	(-3.47)	(-6.77)	(1.98)	(-3.46)
Log Market Share	-0.012***	0.028***	0.023	0.218**
	(-3.53)	(3.69)	(1.09)	(2.50)
Log Words	-0.024***	0.384***	0.883***	1.377***
	(-3.55)	(26.82)	(22.14)	(8.37)
Observations	3918	3918	3918	3918

(b) With Sector Fixed Effects

	(1) Policy Visible	(2) Legal Sophistication	(3) Fog Index	(4) Cookies
Log Market Value	0.013***	0.036*	0.040	0.795
	(4.96)	(2.11)	(0.74)	(1.81)
Knowledge Share	0.240**	1.689***	-0.561	12.272**
	(2.42)	(6.37)	(-0.40)	(2.46)
Knowledge Share^2	-0.351***	-2.601***	1.279	-18.362**
	(-3.37)	(-6.60)	(1.06)	(-2.27)
Log Market Share	-0.008	0.038***	0.033	0.233
	(-1.25)	(4.85)	(0.37)	(0.80)
Log Words	-0.024	0.380***	0.879***	1.209**
	(-1.36)	(4.76)	(7.71)	(2.86)
Observations	3918	3918	3918	3918

(c) Controlling For Cookies

	(1) Policy Visible	(2) Policy Visible	(3) Legal Sophistication	(4) Legal Sophistication	(5) Fog Index	(6) Fog Index
Cookies	0.005***	0.004***	0.006***	0.005***	-0.020***	-0.017***
	(7.59)	(6.98)	(4.49)	(3.43)	(-5.08)	(-5.19)
Log Market Value	0.014***	0.009***	0.040***	0.032	0.050**	0.053
	(3.40)	(3.63)	(4.51)	(1.75)	(2.04)	(1.17)
Knowledge Share	0.433***	0.187	1.629***	1.630***	-1.456**	-0.350
	(3.76)	(1.66)	(6.60)	(6.76)	(-2.12)	(-0.24)
Knowledge Share^2	-0.576***	-0.272*	-2.628***	-2.512***	1.912*	0.963
	(-3.07)	(-2.06)	(-6.53)	(-7.12)	(1.71)	(0.75)
Log Market Share	-0.014***	-0.009	0.027***	0.037***	0.027	0.037
	(-3.85)	(-1.58)	(3.52)	(5.04)	(1.29)	(0.43)
Log Words	-0.031***	-0.029	0.375***	0.374***	0.910***	0.900***
	(-4.55)	(-1.68)	(26.05)	(4.72)	(22.69)	(7.90)
Sector FE	No	Yes	No	Yes	No	Yes
N	3,918	3,918	3,918	3,918	3,918	3,918

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b) and (c).

Table C.3. Policy Attributes: Further Firm Controls

(a) Without Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.044***	0.051***	-0.009	0.044***	0.021	0.838***
	(12.25)	(12.30)	(-0.90)	(4.39)	(0.79)	(10.38)
Knowledge Share	0.715***	0.593***	1.083***	1.676***	-0.417	15.675***
	(5.21)	(3.74)	(2.93)	(4.62)	(-0.43)	(5.06)
Knowledge Share^2	-0.596***	-0.531**	-1.343**	-2.689***	-0.043	-20.689***
	(-3.03)	(-2.34)	(-2.53)	(-5.17)	(-0.03)	(-4.66)
Zero Knowledge Capital	-0.023	-0.018	-0.170***	-0.153***	-0.185	0.906**
	(-1.35)	(-0.93)	(-3.70)	(-3.40)	(-1.52)	(2.39)
Log Market Share	0.014***	-0.012***	0.086***	0.057***	0.096***	0.355***
	(4.63)	(-3.47)	(9.82)	(6.69)	(4.17)	(5.14)
Marketing / Assets	0.328	0.125	2.936***	1.399**	0.669	20.580***
	(1.51)	(0.50)	(5.11)	(2.48)	(0.44)	(4.20)
Marketing Missing	-0.045***	-0.125***	-0.024	-0.049	0.798***	-3.483***
	(-3.61)	(-8.62)	(-0.69)	(-1.46)	(8.81)	(-12.29)
Market to book	-0.007**	-0.013***	0.034***	0.002	0.048**	0.097
	(-2.44)	(-3.86)	(3.94)	(0.19)	(2.11)	(1.47)
Observations	5109	5109	3902	3902	3902	5109

(b) With Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.046***	0.047***	0.011	0.042*	0.037	0.903***
	(7.94)	(6.28)	(0.42)	(1.93)	(0.58)	(3.45)
Knowledge Share	0.651***	0.483***	0.907**	1.647***	-0.094	14.708**
	(13.83)	(4.16)	(2.95)	(3.83)	(-0.07)	(2.98)
Knowledge Share^2	-0.514***	-0.401**	-1.087***	-2.511***	0.080	-18.460**
	(-8.05)	(-2.69)	(-3.83)	(-4.44)	(0.06)	(-2.47)
Zero Knowledge Capital	-0.007	0.009	-0.119**	-0.175	-0.326*	0.774*
	(-0.90)	(0.28)	(-2.27)	(-1.78)	(-1.94)	(2.06)
Log Market Share	0.011**	-0.009	0.066*	0.060***	0.093	0.270*
	(2.59)	(-0.87)	(2.06)	(5.12)	(1.02)	(1.87)
Marketing / Assets	0.195	0.203	2.347***	1.153***	-0.001	15.414
	(1.04)	(0.51)	(4.70)	(5.79)	(-0.00)	(1.16)
Marketing Missing	-0.029	-0.085*	-0.029	-0.004	0.687**	-3.017***
	(-1.50)	(-1.85)	(-0.32)	(-0.13)	(2.37)	(-6.34)
Market to book	-0.010**	-0.014**	0.019	-0.006	0.034	-0.011
	(-2.70)	(-2.62)	(1.20)	(-0.76)	(0.90)	(-0.18)
Observations	5109	5109	3902	3902	3902	5109

(c) Controlling For Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
Cookies	0.007***	0.007***	0.008***	0.008***	0.012***	0.011***	0.011***	0.009***	-0.000	-0.001
	(11.69)	(7.12)	(11.35)	(9.62)	(7.63)	(4.58)	(6.99)	(5.21)	(-0.07)	(-0.58)
Log Market Value	0.038***	0.040***	0.044***	0.040***	-0.019*	0.001	0.036***	0.034	0.022	0.038
	(10.60)	(6.94)	(10.69)	(5.29)	(-1.83)	(0.05)	(3.54)	(1.53)	(0.79)	(0.62)
Knowledge Share	0.603***	0.553***	0.467***	0.373**	0.895**	0.739**	1.506***	1.505***	-0.412	-0.074
	(4.44)	(8.28)	(2.98)	(3.23)	(2.43)	(2.35)	(4.16)	(3.64)	(-0.42)	(-0.06)
Knowledge Share^2	-0.448**	-0.390***	-0.364	-0.262*	-1.091**	-0.877**	-2.462***	-2.334***	-0.049	0.055
	(-2.30)	(-5.38)	(-1.62)	(-2.03)	(-2.07)	(-3.01)	(-4.75)	(-4.36)	(-0.03)	(0.04)
Zero Knowledge Capital	-0.029*	-0.012*	-0.025	0.003	-0.188***	-0.132**	-0.170***	-0.186*	-0.184	-0.325*
	(-1.76)	(-1.88)	(-1.32)	(0.10)	(-4.12)	(-2.44)	(-3.78)	(-1.87)	(-1.51)	(-1.95)
Log Market Share	0.012***	0.009*	-0.015***	-0.011	0.081***	0.063*	0.053***	0.057***	0.096***	0.093
	(3.84)	(2.24)	(-4.32)	(-1.08)	(9.35)	(1.91)	(6.23)	(5.20)	(4.17)	(1.02)
Marketing / Assets	0.180	0.092	-0.041	0.087	2.662***	2.145***	1.151**	0.984***	0.676	0.023
	(0.84)	(0.59)	(-0.16)	(0.26)	(4.66)	(4.21)	(2.05)	(5.14)	(0.44)	(0.01)
Marketing Missing	-0.020	-0.009	-0.097***	-0.062	0.023	0.007	-0.007	0.026	0.797***	0.682**
	(-1.63)	(-0.56)	(-6.67)	(-1.45)	(0.66)	(0.08)	(-0.22)	(0.77)	(8.66)	(2.41)
Market to book	-0.008***	-0.010**	-0.014***	-0.014**	0.033***	0.020	0.000	-0.006	0.048**	0.034
	(-2.71)	(-2.73)	(-4.14)	(-2.75)	(3.82)	(1.19)	(0.05)	(-0.70)	(2.11)	(0.90)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	5,109	5,109	5,109	5,109	3,902	3,902	3,902	3,902	3,902	3,902

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b)and (c).

Table C.4. Policy Attributes: Excluding Manufacturing Firms

(a) Without Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.042***	0.052***	-0.032**	0.025**	-0.038	1.159***
	(9.29)	(10.20)	(-2.47)	(2.39)	(-1.11)	(11.22)
Knowledge Share	1.205***	0.971***	4.722***	4.361***	5.190***	47.836***
	(5.38)	(3.81)	(7.91)	(8.84)	(3.28)	(9.28)
Knowledge Share^2	-1.351***	-1.328**	-6.668***	-6.100***	-7.560**	-75.825***
	(-2.63)	(-2.27)	(-4.71)	(-5.21)	(-2.01)	(-6.41)
Log Market Share	0.012***	-0.020***	0.107***	0.058***	0.176***	0.150*
	(3.27)	(-4.81)	(10.05)	(6.62)	(6.21)	(1.75)
Observations	3380	3380	2515	2515	2515	3380

(b) With Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.043***	0.043**	-0.003	0.023	-0.015	1.279***
	(5.06)	(3.20)	(-0.08)	(0.68)	(-0.15)	(4.21)
Knowledge Share	0.673***	0.422	2.754***	2.979***	4.006	30.761***
	(4.90)	(1.80)	(4.16)	(14.07)	(1.70)	(5.54)
Knowledge Share^2	-0.635*	-0.622	-3.817***	-4.178***	-5.466	-53.162***
	(-2.25)	(-1.54)	(-4.18)	(-6.48)	(-1.63)	(-5.15)
Log Market Share	0.009	-0.013	0.075	0.058**	0.169	-0.048
	(1.40)	(-0.69)	(1.62)	(2.60)	(1.07)	(-0.34)
Observations	3380	3380	2515	2515	2515	3380

(c) Controlling For Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
Cookies	0.008***	0.007***	0.010***	0.009***	0.012***	0.010**	0.010***	0.008***	-0.013**	-0.008
	(11.15)	(6.01)	(11.76)	(7.90)	(6.18)	(3.17)	(6.42)	(4.76)	(-2.54)	(-1.60)
Log Market Value	0.032***	0.034***	0.041***	0.032**	-0.045***	-0.016	0.014	0.012	-0.023	-0.005
	(7.18)	(5.66)	(7.99)	(3.00)	(-3.48)	(-0.49)	(1.30)	(0.35)	(-0.67)	(-0.05)
Knowledge Share	0.813***	0.444**	0.502**	0.156	4.185***	2.449**	3.900***	2.732***	5.780***	4.250
	(3.65)	(3.22)	(1.98)	(0.66)	(6.99)	(3.23)	(7.88)	(16.35)	(3.62)	(1.73)
Knowledge Share^2	-0.730	-0.239	-0.584	-0.162	-5.782***	-3.264**	-5.339***	-3.731***	-8.534**	-5.908
	(-1.44)	(-0.86)	(-1.01)	(-0.40)	(-4.09)	(-3.06)	(-4.57)	(-6.65)	(-2.26)	(-1.68)
Log Market Share	0.011***	0.009	-0.022***	-0.013	0.106***	0.076	0.057***	0.059**	0.178***	0.168
	(2.99)	(1.71)	(-5.26)	(-0.71)	(9.96)	(1.61)	(6.51)	(2.56)	(6.28)	(1.08)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	3,380	3,380	3,380	3,380	2,515	2,515	2,515	2,515	2,515	2,515

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b) and (c).

Table C.5. Policy Attributes: Service Sector Only

(a) Without Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.033***	0.025***	0.047**	0.100***	0.089	1.469***
	(4.32)	(2.59)	(2.12)	(4.72)	(1.53)	(6.03)
Knowledge Share	0.644**	0.606*	1.830**	2.451***	1.962	33.662***
	(2.54)	(1.92)	(2.55)	(3.58)	(1.04)	(4.14)
Knowledge Share^2	-0.499	-0.573	-2.579*	-2.791**	-2.930	-54.712***
	(-0.96)	(-0.88)	(-1.75)	(-1.99)	(-0.76)	(-3.28)
Log Market Share	0.018**	0.020**	0.013	0.019	0.022	0.058
	(2.51)	(2.25)	(0.63)	(0.97)	(0.40)	(0.25)
Observations	730	730	617	617	617	730

(b) With Sector Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.012**	0.012	0.036*	0.077***	0.052	1.141***
	(2.60)	(1.62)	(1.95)	(3.89)	(0.75)	(6.17)
Knowledge Share	0.595***	0.561*	1.760***	2.254***	2.124***	33.482***
	(6.02)	(1.92)	(4.76)	(5.67)	(3.63)	(8.75)
Knowledge Share^2	-0.363	-0.468	-2.417***	-2.550**	-3.492***	-53.556***
	(-1.35)	(-0.61)	(-3.36)	(-2.86)	(-4.52)	(-3.98)
Log Market Share	0.043***	0.036***	0.025	0.045**	0.056	0.449*
	(5.28)	(4.29)	(1.09)	(2.77)	(0.82)	(2.07)
Observations	730	730	617	617	617	730

(c) Controlling For Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
Cookies	0.005***	0.005***	0.008***	0.008***	0.012***	0.012***	0.008**	0.007***	-0.002	-0.003
	(4.61)	(12.23)	(5.66)	(11.64)	(3.96)	(6.11)	(2.58)	(3.47)	(-0.21)	(-0.55)
Log Market Value	0.025***	0.006	0.013	0.002	0.029	0.021	0.089***	0.068**	0.091	0.056
	(3.26)	(1.18)	(1.34)	(0.26)	(1.32)	(1.19)	(4.13)	(3.03)	(1.54)	(0.76)
Knowledge Share	0.466*	0.424***	0.336	0.289	1.428**	1.377***	2.200***	2.017***	2.018	2.228***
	(1.84)	(4.65)	(1.07)	(1.10)	(1.99)	(3.81)	(3.20)	(4.87)	(1.06)	(3.24)
Knowledge Share^2	-0.211	-0.090	-0.134	-0.033	-1.896	-1.769**	-2.363*	-2.148**	-3.027	-3.667***
	(-0.41)	(-0.41)	(-0.21)	(-0.05)	(-1.30)	(-2.33)	(-1.68)	(-2.63)	(-0.78)	(-4.27)
Log Market Share	0.018**	0.041***	0.019**	0.032***	0.014	0.023	0.020	0.044**	0.022	0.057
	(2.50)	(5.00)	(2.25)	(4.06)	(0.65)	(1.01)	(0.98)	(2.54)	(0.40)	(0.84)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	730	730	730	730	617	617	617	617	617	617

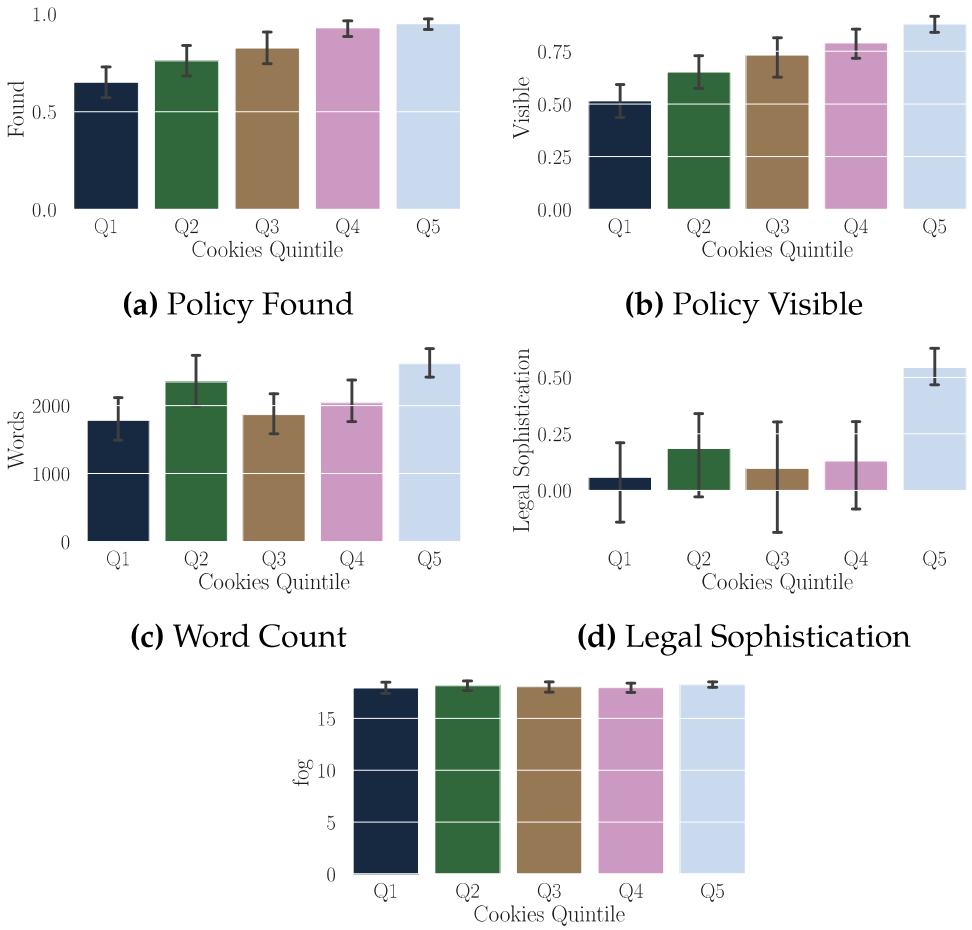
Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the SIC2 industry level in panel (b) and (c).

Table C.6. Policy Attributes: 2-digit SIC Code Fixed Effects

	(1) Policy Found	(2) Policy Visible	(3) Log Words	(4) Legal Sophistication	(5) Fog Index	(6) Cookies
Log Market Value	0.034***	0.033***	0.035**	0.041	0.077*	0.752***
	(5.26)	(5.75)	(2.39)	(1.65)	(1.71)	(5.14)
Knowledge Share	0.691***	0.588***	0.994***	1.823***	-0.059	12.843**
	(7.56)	(4.10)	(3.00)	(3.64)	(-0.06)	(2.38)
Knowledge Share^2	-0.536***	-0.607***	-1.177***	-2.521***	0.218	-14.684**
	(-4.88)	(-3.24)	(-3.01)	(-4.49)	(0.18)	(-2.04)
Log Market Share	0.024***	0.008	0.039***	0.066***	0.036	0.455***
	(4.20)	(1.03)	(2.67)	(3.19)	(0.80)	(3.90)
Observations	5140	5140	3918	3918	3918	5140

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the SIC2 industry level.

C.6 Policy Attributes: Service Sector



(a) Policy Found

(b) Policy Visible

(c) Word Count

(d) Legal Sophistication

(e) Gunning Fog Index

Figure C.3. Privacy Policies and Cookies

C.7 First and Third-Party Data Extraction

Table C.7. First and Third-Party Data Extraction

(a) Without Sector Fixed Effects

	(1) Cookies	(2) 1^{st}-Party Trackers	(3) 3^{rd}-Party Trackers	(4) TP Requests
Log Market Value	0.857***	0.555***	0.323***	1.762***
	(10.86)	(10.96)	(8.07)	(5.18)
Knowledge Share	14.562***	8.921***	4.614***	62.430***
	(6.25)	(5.99)	(3.92)	(6.24)
Knowledge Share^2	-22.301***	-13.729***	-7.333***	-99.005***
	(-5.87)	(-5.65)	(-3.82)	(-6.06)
Log Market Share	0.316***	0.208***	0.123***	1.206***
	(4.75)	(4.86)	(3.64)	(4.20)
Observations	5140	4983	4983	4983

(b) With Sector Fixed Effects

	(1) Cookies	(2) 1^{st}-Party Trackers	(3) 3^{rd}-Party Trackers	(4) TP Requests
Log Market Value	0.880**	0.588**	0.323*	2.140
	(2.53)	(2.79)	(1.96)	(1.38)
Knowledge Share	14.321**	8.436**	5.155***	57.188***
	(2.58)	(2.49)	(3.52)	(3.32)
Knowledge Share^2	-20.522**	-12.430**	-7.106***	-86.596**
	(-2.43)	(-2.45)	(-3.41)	(-3.19)
Log Market Share	0.261	0.149	0.114	0.739
	(1.13)	(1.07)	(1.03)	(0.95)
Observations	5140	4983	4983	4983

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Table C.8. Controlling For First and Third-Party Cookies

(a) All Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
Cookies	0.007***	0.007***	0.009***	0.008***	0.013***	0.011***	0.011***	0.009***	-0.008**	-0.007
	(12.45)	(6.58)	(12.89)	(8.10)	(8.37)	(4.10)	(7.40)	(5.39)	(-1.96)	(-1.47)
Log Market Value	0.036***	0.037***	0.041***	0.035***	-0.016	0.007	0.034***	0.034	0.036	0.060
	(10.39)	(8.31)	(10.25)	(5.29)	(-1.63)	(0.30)	(3.54)	(1.46)	(1.37)	(0.91)
Knowledge Share	0.738***	0.560***	0.564***	0.347*	2.255***	1.341**	2.475***	2.132***	0.595	0.857
	(7.34)	(7.77)	(4.85)	(2.02)	(8.31)	(2.31)	(9.34)	(6.63)	(0.82)	(0.46)
Knowledge Share^2	-0.647***	-0.438***	-0.593***	-0.316**	-2.567***	-1.560**	-3.591***	-3.096***	-0.423	-0.441
	(-3.94)	(-6.51)	(-3.12)	(-2.28)	(-5.77)	(-2.55)	(-8.27)	(-6.65)	(-0.36)	(-0.26)
Log Market Share	0.013***	0.012**	-0.013***	-0.008	0.083***	0.058*	0.058***	0.059***	0.103***	0.090
	(4.66)	(2.62)	(-4.01)	(-0.61)	(10.04)	(1.87)	(7.15)	(5.39)	(4.66)	(0.80)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	5,140	5,140	5,140	5,140	3,918	3,918	3,918	3,918	3,918	3,918

(b) First-Party Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
1st-Party Trackers	0.011***	0.010***	0.012***	0.010***	0.019***	0.016***	0.017***	0.014***	-0.019***	-0.017
	(11.75)	(6.82)	(10.50)	(5.62)	(7.51)	(3.25)	(6.90)	(5.71)	(-2.84)	(-1.57)
Log Market Value	0.036***	0.037***	0.042***	0.037***	-0.016	0.007	0.035***	0.035	0.042	0.066
	(10.45)	(7.51)	(10.34)	(5.60)	(-1.58)	(0.36)	(3.54)	(1.53)	(1.57)	(1.11)
Knowledge Share	0.718***	0.556***	0.569***	0.370*	2.243***	1.338**	2.499***	2.159***	0.630	0.952
	(7.10)	(6.17)	(4.83)	(1.92)	(8.19)	(2.37)	(9.24)	(6.54)	(0.86)	(0.50)
Knowledge Share^2	-0.635***	-0.436***	-0.596***	-0.335*	-2.575***	-1.566**	-3.639***	-3.141***	-0.534	-0.627
	(-3.85)	(-4.40)	(-3.11)	(-1.95)	(-5.74)	(-2.62)	(-8.21)	(-6.54)	(-0.44)	(-0.36)
Log Market Share	0.013***	0.012**	-0.012***	-0.006	0.084***	0.060*	0.058***	0.060***	0.105***	0.091
	(4.38)	(2.44)	(-3.45)	(-0.50)	(10.01)	(2.02)	(7.07)	(5.74)	(4.68)	(0.80)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	4,983	4,983	4,983	4,983	3,822	3,822	3,822	3,822	3,822	3,822

(c) Third-Party Cookies

	(1) Policy Found	(2) Policy Found	(3) Policy Visible	(4) Policy Visible	(5) Log Words	(6) Log Words	(7) Legal Sophistication	(8) Legal Sophistication	(9) Fog Index	(10) Fog Index
3rd-Party Trackers	0.011***	0.010***	0.015***	0.013***	0.021***	0.018***	0.018***	0.014***	-0.010	-0.007
	(9.22)	(7.26)	(10.50)	(16.94)	(6.71)	(3.79)	(5.89)	(4.10)	(-1.18)	(-1.12)
Log Market Value	0.039***	0.040***	0.043***	0.039***	-0.012	0.011	0.038***	0.038	0.035	0.058
	(11.21)	(8.59)	(10.82)	(5.87)	(-1.22)	(0.51)	(3.90)	(1.69)	(1.32)	(0.88)
Knowledge Share	0.767***	0.591***	0.605***	0.391**	2.305***	1.378**	2.558***	2.199***	0.527	0.844
	(7.56)	(9.45)	(5.15)	(2.38)	(8.42)	(2.36)	(9.46)	(6.08)	(0.72)	(0.47)
Knowledge Share^2	-0.707***	-0.493***	-0.648***	-0.373**	-2.661***	-1.628**	-3.723***	-3.203***	-0.372	-0.468
	(-4.27)	(-7.89)	(-3.39)	(-2.76)	(-5.93)	(-2.59)	(-8.39)	(-6.13)	(-0.31)	(-0.29)
Log Market Share	0.014***	0.012**	-0.011***	-0.006	0.085***	0.060*	0.060***	0.060***	0.102***	0.089
	(4.69)	(2.56)	(-3.28)	(-0.52)	(10.18)	(2.01)	(7.24)	(5.36)	(4.55)	(0.79)
Sector FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	4,983	4,983	4,983	4,983	3,822	3,822	3,822	3,822	3,822	3,822

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level.

C.8 Residuals

We first regress both legal sophistication and the number of cookies on a broad set of attributes. We then regress the residuals on each other.

Table C.9. Policy Attributes and Cookies

(a) First Stage

	(1) Legal Sophistication	(2) Cookies
Log Market Value	0.044***	0.838***
	(4.39)	(10.38)
Knowledge Share	1.676***	15.675***
	(4.62)	(5.06)
Knowledge Share^2	-2.689***	-20.689***
	(-5.17)	(-4.66)
Zero Knowledge Capital	-0.153***	0.906**
	(-3.40)	(2.39)
Log Market Share	0.057***	0.355***
	(6.69)	(5.14)
Marketing / Assets	1.399**	20.580***
	(2.48)	(4.20)
Marketing Missing	-0.049	-3.483***
	(-1.46)	(-12.29)
Market to book	0.002	0.097
	(0.19)	(1.47)
Observations	3902	5109

(b) Second Stage

	(1) Legal Sophistication Residuals
Cookies Residuals	0.011***
	(6.99)
Constant	-0.008
	(-0.54)
Observations	3902

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01.

C.9 Alternative Measure of Legal Clarity

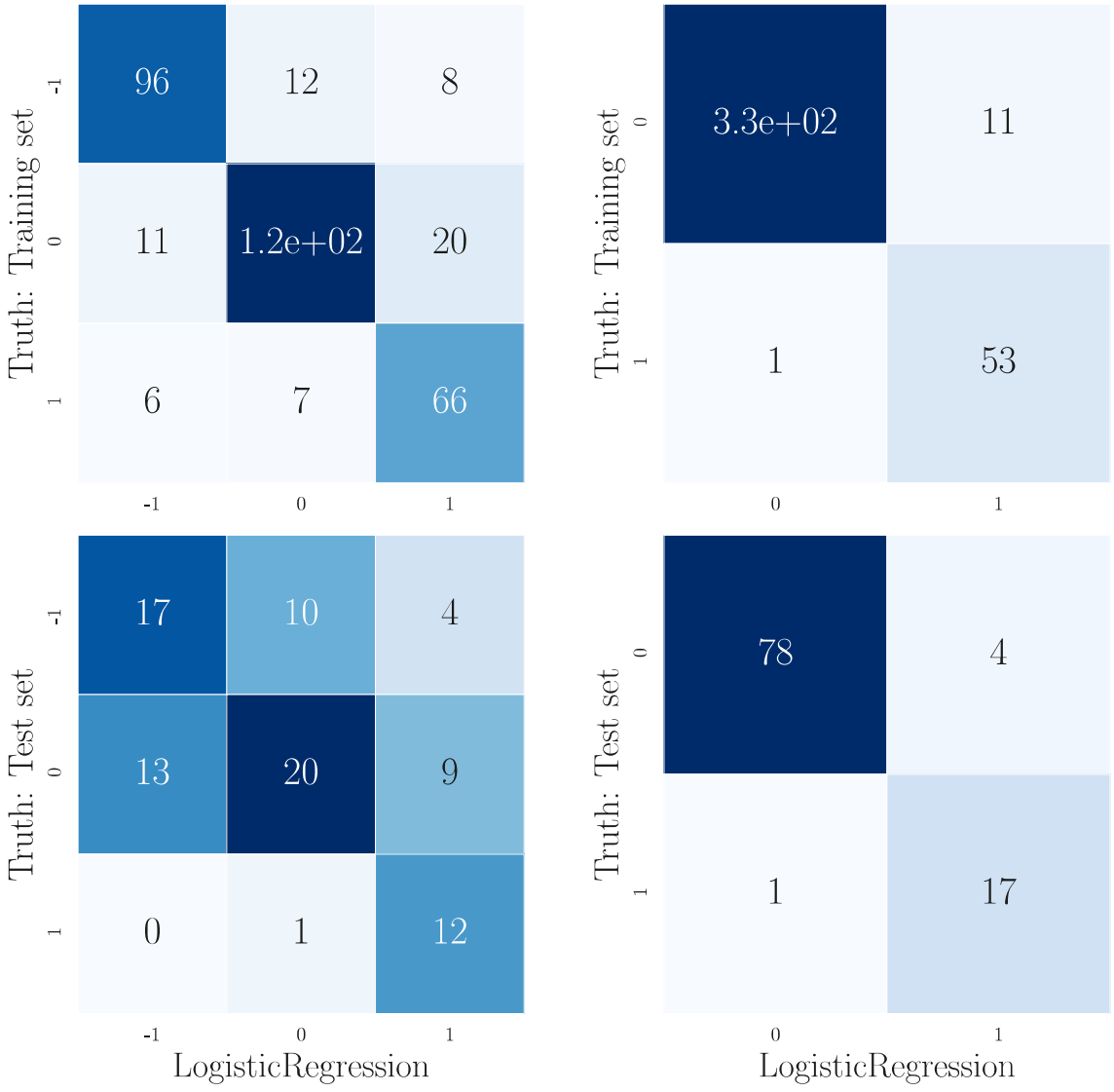
We fit a penalized logistic regression model which predicts the Overall score (High, Medium or Low) that our expert attached to each policy in the reviewed subsample, as a function of the TF.IDF weights on the most common 5000 words in our corpus of documents. We split the reviewed sample into a training set containing 80% ($N = 341$) of the observations, and a test set containing the remaining 20% ($N = 86$). On the training set, we first use SMOTE oversampling (Chawla, Bowyer, Hall, and Kegelmeyer, 2002) in order to obtain a balanced sample of High-, Medium- and Low-scoring policies. On this training set, we estimate a three-class logistic model, with an L_2 regularization penalty that is proportional to the squared norm of the parameter vector.[1] Based on a cross-validation exercise, we choose the weight on the regularization penalty to be equal to $C = 1$. Then, we check the performance of the classifier on the test set, and extrapolate its predictions to all privacy policies that our expert did not evaluate. In addition, we use the same methodology to predict a dummy variable provided by the expert to indicate policies that do not contain privacy-relevant text. Figure C.4 shows the performance of both classifiers on the training and test data, in terms of a "confusion matrix" that compares true scores (in the rows of the matrix) with predicted scores (in the columns).

Out of sample, our classifier is quite good at distinguishing High scores form Medium and Low ones, but not very good at distinguishing Medium and Low scores from each other. Hence, we use the predicted probability that the score is High as our alternative measure of clarity.

Table C.10 shows a regression of these alternative measures on our baseline set of firm

[1] We use the *sklearn* package in Python to fit this model. For further details, see: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

characteristics. We consider two cases: In the first column, we include all policies in our sample. In the second column, we exclude policies with a probability greater than 50% of not containing relevant text.



(a) Predicting Overall Score **(b)** Predicting "irrelevant" dummy

Figure C.4. Performance of Classifiers

Table C.10. Policy Attributes: Alternative Legal Clarity Measure

(a) Without Sector Fixed Effects

	(1) Pr[High Score]	(2) Pr[High Score]
Log Market Value	0.005	0.005
	(0.92)	(0.96)
Knowledge Share	0.829***	0.767***
	(4.38)	(3.48)
Knowledge Share^2	-1.060***	-1.007***
	(-4.23)	(-3.53)
Log Market Share	0.012**	0.012*
	(2.31)	(1.99)
Filtered	No	Yes
N	3,918	3,099

(b) With Sector Fixed Effects

	(1) Pr[High Score]	(2) Pr[High Score]
Log Market Value	0.008	0.008
	(1.53)	(1.75)
Knowledge Share	0.521***	0.495**
	(3.94)	(3.11)
Knowledge Share^2	-0.707***	-0.687***
	(-4.38)	(-3.51)
Log Market Share	0.009	0.009
	(1.77)	(1.75)
Filtered	No	Yes
N	3,918	3,099

Note: t-statistics in parentheses. * p<0.10, ** p<0.05, *** p<0.01. Standard errors are clustered at the Sector level in panel (b).

Bibliography

Acquisti, A., L. Brandimarte, and G. Loewenstein, 2015, "Privacy and human behavior in the age of information," *Science*, 347(6221), 509–514.

Acquisti, A., L. K. John, and G. Loewenstein, 2013, "What is privacy worth?," *The Journal of Legal Studies*, 42(2), 249–274.

Acquisti, A., C. Taylor, and L. Wagman, 2016, "The economics of privacy," *Journal of Economic Literature*, 54(2), 442–92.

Adjerid, I., A. Acquisti, L. Brandimarte, and G. Loewenstein, 2013, "Sleights of privacy: Framing, disclosures, and the limits of transparency," in *Proceedings of the ninth symposium on usable privacy and security*, pp. 1–11.

AFP, 2023, "Riots erupt in Nigerian cities as bank policy leads to scarcity of cash," *The Guardian*.

Agarwal, S., W. Kuang, L. Wang, and Y. Yang, 2020, "The Role of Agents in Fraudulent Activities: Evidence from the Housing Market in China," .

Agarwal, S., K. Li, Y. Qin, J. Wu, and J. Yan, 2020, "Tax evasion, capital gains taxes, and the housing market," *Journal of Public Economics*, 188, 104222.

Ahn, S. C., and A. R. Horenstein, 2013, "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81(3), 1203–1227.

Alvarez, F., and D. Argente, 2022, "On the effects of the availability of means of payments: The case of uber," *The Quarterly Journal of Economics*, 137(3), 1737–1789.

Amornsiripanitch, N., 2020, "Why Are Residential Property Tax Rates Regressive?," *Available at SSRN 3729072*.

Amos, R., G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, 2020, "Privacy Policies over Time: Curation andAnalysis of a Million-Document Dataset," *arXiv preprint arXiv:2008.09159*.

Anagol, S., F. V. Ferreira, and J. M. Rexer, 2021, "Estimating the Economic Value of Zoning Reform," working paper, National Bureau of Economic Research.

Anagol, S., B. B. Lockwood, A. Davids, and T. Ramadorai, 2022, "Diffuse Bunching with Frictions: Theory and Estimation," .

Andersen, S., C. Badarinza, L. Liu, J. Marx, and T. Ramadorai, 2022, "Reference dependence in the housing market," *American Economic Review*, 112(10), 3398–3440.

Antweiler, W., and M. Z. Frank, 2004, "Is all that talk just noise? The information content of Internet stock message boards," *Journal of Finance*, 59(3), 1259–1294.

Artavanis, N., A. Morse, and M. Tsoutsoura, 2016, "Measuring income tax evasion using bank credit: Evidence from Greece," *The Quarterly Journal of Economics*, 131(2), 739–798.

Athey, S., C. Catalini, and C. Tucker, 2017, "The digital privacy paradox: Small money, small costs, small talk," working paper, National Bureau of Economic Research.

Avenancio-León, C. F., and T. Howard, 2022, "The assessment gap: Racial inequalities in property taxation," *The Quarterly Journal of Economics*, 137(3), 1383–1434.

Baker, M., and J. Wurgler, 2006, "Investor sentiment and the cross-section of stock returns," *Journal of Finance*, 61(4), 1645–1680.

Balasubramaniam, V., J. Y. Campbell, T. Ramadorai, and B. Ranish, 2022, "Who Owns What? A Factor Model for Direct Stock Holding," *Journal of Finance (forthcoming)*.

Barber, B. M., X. Huang, T. Odean, and C. Schwarz, 2022, "Attention-Induced Trading and Returns: Evidence from Robinhood Users," *The Journal of Finance*, 77(6), 3141–3190.

Barber, B. M., and T. Odean, 2008, "All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors," *Review of Financial Studies*, 21(2), 785–818.

Barrot, J. N., R. Kaniel, and D. Sraer, 2016, "Are retail traders compensated for providing liquidity?," *Journal of Financial Economics*, 120(1), 146–168.

Bartlett, R., J. McCrary, and O. Maureen, 2022, "A Fractional Solution to a Stock Market Mystery," .

Basu, K., 2015, *An economist in the real world: the art of policymaking in India*. MIT Press.

Bergemann, D., and A. Bonatti, 2018, "Markets for information: An introduction," .

Best, M. C., and H. J. Kleven, 2018, "Housing market responses to transaction taxes: Evidence from notches and stimulus in the UK," *The Review of Economic Studies*, 85(1), 157–193.

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang, 2021, "Tracking Retail Investor Activity," *Journal of Finance*, 76(5), 2249–2305.

Bonneau, J., and S. Preibusch, 2010, "The privacy jungle: On the market for data protection in social networks," in *Economics of information security and privacy*. Springer, pp. 121–167.

Braguinsky, S., S. Mityakov, and A. Liscovich, 2014, "Direct estimation of hidden earnings: Evidence from Russian administrative data," *The Journal of Law and Economics*, 57(2), 281–319.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, 2020, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Bryzgalova, S., A. Pavlova, and T. Sikorskaya, 2022, "Retail Trading in Options and the Rise of the Big Three Wholesalers," .

Campbell, J. Y., T. Ramadorai, and B. Ranish, 2015, "The impact of regulation on mortgage risk: Evidence from India," *American Economic Journal: Economic Policy*, 7(4), 71–102.

Casaburi, L., and U. Troiano, 2016, "Ghost-house busters: The electoral response to a large anti–tax evasion program," *The Quarterly Journal of Economics*, 131(1), 273–314.

Chang, Y.-c., 2012, "Self-Assessment of Takings Compensation: An Empirical Study," *The Journal of Law, Economics, & Organization*, 28(2), 265–285.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, 2002, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, C. Y., R. Despres, L. Guo, and T. Renault, 2019, "What Makes Cryptocurrencies Special? Investor Sentiment and Return Predictability During the Bubble," .

Chen, H., P. De, Y. Hu, and B. H. Hwang, 2014, "Wisdom of crowds: The value of stock opinions transmitted through social media," *Review of Financial Studies*, 27(5), 1367–1403.

Chen, L., Y. Huang, S. Ouyang, and W. Xiong, 2021, "The data privacy paradox and digital demand," working paper, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri, 2011, "Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records," *The quarterly journal of economics*, 126(2), 749–804.

Chodorow-Reich, G., G. Gopinath, P. Mishra, and A. Narayanan, 2020, "Cash and the economy: Evidence from India's demonetization," *The Quarterly Journal of Economics*, 135(1), 57–103.

Coffee Jr, J. C., 2002, "Racing towards the top: The impact of cross-listing and stock market competition on international corporate governance," *Colum. L. Rev.*, 102, 1757.

Comptroller and Auditor General of India, 2016, "Report of the Comptroller and Auditor General of India on Revenue Sector for Year Ended 21 March 2015," working paper.

Cookson, J. A., J. E. Engelberg, and W. Musslins, 2021, "Echo Chambers," .

Crouzet, N., and J. Eberly, 2018, "Intangibles, Investment, and Efficiency," *AEA Papers and Proceedings, 108 : 426-31.*

Dachis, B., G. Duranton, and M. A. Turner, 2012, "The effects of land transfer taxes on real estate markets: evidence from a natural experiment in Toronto," *Journal of economic Geography*, 12(2), 327–354.

Das, S., 2020, "Why Arun Jaitley Gave University Basic Income a Miss in Union Budget 2017," *Livemint*.

Dorn, D., and P. Sengmueller, 2009, "Trading as entertainment?," *Management Science*, 55(4), 591–603.

Eaton, G. W., T. C. Green, B. S. Roseman, and Y. Wu, 2022, "Retail Trader Sophistication and Stock Market Quality: Evidence from Brokerage Outages," *Journal of Financial Economics*, 146(2), 502–528.

Eisfeldt, A., and D. Papanikolau, 2014, "The value and ownership of intangible capital," *American Economic Review: Papers and Proceedings 104, 1-8*.

Engelberg, J., C. Sasseville, and J. Williams, 2012, "Market madness? the case of Mad Money," *Management Science*, 58(2), 351–364.

Englehardt, S., and A. Narayanan, 2016, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of ACM CCS 2016*.

Fabian, B., T. Ermakova, and T. Lentz, 2017, "Large-scale readability analysis of privacy policies," in *Proceedings of the International Conference on Web Intelligence*, pp. 18–25. ACM.

Fan, Z., X. Wang, and H. Zhang, 2022, "Understanding Misreporting: New Evidence from Responses to a Housing Transaction Tax," .

Farrell, M., R. Jame, and T. Qiu, 2020, "The Cross-Section of Non-Professional Analyst Skill," .

Fedyk, V., 2022, "This Time is Different : Investing Preferences in the Age of Robinhood," .

Finkel, J. R., T. Grenager, and C. Manning, 2005, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 363–370. Association for Computational Linguistics.

Fisman, R., and S.-J. Wei, 2004, "Tax rates and tax evasion: evidence from "missing imports" in China," *Journal of political Economy*, 112(2), 471–496.

Fong, K. Y. L., D. R. Gallagher, and A. D. Lee, 2013, "Individual Investors and Broker Types," *Journal of Financial and Quantitative Analysis*, 49(2), 431–451.

Foucault, T., D. Sraer, and D. J. Thesmar, 2011, "Individual Investors and Volatility," *The Journal of Finance*, 66(4), 1369–1406.

Gabaix, X., and D. Laibson, 2006, "Shrouded attributes, consumer myopia, and information suppression in competitive markets," *The Quarterly Journal of Economics*, 121(2), 505–540.

Garmaise, M. J., and T. J. Moskowitz, 2004, "Confronting information asymmetries: Evidence from real estate markets," *The Review of Financial Studies*, 17(2), 405–437.

Gechter, M., and N. Tsivanidis, 2020, "Spatial spillovers from urban renewal: evidence from the Mumbai Mills Redevelopment," *Penn State University (mimeo)*.

Genesove, D., and C. Mayer, 2001, "Loss aversion and seller behavior: Evidence from the housing market," *The quarterly journal of economics*, 116(4), 1233–1260.

Gentzkow, M., B. Kelly, and M. Taddy, 2018, "Text as Data," .

Goldfarb, A., and C. Tucker, 2012, "Shifts in privacy concerns," *American Economic Review*, 102(3), 349–53.

Gopalakrishnan, S., D.-G. A., 1986, "Stamp Duties and Registration Fees in West Bengal," Unpublished Manuscript.

Gordon, R., and W. Li, 2009, "Tax structures in developing countries: Many puzzles and a possible explanation," *Journal of public Economics*, 93(7-8), 855–866.

Green, J., J. R. Hand, and X. F. Zhang, 2017, "The characteristics that provide independent information about average u.s. monthly stock returns," *Review of Financial Studies*, 30(12), 4389–4436.

Griffin, J. M., and G. Maturana, 2016, "Who facilitated misreporting in securitized loans?," *The Review of Financial Studies*, 29(2), 384–419.

Grinblatt, M., S. Ikäheimo, M. Keloharju, and S. Knüpfer, 2016, "IQ and mutual fund choice," *Management Science*, 62(4), 924–944.

Grinblatt, M., and M. Keloharju, 2001, "What makes investors trade?," *Journal of Finance*, 56(2), 589–616.

——— , 2009, "Sensation seeking, overconfidence, and trading activity," *Journal of Finance*, 64(2), 549–578.

Gu, S., B. Kelly, and D. Xiu, 2020, "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 33(5), 2223–2273.

Gunning, R., 1952, *The technique of clear writing*. McGraw-Hill, New York.

Harari, M., and M. Wong, 2018, "Slum upgrading and long-run urban development: Evidence from Indonesia," in *2018 Meeting Papers. Society for Economic Dynamics*.

Harberger, A. C., 1965, *Issues of tax reform for Latin America*. PAU, Department of Economic Affairs.

Hoofnagle, C. J., and J. M. Urban, 2014, "Alan Westin's privacy homo economicus," *Wake Forest L. Rev.*, 49, 261.

Hu, D., C. M. Jones, V. Zhang, and X. Zhang, 2021, "The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery," .

Jame, R., R. Johnston, S. Markov, and M. C. Wolfe, 2016, "The Value of Crowdsourced Earnings Forecasts," *Journal of Accounting Research*, 54(4), 1077–1110.

Kang, R., S. Brown, and S. Kiesler, 2013, "Why do people seek anonymity on the internet? Informing policy and design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2657–2666.

Kang, R., L. Dabbish, N. Fruchter, and S. Kiesler, 2015, ""my data just goes everywhere:" user mental models of the internet and implications for privacy and security," in *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pp. 39–52.

Kaniel, R., G. Saar, and S. Titman, 2008, "Individual investor trading and stock liquidity," *Journal of Finance*, 63(1), 273–310.

Karmakar, S., and A. Narayanan, 2020, "Do households care about cash? Exploring the heterogeneous effects of India's demonetization," *Journal of Asian Economics*, 69, 101203.

Kelley, E. K., and P. C. Tetlock, 2013, "How Wise Are Crowds? Insights from Retail Orders and Stock Returns," *Journal of Finance*, 68(3), 1229–1265.

Khan, A. Q., A. I. Khwaja, and B. A. Olken, 2016, "Tax farming redux: Experimental evidence on performance pay for tax collectors," *The Quarterly Journal of Economics*, 131(1), 219–271.

Kleven, H. J., 2016, "Bunching," *Annual Review of Economics*, 8, 435–464.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez, 2011, "Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark," *Econometrica*, 79(3), 651–692.

Kleven, H. J., C. T. Kreiner, and E. Saez, 2016, "Why can modern governments tax so much? An agency model of firms as fiscal intermediaries," *Economica*, 83(330), 219–246.

Kleven, H. J., and M. Waseem, 2013, "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan," *The Quarterly Journal of Economics*, 128(2), 669–723.

Kopczuk, W., and D. Munroe, 2015, "Mansion tax: The effect of transfer taxes on the residential real estate market," *American economic Journal: economic policy*, 7(2), 214–57.

Krosnick, J. A., 1991, "Response strategies for coping with the cognitive demands of attitude measures in surveys," *Applied Cognitive Psychology*, 5(3), 213–236.

Kumar, A., 2009, "Who gambles in the stock market?," *Journal of Finance*, 64(4), 1889–1933.

Kummer, M., and P. Schulte, 2019, "When private information settles the bill: Money and privacy in Googles market for smartphone applications," *Management Science*, 65(8), 3470–3494.

La Porta, R., F. Lopez-de Silanes, and A. Shleifer, 2002, "Government ownership of banks," *The Journal of Finance*, 57(1), 265–301.

Lahiri, A., 2020, "The great indian demonetization," *Journal of Economic Perspectives*, 34(1), 55–74.

Li, F., 2010, "The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach," *Journal of Accounting Research*, 48(5), 1049–1102.

Liu, Z., M. Sockin, and W. Xiong, 2020, "Data privacy and temptation," working paper, National Bureau of Economic Research.

Loughran, T., and B. Mcdonald, 2011, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 66(1), 35–65.

Lush, M., A. Fontes, M. Zhu, O. Valdes, and G. Mottola, 2021, "Investing 2020: New Accounts and the People Who Opened Them," working paper February, FINRA Foundation and NORC at the University of Chicago.

Mayle, P., 2000, *A year in Provence*. Penguin UK.

McDonald, A. M., and L. F. Cranor, 2008, "The cost of reading privacy policies," *Isjlp*, 4, 543.

Mishra, P., N. Prabhala, and R. G. Rajan, 2022, "The Relationship Dilemma: Why Do Banks Differ in the Pace at Which They Adopt New Technology?," *The Review of Financial Studies*.

Montalvo, J. G., A. Piolatto, and J. Raya, 2020, "Transaction-tax evasion in the housing market," *Regional Science and Urban Economics*, 81, 103526.

Nagel, S., 2012, "Evaporating liquidity," *Review of Financial Studies*, 25(7), 2005–2039.

Narayanan, A., and V. Shmatikov, 2008, "Robust de-anonymization of large sparse datasets," *Proceedings - IEEE Symposium on Security and Privacy*, pp. 111–125.

National Institute of Public Finance and Policy, 1995, "Black Money in the Real Estate Sector: A Study," .

Odean, T., 1998, "Are investors reluctant to realize their losses?," *The Journal of Finance*, 53(5), 371–392.

——— , 1999, "Do investors trade too much?," *American Economic Review*, pp. 1279–1298.

OECD, 2007, "Report on tax fraud and money laundering vulnerabilities involving the real estate sector," OECD Centre for Tax Policy and Administration.

Okunogbe, O., and F. Santoro, 2022, "The promise and limitations of information technology for tax mobilisation," .

Panchapagesan, V., K. N., 2017, "Revenue Neutral Approach to Lower Stamp Duty and Registration Charges for Affordable Housing," Unpublished Manuscript.

Panigrahi, P., 2021, "Endogenous Spatial Production Networks: Quantitative Implications for Trade and Productivity," *Available at SSRN 3982029*.

Pedersen, L. H., 2022, "Game on: Social networks and markets," *Journal of Financial Economics*, 146(3), 1097–1119.

Peters, R. H., and L. A. Taylor, 2017, "Intangible capital and the investment-q relation," *Journal of Financial Economics*, 123(2), 251–272.

Piketty, T., 2015, "Capital and wealth taxation in the 21 st century," *National Tax Journal*, 68(2), 449–458.

Piskorski, T., A. Seru, and J. Witkin, 2015, "Asset quality misrepresentation by financial intermediaries: Evidence from the RMBS market," *The Journal of Finance*, 70(6), 2635–2678.

Pissarides, C. A., and G. Weber, 1989, "An expenditure-based estimate of Britain's black economy," *Journal of public economics*, 39(1), 17–32.

Pomeranz, D., and J. Vila-Belda, 2019, "Taking state-capacity research to the field: Insights from collaborations with tax authorities," *Annual Review of Economics*, 11, 755–781.

Posner, E. A., and E. G. Weyl, 2019, *Radical Markets*. Princeton University Press.

Rainie, L., S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish, 2013, "Anonymity, privacy, and security online," *Pew Research Center*, 5.

Rajaraman, A., and J. D. Ullman, 2011, *Mining of massive datasets*. Cambridge University Press.

Rocha, R., T. Scot, and J. Feinmann, 2023, "Property Taxes in Sao Paulo," .

Rock, D., 2019, "Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence," *Available at SSRN 3427412*.

Saez, E., 2010, "Do taxpayers bunch at kink points?," *American economic Journal: economic policy*, 2(3), 180–212.

Semenova, V., and J. Winkler, 2021, "Reddit ' s self-organised bull runs : Social contagion and asset prices," .

Slemrod, J., 2007, "Cheating ourselves: The economics of tax evasion," *Journal of Economic perspectives*, 21(1), 25–48.

Slemrod, J., C. Weber, and H. Shan, 2017, "The behavioral response to housing transfer taxes: Evidence from a notched change in DC policy," *Journal of Urban Economics*, 100, 137–153.

Sood, A., 2020, "Land Market Frictions in Developing Countries: Evidence from Manufacturing Firms in India," working paper, Mimeo. https://aradhyasood. github. io/Sood_Land_Frictions_India1. pdf.

Spence, A. M., 1975, "Monopoly, Quality, and Regulation," *Bell Journal of Economics*, 6(2), 417–429.

Stiglitz, J. E., 2015, "The origins of inequality, and policies to contain it," *National Tax Journal*, 68(2), 425–448.

Tandon, N., 1987, "Under-reporting of Real Estate Transactions in India," Unpublished Manuscript.

Tang, H., 2019, "The value of privacy: Evidence from online borrowers," working paper, Technical report, Working Paper, HEC Paris.

Tetlock, P. C., 2007, "Giving content to investor sentiment: The role of media in the stock market," *Journal of Finance*, 62(3), 1139–1168.

Tsivanidis, N., 2019, "Evaluating the impact of urban transit infrastructure: Evidence from bogota's transmilenio," *Unpublished manuscript*.

Welch, I., 2022, "The Wisdom of the Robinhood Crowd," *Journal of Finance*, 77(3), 1489–1527.